Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

Entschlüsselung der Genomsequenz von *Escherichia blattae* und komparative Bioinformatik mikrobieller Genome

vorgelegt von
Arnim Wiezer
aus Göttingen

Göttingen 2004

D 7

Referent:	Prof. Dr. G. Gottschalk
Korreferent:	PD Dr. R. Daniel
Tag der mündlichen Prüfung:	1.7.2004

Inhaltsverzeichnis

1. Einleitung **1**

2. Materialien und Methoden **4**

2.1. Organismen und Plasmide 4
2.2. Nährmedien 4
2.3. Zellanzucht und Stammhaltung 6
2.3.1. Zellanzucht 6
2.3.2. Stammhaltung 6
2.4. Techniken für die Arbeit mit Nukleinsäuren 6
2.4.1. Geräte und Lösungen 6
2.4.2. Isolierung chromosomaler DNA 7
2.4.3. Isolierung von Plasmid DNA 7
2.4.4. Isolierung von Cosmid DNA 7
2.4.5. Aufreinigung von PCR-Produkten 8
2.4.6. Isolation von DNA-Fragmenten aus Agarosegelen 8
2.4.7. Scheren von DNA 8
2.4.8. DNA Konzentrationsbestimmung und Reinheitskontrolle . . . 9
2.4.9. Dialyse von Nukleinsäuren 10
2.4.10. Saccharose-Dichtegradientenzentrifugation 10
2.4.11. Agarose Gelelektrophorese 11
2.5. Modifikaton von Nukleinsäuren 11
2.5.1. Restriktion 11
2.5.2. Dephosphorylierung 12
2.5.3. Herstellung von „blunt ends“ 12
2.5.4. Herstellung von Insert-DNA mit A-Überhang 13
2.5.5. Ligation 13
2.5.6. „TOPO“-Klonierung 14

2.5.7. Transformation von *E. coli* TOP10-Zellen 14
2.5.8. Transformation von *E. coli* DH5α-Zellen 14
2.5.9. Blau-Weiß Screening (X-Gal-Test) 15
2.6. Polymerasekettenreaktion (PCR) 16
2.7. Genbanken 18
2.7.1. Plasmidbanken 18
2.7.2. Cosmidbanken 18
2.7.3. Cosmid-Subklonierungen 19
2.8. Sequenzierung 19
2.8.1. Sequenzierung mit MegaBace 20
2.8.2. Sequenzierung mit ABI377 21
2.8.3. Sequenzabbrüche 22
2.9. Sequenzierungs-Software 25
2.10. Lückenschluss - *gap closure* 25
2.10.1. „Plasmid walking“ 25
2.10.2. Cosmid-Brücken zum Schließen von Lücken 26
2.10.3. ERGO zum Schließen von Lücken 27
2.10.4. Lückenschluss mit BLAST 29
2.10.5. Multiplex-PCR 30
2.11. BLAST 31
2.11.1. Bi-direktionales BLAST 31
2.12. Annotation 32
2.12.1. ORF Vorhersage 33
2.12.2. ERGO 33

3. Ergebnisse **35**
3.1. Genomprojekt *Escherichia blattae* 35
3.1.1. Der Organismus 35
3.1.2. Sequenzierung 36
3.1.3. Rohsequenzierung 36
3.1.4. Lückenschluss 38
3.1.5. Qualitätsverbesserung 40
3.1.6. Annotation 40
3.1.7. Horizontaler Gentransfer in *Escherichia blattae* 45
3.1.8. Vergleich mit anderen Enterobakterien 47

3.2. Untersuchung des *Methanosarcina mazei* Gö1 Genoms 49
3.2.1. Lokalisation des Replikationsursprungs 50
3.2.2. Vergleich mit *Methanosarcina acetivorans* C2A 56
3.2.3. Horizontaler Gentransfer in *Ms. mazei* - „first hit"-Methode . . 64
3.2.4. Horizontaler Gentransfer II - Vorhersagen mit SIGI 65
3.3. Untersuchung des *Thermus thermophilus* HB27 Genoms 66
3.3.1. Bestimmung des Replikationsursprungs 67
3.3.2. Vergleich mit *Deinococcus radiodurans* R1 68
3.3.3. Horizontaler Gentransfer in *Thermus thermophilus* 76
3.4. Programme 79
3.4.1. Quann 79
3.4.2. Cogger 81
3.4.3. secureBLAST 83

4. Diskussion **85**
4.1. Genomsequenzierungsprojekt *Escherichia blattae* 85
4.1.1. Sequenzierung 86
4.1.2. Annotation 86
4.1.3. Horizontaler Gentransfer in *Escherichia blattae* 87
4.1.4. Vergleich mit anderen Enterobakterien 88
4.2. Untersuchung des *Methanosarcina mazei* Gö1-Genoms 89
4.2.1. Der Replikationsursprung 90
4.2.2. Vergleich mit *Ms. acetivorans* 92
4.2.3. Horizontaler Gentransfer in *Ms. mazei* 95
4.3. Untersuchung des *Thermus thermophilus* HB27-Genoms 98
4.3.1. Lokalisation des Replikationsursprungs 98
4.3.2. Vergleich mit *Deinococcus radiodurans* R1 99
4.3.3. Horizontaler Gentransfer in *Thermus thermophilus* 103
4.4. Programme 105
4.4.1. Quann 105
4.4.2. Cogger 106
4.5. Ausblick 108

5. Zusammenfassung **110**
5.1. Genomprojekt *Escherichia blattae* 110

5.2. Das *Methanosarcina mazei* Gö1-Genom 110
5.3. Das *Thermus thermophilus* HB27-Genom 111
5.4. Programmentwicklungen 111

Literaturverzeichnis **113**

A. Anhang **124**
A.1. bi-direktionales BLAST 124
A.2. *Ms. mazei*-spezifische Proteine 127
A.3. *Ms. acetivorans*-spezifische Proteine 131
A.3.1. COG Gruppen 143
A.4. Bakterielle Gene nach der *first-hit* Methode in *Ms. mazei* 143
A.5. Fremdgene in *Ms. mazei* SIGI-Vorhersage 149
A.6. *T. thermophilus* Proteine ohne Homolog in *D. radiodurans* 153
A.7. Fremdgene in *T. thermophilus* 162
A.8. *T. thermophilus*-spezifische Proteine (COG-Gruppe C) 163
A.9. Gecoggte *D. radiodurans*-spezifische Proteine 164
A.10. *D. radiodurans*-spezifische Proteine (COG-Gruppe K) 171
A.11. *D. radiodurans*-spezifische Proteine (COG-Gruppe V) 172
A.12. Gecoggte *T. thermophilus*-spezifische Proteine 172
A.13. Fremdgene in *E. blattae* 176
A.14. *E. blattae*-spezifische Proteine 180
A.15. *E. blattae*-spezifische Proteine (Fremdgene) 187
A.16. *E. blattae*-spezifische Proteine (Vergleich Enterobakterien) 189

Abkürzungen

A	Adenin
aa	Aminosäuren (engl.: amino acids)
Amp	Ampicillin
B.	Buchnera
(bi)dest.	(zweifach) destilliert
bp	Basenpaare
C	Cytosin
°C	Grad Celsius
ca.	circa
DMSO	Dimethylsulfoxid
DNA	Desoxyribonukleinsäure
E.	Escherichia
E-cup	Eppendorf-Reaktionsgefäß
EDTA	Ethylendiamintetracetat
et al.	et alii (und andere)
EtOH	Ethanol
Fa.	Firma
G	Guanin
g	Gramm
h	Stunde
IPTG	Isopropyl-β-d-thiogalactopyranosid
KAc	Kalium-Acetat
kbp	Kilobasen
Kan	Kanamycin
Konz.	Konzentration
l	Liter
LB	Luria-Bertani

M	molar
Mbp	Mega Basenpaare
MCS	multiple cloning site
mg	Milligramm
μg	Mikrogramm
min	Minuten
ml	Milliliter
mM	millimolar
μm	Mikrometer
NCBI	National Center for Biotechnology Information
OD	optische Dichte
ORF	offener Leserahmen (engl.: open reading frame)
PCR	Polymerasekettenreaktion (engl.: Polymerase Chain Reaction)
r	Resistenz
PEG	Polyethylenglycol
rpm	Umdrehungen pro Minute (engl.: rounds per minute)
RT	Raumtemperatur
S.	Salmonella
Sek	Sekunde
Sh.	Shigella
T	Thymin
Tab.	Tabelle
TEMED	N, N, N, N', N'-Tetramethylethylendiamid
Tris	Tris(hydroxymethyl)-aminoethan
U	Unit (Einheit für Enzymaktivität)
Vol.	Volumen
v/v	Volumeneinheit pro Volumeneinheit
w/v	Masseneinheit pro Volumeneinheit
X-Gal	5-Brom-4-chlor-3-indolyl-β-D-galactopyranosid
Y.	Yersinia

1. Einleitung

Stellen Sie sich vor, Sie müssten an einem Projekt mitarbeiten, das wie folgt beschrieben wird:

> Es geht darum, den Inhalt einer Datei von 3,2 Gigabyte zu verstehen, die im Massenspeicher eines Computers mit unbekannter Konstruktion abgelegt ist. Das Verständniss dieser Datei ist gleichbedeutend mit dem *reverse engineering* des Computersystems bis hin zur vollständigen Beschreibung und seiner Wartung.
>
> Das *reverse engineering* der Datei wird dadurch erschwert, dass die Datei „ein Haufen von Bytes“ in der Reihenfolge ist, in der sie in das Gerät eingegeben wurden. Außerdem ist bekannt, dass die Dateien fragmentiert vorliegen und manche Teile des Geräts enthalten auch gelöschte Dateien und anderen „Müll“. Nachdem man den Abfall identifiziert und verworfen hat, und nachdem dann die fragmentierten Dateien zusammengefügt wurden, kann man mit dem *reverse engineering* des Codes beginnen. Dabei verfügt man nur über unvollständige und manchmal auch falsche Erkenntnisse über die CPU (Zentrale Recheneinheit), auf der der Code läuft. Die Aufklärung von Aufbau und Funktion der CPU ist ebenfalls ein Teil des Projekts, denn manche Teile der 3,2 Gigabyte sind binäre Anweisungen für den computergestützten Produktionsprozeß, durch den die CPU hergestellt wird. Darüber hinaus gilt es zu berücksichtigen, dass zu dem riesigen Datenbestand auch Code gehört, der bei buchstäblich Millionen von Wartungsarbeiten entstanden ist, und diese Wartungsarbeiten wurden von den schlimmsten, opportunistischsten Hackern ausgeführt, die sich fauler Tricks bedienen, wie Spaghetti verworrenen Code benutzen, sich selbst modifizierenden Code schreiben und sich auf undokumentierte Eigenheiten des Systems verließen.

Diese Vorstellung wurde sinngemäß von Robins [79] übernommen und beschreibt die Aufgabe der Entschlüsselung des menschlichen Genoms verglichen mit der voll-

ständigen Aufklärung eines unbekannten Computers aus der Sicht der Wissenschaft von 1992.

Die erste vollständige mikrobielle Genomsequenz, die von *Haemophilus influenzae* Rd wurde 1995 veröffentlicht [26]. Seit diesem Datum hat nicht nur die Anzahl an DNA-Sequenzen in den öffentlichen Datenbanken stark zugenommen, sondern auch die Anzahl vollständiger Genomsequenzen. Diese immer schneller wachsende Sequenzmenge stellt große Herausforderungen an ihre Verarbeitung. Die Bioinformatik hat diese Herausforderung angenommen.

Die Situation bei Mikroorganismen ist ähnlich kompliziert wie im obigen Beispiel, nur die Genome sind mit einigen Megabasen deutlich kleiner. Nach beendeter Sequenzierung eines Organismus existiert eine Menge von Buchstaben, die letztendlich mit Leben gefüllt werden müssen. Diese Aufgabe beginnt mit dem Finden der offenen Leserahmen (ORFs). Hierfür existieren mittlerweile Programme, die alle ihre Vor- und Nachteile haben, in Kombination angewandt aber gute Ergebnisse erzielen [91]. Der entscheidende Schritt, um einer Sequenz ihre biologische Information zu entlocken, besteht in der Annotation der putativen Proteine, die aus den ORFs abgeleitet wurden. Hierzu existieren einige Programme wie z. B. ERGO oder GenDB [74, 64], mit deren Hilfe diese Aufgabe gelöst werden kann. Hinzu kommen eine Reihe weiterer Werkzeuge, von denen hier nur das COG-System zur Klassifizierung von Proteinen genannt werden soll [90]. Bringt man die vorhandenen Annotationen in einen Gesamtzusammenhang und weist die einzelnen Proteine ihren biochemischen Stoffwechselwegen zu, ist man von der *in silico* Rekonstruktion eines Organismus schon nicht mehr so weit entfernt. Bleibt man im Bild des obigen Beispiels, so ist nach der Annotation der Computer in großen Teilen beschrieben, also *reverse engineered*. Die Informationen, die man bei der Auswertung einer Genomsequenz gewonnen hat, kann man außerdem benutzen, um andere Sequenzen zu beschreiben oder Verwandtschaftsverhältnisse innerhalb der Mikroorganismen zu untersuchen.

In dieser Arbeit werden mehrere bioinformatische Aspekte der Genomanalyse an Mikroorganismen vorgestellt. So wird in die Sequenzierung anhand des *Escherichia blattae* Genoms eingeführt. Die in dieser Arbeit ermittelte Sequenz wird hinsichtlich der Proteine und deren Zugehörigkeit zu Proteinklassen untersucht. Das Proteom des Organismus wird mit denen anderer Enterobakterien verglichen und auf mögliche Fremdgene hin analysiert. Anhand der Genome von *Methanosarcina mazei* Gö1 und *Methanosarcina acetivorans* C2A wird ein Genomvergleich nah verwandter Archaeen vorgestellt. In *Ms. mazei* wird außerdem gezeigt, welche Probleme bei der *in*

silico Bestimmung des Replikationsursprungs von Archaeen auftreten. Das Genom des methanogenen Archaeons *Ms. mazei* wird außerdem mit zwei verschiedenen Methoden auf putativ horizontal transferierte Gene hin untersucht. Weiterhin werden mehrere Aspekte des Genoms des thermophilen Bakterium *Thermus thermophilus* HB27 betrachtet. Dazu gehören die Lokalisation des Replikationsursprungs und die Existenz putativer Fremdgene. In einem weiteren Genomvergleich zwischen *Thermus thermophilus* und *Deinococcus radiodurans* R1 wird die Verwandtschaft dieser beiden Organismen näher beleuchtet.

2. Materialien und Methoden

2.1. Organismen und Plasmide

Im Rahmen dieser Arbeit wurden verschiedene Stämme und Plasmid- sowie Cosmidvektoren verwendet. Sie sind in den Tabellen 2.1 und 2.2 dargestellt.

Tabelle 2.1.: Verwendete Organismen

Stamm	Geno-/Phänotyp	Herkunft
Escherichia blattae		BURGESS et al. [12]
Escherichia coli DH5α	*sup*E44, ϕ80*lac*ZΔM15, Δ*lac*U169, *hsd*R17, *rec*A1, *end*A1, *gyr*A96, *thi*-1, *rel*A1	Gibco, [36]
Escherichia coli TOP10	F-, *mcr*A, ϕ80*lac*ZΔM15, Δ(*mrr*-*hsd*RMS-*mcr*BC), Δ*lac*X74, *rec*A1, *ara*D139, (*ara*-*leu*)7697, *gal*U, *gal*K, *rps*L, (*Str*^{R}), *end*A1, *nup*G	Invitrogen, Karlsruhe, Deutschland

Tabelle 2.2.: Verwendete Plasmide und Cosmide

Plasmid/Cosmid	Eigenschaften	Herkunft
PCR4-TOPO	Amp^{r}, Kan^{r}, *lac*Zα-*ccd*B	Invitrogen
pBluescript SK^{+}	Amp^{r}, *lacPOZ'*, *colE1*, *oriR*	Stratagene
SuperCos I	Amp^{r}, Kan^{r}, *cos*	Stratagene, San Diego, USA

2.2. Nährmedien

Die verwendeten Medien wurden durch Autoklavieren für 30 Minuten bei 121 °C sterilisiert. Zur Herstellung von festen Medien wurde 1,5 % (w/v) Agar hinzugefügt.

Tabelle 2.3.: Luria-Bertani Medium (LB) [82]

Bestandteil	Menge
Trypton	10 g
Hefeextrakt	5 g
NaCl	10 g
$H_2O_{bidest.}$	ad 1000 ml

Tabelle 2.4.: TB-Medium

Bestandteil	Menge
Trypton	12 g
Hefeextrakt	24 g
Glycerin 87 % (v/v)	4 ml
$H_2O_{bidest.}$	ad 1000 ml

Für die Anzucht in Flat-Bottom-Blocks (Qiagen) wurden zu 400 ml TB-Medium je 8 ml 5 x Phosphatpuffer (s. Tabelle 2.6) und Glucose-Stammlösung (10 %, v/v) gegeben.

Tabelle 2.5.: Medienzusätze

Zusatz	Stammlösung	Endkonzentration
Ampicillin	100 mg/ml in 50 % EtOH [a]	100 μg/ml
Kanamycin	100 mg/ml in $H_20_{bidest.}$ [a]	25..50 μg/ml
IPTG	100 mg/ml in $H_20_{bidest.}$ [b]	40 μg/ml
X-GAL	40 mg/ml in N,N-Dimethylformamid [b]	40 μg/ml

a) Lagerung bei 4 °C b) vor Gebrauch frisch angesetzt. Medienzusätze wie Antibiotika, X-GAL oder IPTG wurden nach dem Autoklavieren und Abkühlen der Medien auf ca. 60 °C zugegeben.

Tabelle 2.6.: 5 x Phosphat Puffer

Bestandteil	Menge
KH_2PO_2	57,75 g
K_2HPO_4	313,50 g
$H_2O_{bidest.}$	ad 500 ml

2.3. Zellanzucht und Stammhaltung

2.3.1. Zellanzucht

Escherichia blattae wurde über Nacht in Reagenzgläsern in 5 ml LB-Medium bei 37 °C inkubiert. Zur Isolation genomischer DNA wurde der Organismus in 100 ml Kulturen aerob und bei 120 rpm im Schüttler angezogen. Alle Kulturen wurden mit 1,5 % Inokulum angeimpft. Die Anzucht von rekombinanten *Escherichia coli* DH5α Zellen erfolgte in Flat-Bottom-Blocks (Qiagen), die nach dem Animpfen mit luftdurchlässigen Folien (Aerosheet) abgeklebt wurden. Die einzelnen 96 Wells der Blöcke wurden jeweils mit 1,3 ml TB-Medium mit Ampicillin, Glucose und Phosphatpuffer (s. Tabellen 2.4, 2.6 und 2.5) befüllt. Die Inkubation erfolgte für 20..26 Stunden bei 37 °C mit 360 rpm im Brutschrank.

2.3.2. Stammhaltung

Zur kurzfristigen Stammhaltung wurden *E. blattae* sowie die rekombinanten *E. coli*-Stämme auf LB-Agar-Platten ausgestrichen, über Nacht bei 37 °C inkubiert und dann bei 4 °C gelagert. Zur langfristige Lagerung von *E. blattae* wurde der Organismus in 5 ml LB-Medium bei 37 °C inkubiert. In der späten logarithmischen Wachstumsphase wurden 750 μl Zellsuspension mit 250 μl DMSO versetzt und bei -80 °C gelagert.

Für die Stammhaltung der rekombinanten *E. coli*-Klone der Sequenzierungs-Genbanken im 96er Format wurden 1,3 ml Kulturen wie oben beschrieben inkubiert und in Mikrotiterplatten mit 230 μl TB-Medium/Well überstempelt. Diese Mikrotiterplatten wurden über Nacht bei 37 °C inkubiert, anschließend mit 23 μl DMSO/Well versetzt und bei -80 °C gelagert.

2.4. Techniken für die Arbeit mit Nukleinsäuren

2.4.1. Geräte und Lösungen

Zur Inaktivierung von Nukleasen wurden alle hitzestabilen Geräte und Lösungen durch Autoklavieren (20 Minuten bei 121 °C) sterilisiert. Materialien, die nicht autoklavierbar sind, wurden mit 70 % (v/v) Ethanol gespült oder abgeflammt. Hitzelabile Biocheminkalien, wie z. B. Antibiotika, wurden als Lösung sterilfiltriert.

2.4.2. Isolierung chromosomaler DNA

Zur Isolierung chromosomaler DNA wurde *E. blattae* in 300 ml LB-Medium über Nacht bei 37 °C und 120 rpm im Schüttler inkubiert. Die Isolierung der DNA wurde am nächsten Tag mit Hilfe des *AquaPure DNA*-Kits der Firma Biorad nach Protokoll durchgeführt, abschließend allerdings wurde die DNA in Wasser statt TE-Puffer gelöst.

2.4.3. Isolierung von Plasmid DNA

Zur Isolierung von Plasmid DNA, die nicht zur Sequenzierung bestimmt war, wurde das *Qiagen Mini-Prep Kit* (QIAprep Spin Miniprep Kit, Quiagen) nach Anleitung benutzt.

Die Plasmide für die Rohsequenzierungsphase des Genomprojektes wurden mit Hilfe von Qiagen BioRobot 8000 Robotern isoliert. Dabei kam das *QIAprep 96 Turbo BioRobot Kit* zum Einsatz. Hierzu wurden die Klone wie unter 2.3.1 beschrieben im 96er Maßstab in Flat Bottom Blocks angezogen und dann ebenfalls im 96er Maßstab im BioRobot weiterverarbeitet.

Die Flat Bottom Blocks wurden nach der Inkubation (s. 2.3.1) für 20 Minuten bei 3300 rpm abzentrifugiert (Megafuge 2.0, Fa. Heraeus) und der Überstand verworfen. Der Block mit den Zellpellets wurde in den BioRobot 8000 gestellt und dann automatisch weiterbearbeitet. Die Schritte der Plasmidisolation entsprechen der Alkalischen Lyse mit anschließender Aufreinigung der DNA durch eine Silikat-Gel-Membran, die bis zu 20 μg Plasmid-DNA binden kann. Zellreste und denaturierte Proteine werden über einen Filter vor der Auftragung des Lysats auf die Membran abgetrennt. Nach einem Waschschritt erfolgt die Elution der Plasmid-DNA mit 120 μl Millipore-Wasser in Mikrotiterplatten in 96-Well Mikrotiterplatten mit Flachboden (UV-geeignet, Fa. Greiner).

2.4.4. Isolierung von Cosmid DNA

Zur Anzucht und Isolation der Cosmide im 96er Maßstab wurden die Cosmidklone in Reagenzgläsern mit 5 ml TB-Medium, Glucose und Phosphatpuffer (s. Tabelle 2.4) über Nacht bei 37 °C und 180 rpm angezogen. Zur Isolierung der DNA wurde das *QIAprep Spin Miniprep Kit* verwendet. Abweichend vom Protokoll wurden 3 ml Kultur abzentrifugiert und die Cosmide mit 70 °C heißem HPLC-Wasser eluiert.

2.4.5. Aufreinigung von PCR-Produkten

Zur Sequenzierung von PCR-Produkten ist es erforderlich, noch vorhandene dNTPs, Primer, Puffer und Enzym-Reste aus dem PCR-Ansatz zu entfernen. Dazu wurde das *QIAquick PCR Purification Kit* verwendet. Die Aufreinigung wurde analog zum Protokoll durchgeführt mit der Abweichung, dass die Säule nach dem Waschen mit PE-Puffer zusätzlich zwei mal zwei Minuten zentrifugiert (13.000 rpm) wurde, um alle Spuren von PE-Puffer zu entfernen.

2.4.6. Isolation von DNA-Fragmenten aus Agarosegelen

Die direkte Isolation von DNA-Banden aus Agarosegelen wurde unter anderem dann erforderlich, wenn eine PCR-Reaktion mehrere Banden ergab. Um alle Banden einzeln zu sequenzieren, wurde jede Bande mit einem Skalpel aus dem Agarosegel ausgeschnitten. Mit dem *QIAquick Gel Extraction Kit* wurde die DNA aus den Agarose-Stücken isoliert und war danach direkt zur Sequenzierung einsetzbar.

2.4.7. Scheren von DNA

Zur Erstellung von Genbanken wird das Scheren von DNA häufig angewendet. Dabei wird die DNA durch mechanische Kräfte gebrochen und führt somit zu DNA-Fragmenten, deren Größenverteilung unabhängig von Erkennungsstellen ist. Dieses ist vorteilhaft, da man davon ausgehen kann, dass kein Bereich der chromosomalen DNA in Genbanken über- oder unterproportional vertreten ist.

In dieser Arbeit wurden Nebulizer (Fa. Invitrogen) zum Scheren von DNA verwendet. Bei diesen Geräten (s. Abbildung 2.1) wird die DNA-Lösung in einen nach unten spitz zulaufenden Zylinger gegeben und durch Druckluft oder Stickstoff aus der Spitze des Zylinders in einen dünnen Schlauch gepresst, der sich am oberen Ende des Zylinders öffnet und die Lösung wieder in den Boden des Zylinders freigibt.

In den Nebulizer wurden 750 μl *shearing buffer* (s. Tabelle 2.7) und 15 μg DNA gegeben. Dann wurde der Nebulizer fest verschraubt und während des ganzen Schervorgangs auf Eis gelagert. Als nächstes wurde der Nebulizer an die Stickstoffversorgung angeschlossen und ein Druck von 0,62 bar angelegt. Die Dauer des Scherens wurde in Abhängigkeit von der gewünschten Fragmentgröße empirisch ermittelt. Zur Weiterverarbeitung der gescherten DNA war es erforderlich, die DNA-Lösung aus dem Nebulizer in einer SpeedVac auf ca. 200 μl einzudampfen.

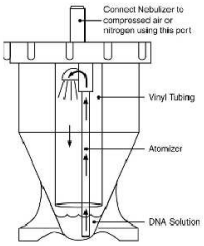

Abbildung 2.1.: Aufbau eines Nebulizers (Fa. Invitrogen)

Tabelle 2.7.: *Shearing Buffer*

Bestandteil	Menge
Tris	121,1 mg
EDTA pH 8,0	37,23 mg
Glycerin	10 ml

2.4.8. DNA Konzentrationsbestimmung und Reinheitskontrolle

Zur Konzentrationsbestimmung wässriger DNA-Lösungen wurde die Absorption im Photometer (Uvikon 810 Photometer, Kontron Instruments, Eching) bei 260 nm gemessen. Dabei kam eine 1 ml Quarzküvette mit 1 cm Schichtdicke zum Einsatz. Der Nullwert wurde mit Wasser eingestellt. Eine OD_{260} von 1 entspricht 50 μg / ml doppensträngiger DNA, 40 μg / ml einzelsträngiger DNA/RNA oder 31 μg / ml Oligonukleotiden [82].

Die Reinheit von DNA-Lösungen kann durch Bestimmen der optischen Dichte bei 230 nm und 280 nm gemessen werden, da Proteine bei 280 nm absorbieren. Für reine DNA nimmt man einen Koeffizienten von 1,8 (OD_{260}/OD_{280}) an. Proteinverunreinigungen führen zu einer stärkeren Absorption bei 280 nm. Um eine Verunreinigung durch eventuell vorhandene Polysaccharide auszuschließen, wurde das Verhältnis von OD_{230} : OD_{260} : OD_{280} ermittelt. Es sollte bei 0,45 : 1 : 0,515 liegen.

Eine grobe Abschätzung der DNA-Konzentration konnte auch anhand der Intensität von Banden in der Agarose Gelelektrophorese vorgenommen werden.

Plasmide, die mit dem BioRobot isoliert wurden, wurden in 96er Mikrotiterplatten eluiert und ihre Konzentration im Plattenphotometer (PowerWave X, Fa. Bio-Tek Instuments) bestimmt.

2.4.9. Dialyse von Nukleinsäuren

Zur Reinigung von DNA-Lösungen wurde die Mikrodialyse verwendet. Dabei wird eine wässerige DNA-Lösung von niedermolekularen Substanzen wie z. B. Salzen befreit. Dieses Verfahren erhöht die Ligationseffizienz (s. 2.5.5) und ermöglicht die Transformation (s. 2.5.8) von *E. coli* durch Elektroporation. Zur Dialyse wurden 20..50 μl Lösung auf die glänzende Seite eines sterilen Membranfilter (0,025 μm Porengröße, VS-Filter der Fa. Millipore) pipettiert und gegen steriles Wasser dialysiert. Nach 30..60 Minuten wurde die DNA-Lösung vorsichtig vom Filter abgenommen.

2.4.10. Saccharose-Dichtegradientenzentrifugation

Die Auftrennung von DNA-Fragmenten kann durch unterschiedliche Verfahren erfolgen. Am häufigsten verwendet werden dabei die Agarose-Gelelektrophorese und die Dichtegradientenzentrifugation [82]. Zur Herstellung der Sequenzierungs-Genbanken wurde die Zentrifugation im Saccharose-Dichtegradienten [82] verwendet. Diese Methode hat den Vorteil, dass die aufgetrennten DNA-Fragmente nahezu unbeschädigt zurückgewonnen werden können.

Durchführung

Als erstes wurde in 10 ml Zentrifugenröhrchen (Sorvall Polyallomer 3699, Fa. Du Pont, Dreieich) ein Saccharose-Gradient (10..40 %, w/v) hergestellt. Dazu wurden verschiedene Saccharoselösungen angesetzt (32,5 %, 25 %, 17,5 % und 10 %, w/v in 1 M NaCl, 20 mM Tris-HCl, 5 mM EDTA, pH 8,0). Die Lösungen wurden beginnend mit der höchsten Konzentration, in die Zentrifugenröhrchen geschichtet und eingefroren. Auf den Gradienten wurden 100..200 μg partiell verdaute DNA aufgetragen. Die anschließende Zentrifugation erfolgte in einem „swing-out“-Rotor (Sorvall TH641) in einer Ultrazentrifuge (Sorvall OTD 55 B) bei 27.000 rpm und 20 °C für 24 h. Nach der Zentrifugation wurden vorsichtig jeweils 300 μl-Fraktionen von oben nach unten aus dem Zentrifugationsröhrchen entnommen.

Zur Fällung von DNA aus dem Saccharose-Dichtegradienten wurden die Fraktionen aus dem Gradienten mit 0,5 Volumen PEG-NaCl-Lösung (20 % PEG; 6000, 2,5 M NaCl) versetzt und über Nacht bei Raumtemperatur inkubiert. Am nächsten Tag wurde die Lösung 30 Minuten bei 22.650 rpm zentrifugiert. Nachdem der Überstand abgenommen und das Pellet in 500 μl Ethanol (70 % v/v) aufgenommen

wurde, erfolgte eine weitere Zentrifugation unter gleichen Bedingungen. Der Überstand wurde abgenommen, das Pellet getrocknet und in 30 μl sterilem H_2O gelöst. Zur Überprüfung der Fraktionen werden 10 μl von jeder dritten Fraktion auf einem Agarosegel überprüft (s. 2.4.11).

2.4.11. Agarose Gelelektrophorese

Zur analytischen und präperativen Auftrennung von DNA wurden Agarosegele in horizontalen Kammern der Größe 11 cm · 6,5 cm verwendet. Die Agarose (Seakem LE Agarose, Fa. Cambrex) wurde 0,8 prozentig (w/v) in einfach konzentriertem TAE-Puffer (s. Tabelle 2.8) angesetzt. Derselbe Puffer diente auch als Laufpuffer. Die aufzutrennende DNA wurde vor dem Lauf mit 0,2 Vol. *Loading-buffer* Lösung (Fa, Fermentas) versetzt, um sie zu beschweren und die Lauffront während der Auftrennung verfolgen zu können. Als Größenreferenz wurden auf jedes Gel zusätzlich 3,5 μl Größenstandard aufgetragen („GeneRuler DNA Ladder Mix“, Fa. Fermentas). Nach der Elektrophorese wurden die Gele in einem Ethidiumbromid-Bad (1 μg/ml) für ca. 10 Minuten gefärbt und dann gewässert, um überschüssiges Ethidiumbromid zu entfernen. Die Visualisierung der DNA-Banden erfolgte durch UV-Licht (254 nm) mit einem Gel-Dokumentationsystems (MultiImage LightCabinet, Fa. Alpha Innotech).

Tabelle 2.8.: 20 x TAE Puffer

Bestandteil	Menge
Tris	96,9 g
EDTA	7,5 g
Eisessig	22,8 ml
$H_2O_{bidest.}$	ad 1000 ml

2.5. Modifikaton von Nukleinsäuren

2.5.1. Restriktion

In dieser Arbeit wurden Restriktionsnukleasen zu unterschiedlichen Zwecken eingesetzt. Zum einen dienten sie bei der Genbankherstellung dazu, chromosomale DNA partiell in lineare Stücke unterschiedlicher Größen zu schneiden und zum anderen um Restriktionskarten von Plasmiden oder Cosmide zu erstellen. Damit konnte schnell

ein Überblick über die korrekte Anordnung von Plasmiden oder Cosmiden in der Genomdatenbank von *E. blattae* erhalten werden.

Durchführung

Die Enzyme wurden in den jeweiligen Puffersystemen und Temperaturspektren der Hersteller benutzt. Es kamen 3..5 U Enzym pro Mikrogramm DNA zum Einsatz. Zur vollständigen Restriktion von Plasmiden oder Cosmiden wurden die Ansätze für ca. 2 Stunden inkubiert.

Bei der partiellen Restriktion wurde die Inkubationsdauer variiert, um eine möglichst statistische Verteilung der Fragmente zu erreichen. Zur Inaktivierung der Restriktionsenzyme wurden die Ansätze nach der Restriktion für ca. 20 Minuten bei 65 °C inkubiert oder mit 50 mM EDTA versetzt. EDTA sorgt in diesem Fall dafür, dass alle freien Mg^{2+}-Ionen komplexiert werden und so dem Restriktionsenzym nicht mehr als Cofaktor zur Verfügung stehen.

2.5.2. Dephosphorylierung

Bei der Ligation von DNA-Fragmenten mit überhängenden Enden („sticky ends“) wurde das Insert mit alkalischer Phosphatase behandelt. Das Enzym entfernt von linearer DNA die 5’-Phosphatgruppe und verhindert so, dass mehrere Inserts aneinander ligieren.

Durchführung

Die Insert-DNA wurde mit 1 U alkalischer Phosphatase pro Mikrogramm DNA versetzt und eine Stunde bei 37 °C inkubiert. Danach wurde der Reaktionsansatz mit Hilfe des *QIAquick PCR Purification Kit* von Enzym und Puffer befreit (s. 2.4.5).

2.5.3. Herstellung von „blunt ends“

Beim Scheren von DNA wird diese gebrochen, wobei undefinierte Enden entstehen. Um eine Klonierung vornehmen zu können, braucht man allerdings definierte Enden. Zu diesem Zweck werden die Enden der gescherten DNA mit Hilfe der T4-DNA-Polymerase aufgefüllt. Nach dieser Reaktion liegen „blunt ends“ vor und die so „geglättete“ Insert-DNA kann weiterverarbeitet werden. Der folgende Ansatz wurde für 1,5 Stunden bei Raumtemperatur inkubiert

50 μl Ansatz	
Insert DNA	ca. 38 μl
5x Puffer T4-DNA-Polymerase	10 μl
dNTP Mix je 10 mM	1,5 μl
T4-DNA-Polymerase	1 μl

Anschließend wurde der Ansatz mit dem *QIAquick PCR Purification Kit* (s. 2.4.5) aufgereinigt. Die Elution erfolgte mit 50 μl H_2O.

2.5.4. Herstellung von Insert-DNA mit A-Überhang

Bei der Genbankherstellung mit dem TOPO-TA-Kit (Invitrogen) ist es erforderlich, dass die Insert-DNA am 5'-Ende eine Adenin-Base als Überhang aufweist, da der TOPO-Vektor einen entsprechenden Thymidin-Überhang enthält. Nachdem die Enden der gescherten DNA als „blunt end“ vorlagen (s. 2.5.3), wurde mittels Taq-Polymerase und dATP das Verlängern des DNA-Strangs um eine Adenin-Base durchgeführt:

70 μl Ansatz	
„blunt-end“ DNA	50 μl
dATP (0,5 mM)	6 μl
$MgCl_2$	6 μl
10 x *PCR Buffer without* $MgCl_2$	7 μl
Taq DNA-Polymerase (5U/μl)	0,5 μl

Die Reaktion erfolgte 25 Minuten lang bei 72 °C. Anschließend wurde der Ansatz mit dem *QIAquick PCR Purification Kit* aufgereinigt (s. 2.4.5). Die Elution erfolgte mit 50 μl H_2O.

2.5.5. Ligation

DNA-Ligasen schließen die Phosphodiester-Bindung zwischen einer 5'-Phosphatgruppe an einem DNA-Ende und der 3'-OH-Gruppe am anderen DNA-Ende. Um eine effektive Ligation zu gewährleisten, wurde Vektor- und Insert-DNA im Verhältnis 1:3 eingesetzt. Der Ligationsansatz wurde mit 1 U T4-DNA-Ligase (Fa. Promega) sowie dem empfohlenen Puffersystem angesetzt und über Nacht bei 16 °C inkubiert. Zur Berechnung der Menge an Insert-DNA für die Ligation wurde folgende Formel verwendet:

$$\text{ng}_{\text{ Insert}} = \frac{\text{ng}_{\text{ Vektor}} \cdot \text{kb}_{\text{ Insert}} \cdot 3}{\text{kb}_{\text{ Vektor}}}$$

2.5.6. „TOPO“-Klonierung

Zur Subklonierung von Cosmiden wurde das *TOPO-TA Cloning Kit for Sequencing* (Fa. Invitrogen) verwendet. Das Prinzip der TA-Klonierung besteht darin, dass das Insert einen Adenin-Überhang und der Vektor einen Thymidin-Überhang aufweist (s. 2.5.4). Damit ist eine Ligation mehrerer Inserts aneinander ausgeschlossen. Außerdem enthält der Vektor eine kovalent gebundene Topoisomerase I aus dem Vaccina Virus, die die Spaltung der Insert-DNA an speziellen Stellen (5'-CCCTT) und die nachfolgende Re-ligation ermöglicht [86]. Die Vorteile dieser Methode liegen in ihrer Geschwindigkeit und in der Vermeidung von Insert-Insert Ligationen. Der folgende Ansatz wurde 2 Stunden bei RT inkubiert:

4 μl Ansatz	
Insert-DNA mit A-Überhang	3 μl
Salt Soultion (TOPO-Kit, Invitrogen)	0,5 μl
Vector PCR4-TOPO	0,5 μl

Anschließend erfolgte eine Transformation in chemisch kompetente *E. coli*-TOP10-Zellen (s. 2.5.7).

2.5.7. Transformation von *E. coli* TOP10-Zellen

70 μl *E. coli* TOP10-Zellen wurden mit 4 μl „TOPO“-Klonierungsreaktion vorsichtig gemischt (s. 2.5.6). Der Ansatz wurde für 10 Minuten auf Eis und anschließend für 45 Sekunden bei 42 °C inkubiert. Danach wurde der Ansatz sofort auf Eis gestellt und mit 400 μl LB-Medium versehen. Abschließend wurden die frisch transformierten Zellen für 1,5 Stunden bei 37 °C und 180 rpm inkubiert. Die Selektion erfolgte auf LB-Platten mit Ampicillin. Die Platten wurden über Nacht bei 37 °C inkubiert.

2.5.8. Transformation von *E. coli* DH5α-Zellen

Die Transformation von *E. coli* mit Plasmiden ist unter anderem durch Elektroporation möglich [25]. Im Genomprojekt *E. blattae* wurden die Plasmidgenbanken mit elektrokompetenten *E. coli* DH5α-Zellen nach der folgender Methode hergestellt.

Herstellung kompetenter Zellen

E. coli DH5α wurde über Nacht in einer 3 ml Vorkultur bei 37 °C angezogen. Am nächsten Morgen wurde mit dieser Vorkultur eine 300 ml Hauptkultur angeimpft und unter starkem Schütteln bis zu einer OD_{600} von 0,7..0,8 inkubiert. Die Zellen wurden dann für 15 Minuten auf Eis inkubiert, geerntet (5.000 rpm, 10 Min, 4 °C) und zwei mal mit eiskaltem $H_2O_{bidest.}$ gewaschen. Das Pellet wurde mit 20 ml einer 10 prozentigen Glycerin-Lösung gewaschen und in 800 μl der gleichen Lösung aufgenommen. Die Zellen wurden nun in Aliquots von 70 μl abgefüllt und bei -80 °C gelagert.

Transformation

Es wurden 0,5..2 μl Ligationsansatz mit 70 μl kompententen Zellen in eine auf Eis vorgekühlte Elektroporationsküvette gegeben. Zur Elektroporation wurde ein *Gene Pulser* (Fa. Biorad) verwendet, wobei eine Spannung von 2,5 kV, eine Kapazität von 2,5 μF und ein Widerstand von 200 Ω eingestellt wurden. Direkt nach der Elektroporation wurden der Ansatz zusammen mit 1 ml LB-Medium in ein E-cup gegeben und für eine Stunde bei 30 °C und 180 rpm inkubiert. Danach wurden die Zellen auf Selektionsmedium ausplattiert.

2.5.9. Blau-Weiß Screening (X-Gal-Test)

Zur Selektion rekombinanter Klone ist es entscheidend, nur solche auszuwählen, die sowohl den transformierten Vektor als auch ein Insert im Vektor tragen. Hierzu wurde das sogenannte „blau-weiß“ Screening verwendet. Für dieses Selektionsverfahren sind *E. coli*-Stämme notwendig, die eine Deletion im *lacZ*-Gen enthalten (*lacZ*ΔM15) und somit keine β-Galactosidase mehr bilden können. Klonierungsvektoren wie pBluescript SK^+ und seine Derivate führen durch das *lacZ'*-Gen zur Komplementation des β-Galactosidase-Defekts. Zellen, die mit einem solchen Vektor transformiert wurden, können das chromogene β-Galactosid X-Gal spalten, wobei ein Farbstoff freigesetzt wird, der durch Luftoxidation blau wird und die Zellen anfärbt. Durch Zugabe von Isopropyl-β-d-thiogalactopyranosid wird indirekt (Bindung des Lac-Repressors) die Transkription der β-Galactosidase induziert. Die bekannten Klonierungsvektoren enthalten im *lacZ* eine multiple Klonierungsstelle (MCS), so dass die Ligation des Vektors mit einem Insert dazu führt, dass das *lacZ*-Gen unter-

brochen und somit die Bildung einer funktionierenden β-Galactosidase unmöglich wird. Die Zellen bleiben also weiß, wenn ein Insert enthalten ist. Zellen ohne Vektor, bei denen die Transformation nicht erfolgreich war, werden durch eine plasmidkodierte Antibiotikaresistenz negativ selektiert.

2.6. Polymerasekettenreaktion (PCR)

Die PCR (*polymerase chain reaction*) dient zur *in vitro* Amplifikation von DNA. Dafür sind ein Paar Primer, Matrizen-DNA (*Template*) soweit eine thermostabile DNA-Polymerase notwendig. In dieser Arbeit wurde die PCR hauptsächlich zu zwei Zwecken eingesetzt; zum Schließen von Sequenzierungslücken (*gap closure*) und zur Überprüfung bereits sequenzierter Bereiche im Genom von *E. blattae*. Als Polymerasen kamen hauptsächlich zwei Enzyme zum Einsatz, die *Taq*-Polymerase (Fa. Eppendorf) sowie ein Enzymmix (*TripleMaster*, Fa. Eppendorf) unbekannten Inhalts für die Amplifikation besonders großer DNA-Fragmente bis zu einer Größe von 15 kb. Als Matrize diente die genomische DNA von *E. blattae* bzw. in Ausnahmefällen Cosmid-DNA. Als Primer wurden Oligonukleotide zwischen 17 und 18 Basen mit einer Schmelztemperatur zwischen 56 und 59 °C verwendet. Die erfolgreiche Amplifikation von DNA ist unter anderem von einer nicht zu unterschiedlichen Schmelztemperatur der Primer abhängig und wurde nach folgender Formel [13] berechnet

$$\mathrm{T_m} = 69,3 + 0,41 \cdot (\mathrm{G + C - Gehalt}) - 650/\mathrm{L}$$

G+C-Gehalt = prozentualer Gehalt an G- und C-Basen im Primer.

L = Länge des Primers in Basen.

Durchführung

Für die beiden genannten Enzyme wurden unterschiedliche Ansätze gewählt. Für die Standard-PCR für erwartete Fragmente zwischen 500 bp und 5 kb wurde der Standard- und für größere Fragmente der „long-range“-Ansatz verwendet.

Standard Ansatz		**long-range Ansatz**	
Matrizen-DNA	1 μl	Matrizen-DNA	120 ng
Vorwärts-Primer (5 pmol)	3 μl	Vorwärts-Primer (5 pmol)	3 μl
Rückwärts-Primer (5 pmol)	3 μl	Rückwärts-Primer (5 pmol)	3 μl
10 x Reaktions-Puffer	5 μl	10 x *tuning-buffer*	3 μl
dNTP Gemisch (10 mM)	1 μl	dNTP Gemisch (10 mM)	1 μl
DNA-Polymerase (2 U/μl)	0,5 μl	*TripleMaster*	0,16 μl
DMSO	2,5 μl		
$H_2O_{bidest.}$	ad 50 μl	$H_2O_{bidest.}$	ad 30 μl

Die PCR-Reaktion wurde in Thermocyclern (Fa. MWG-Biotech) direkt nach der Zugabe des jeweiligen Enzyms gestartet. Abhängig vom gewählten PCR-Ansatz kam eines der beiden folgenden PCR-Programme zum Einsatz. Diese Programme sollen hier nur als Anhaltspunkt dienen. Im Einzelfall muss das Programm optimiert werden, wobei sich eine Variation der „extension time“, der Anlagerungstemperatur und der Anzahl der Zyklen anbieten.

Standard-PCR			**long-range PCR**		
LIDTH	110 °C	ON	LIDTH	110 °C	ON
PAUSE	98 °C	FOREVER	PAUSE	96 °C	FOREVER
TEMP	98 °C	2 min	TEMP	96 °C	2 min
LOOP [	30 X		LOOP [	10 X	
TEMP	96 °C	20 Sek	TEMP	96 °C	15 Sek
TEMP	50 °C	15 Sek	TEMP	55 °C	30 Sek
TEMP	72 °C	4 min	TEMP	68 °C	300 Sek
LOOP]			LOOP]		
LIDTH	OFF		LOOP [	10 X	
TEMP	8 °C	FOREVER	TEMP	96 °C	15 Sek
END			TEMP	55 °C	30 Sek
			TEMP	68 °C	(inc.20)[a] 300 Sek
			LOOP]		
			LIDTH	OFF	
			TEMP	8 °C	FOREVER
			END		

[a] inc. bedeutet, dass bei jedem Durchlauf die Zeit um 20 Sek. verlängert wird.

2.7. Genbanken

Für die Sequenzierung mit dem „whole shotgut approach“ ist es entscheidend, mehrere Genbanken zu verwenden, die das komplette Genom, das sequenziert werden soll, über die gesamte Länge repräsentieren. Dabei sollte kein Bereich überproportional häufig vertreten sein. Um dieses zu erreichen, verwendet man Genbanken, die nach unterschiedlichen Methoden hergestellt werden. Im *E. blattae* Genomsequenzierungsprojekt wurden mehrere Genbanken verwendet, die in diesem Abschnitt beschrieben werden.

Bei der Sequenzierung einer Genbank, in der das Genom statistisch verteilt kloniert vorliegt, nimmt bei steigender Anzahl von Sequenzläufen die Anzahl der Contigs ab [48]. Wenn die Anzahl der Contigs trotz zunehmender Anzahl von Sequenzläufen nicht mehr abnimmt, wird im Normalfall die Sequenzierstrategie auf Lückenschluss umgestellt.

2.7.1. Plasmidbanken

Im Genomprojekt wurden Plasmidbanken verwendet, bei denen chromosomale DNA aus *E. blattae* partiell mit den Restriktionsenzymen *Bsp*143I bzw. *Alu*I geschnitten wurde. Außerdem kam eine Genbank zum Einsatz, die mechanisch gescherte DNA enthält. Die Insertgröße betrug in jedem Fall zwischen zwei und drei kbp. Die DNA-Fragmente wurden über einen Saccharose-Dichtegradienten aufgetrennt (s. 2.4.10), anschließend gefällt, dephosphoryliert (s. 2.5.2), zur Ligation (s. 2.5.5) eingesetzt und letztendlich in *E. coli* DH5α transformiert (s. 2.5.8). Die Klonierung erfolgte im Vektor pBluescript SK$^+$. Die gescherte DNA wurde mit Hilfe der T4-DNA-Polymerase behandelt (s. 2.5.3), um *blunt ends* zu erzeugen.

2.7.2. Cosmidbanken

Um eine Anordnung der Contigs zu erleichtern, wurde außerdem eine Cosmid-Genbank mit Inserts im Bereich von 20..50 kbp erstellt. Die chromosomale DNA wurde für diese Genbank ebenfalls partiell mit *Bsp*143I geschnitten. Die Fragmente wurden anschließend wie unter 2.7.1 beschrieben weiterbehandelt. Die Ligation erfolge in den Vektor SuperCos I. Zur Transduktion wurde der *Gigapack III XL Packaging Extract* (Fa. Stratagene) verwendet. Die Durchführung erfolgte wie im Handbuch des Kits beschrieben.

2.7.3. Cosmid-Subklonierungen

Wenn ein oder mehrere Cosmide eine Brücke zwischen zwei Contigs bildeten und keine Information über die Sequenz zwischen den Contigs vorlag, wurde eine Subklonierung eines der Cosmide durchgeführt. Zu diesem Zweck wurde das jeweilige Cosmid geschert (s. 2.4.7), um eine möglichst gleichmäßige Verteilung der zu klonierenden Fragmente zu erreichen.

Durchführung

Arbeitsschritt	Abschnitt
Scheren der Cosmid DNA	2.4.7
Auffüllen der Enden	2.5.3
5'-As anhängen	2.5.4
Dephosphorylierung der Insert-DNA	2.5.2
Klonierung mit dem TOPO-TA Kit	2.5.6
Transformation	2.5.7

Es ist relativ wahrscheinlich, dass der genomische Bereich zwischen den beiden Contigs, der durch das subklonierte Cosmid repräsentiert wird, schon in den Plasmidbanken vorhanden und sequenziert ist. Die Praxis hat gezeigt, dass man nur relativ wenige (ca. 100) Plasmide der Subklonierung sequenzieren musste, um genügend Sequenzinformation zum Schließen der untersuchten Lücke zu erhalten. Hierbei kam es meistens vor, dass kleinere Contigs, die schon in der Genomdatenbank vorhanden waren, „eingesammelt“ wurden und so die Lücke schlossen. War dieses nicht der Fall (z. B. bei einer größeren Lücke), wurden noch zusätzliche Klone der Subklonierung sequenziert, um so mehr Informationen über die Lücke zu erhalten.

2.8. Sequenzierung

Für die Sequenzierungen im *E. blattae* Genomsequenzierungsprojekt kamen drei verschiedene Sequenzierungsgeräte zum Einsatz. Die Plasmide wurden auf dem MegaBace 1000 (Fa. Amersham Bioscience), MegaBace 4000 und ABI377 (Fa. Applied Biosystems) sequenziert, die Cosmide wurden dagegen ausschließlich auf dem ABI377 sequenziert. Diese Geräte arbeiten mit dem Verfahren der Kettenabbruchmethode [84]. Bei diesem Verfahren werden floureszenzmarkierte ddNTPs (*dye ter-*

minators) eingesetzt. Die markierten ddNTPs werden statt eines dNTPs eingebaut, wodurch es zum Abbruch der Kettenverlängerung kommt. Das so entstandene Fragment ist am 5'-Ende markiert. Diese floureszenz-markierten Fragmente werden von einem Laser im Sequenzierautomaten zur Fluoreszenz angeregt. Das emittierte Licht wird in Detektoren erkannt und in Fluoreszenzpeaks umgewandelt, die dann wiederum beim „base calling“ in Sequenzen übersetzt werden.

2.8.1. Sequenzierung mit MegaBace

Die Geräte der MegaBace-Reihe sind Kapillarsequenzierungseinheiten mit 96 (MegaBace 1000) bzw. 384 (MegaBace 4000) Kapillaren. Bei diesem Typ von Sequenzierautomaten wird die elektrophoretische Auftrennung der markierten DNA-Fragmente in dünne Glaskapillaren (ca. 100 μm Durchmesser), die eine Polyacrylamid-Matrix enthalten, durchgeführt. Einer der Vorteile dieser Methode ist es, dass es nicht zu Verwechslungen der einzelnen Sequenzierungsreaktionen kommen kann, wie es bei den Plattensequenzierern z. B. ABI377 möglich ist. Außerdem wird die DNA-Probe nicht vollständig in die Kapillare gegeben, sondern es wird durch Anlegen einer Spannung nur ein Teil der DNA-Moleküle aus dem Sequenzansatz in die Matrix übertragen („injiziert“). Das hat den Vorteil, dass man dieselbe Sequenzierungsreaktion mehrmals sequenzieren kann, ohne die Reaktion neu ansetzen zu müssen, da beim Injizieren der DNA kein Probenvolumen verbraucht wird. Nachteilig kann sich allerdings die Empfindlichkeit der Auftragungsmethode (Injektion) der DNA auswirken. Setzt man zu hohe Stromstärken ein, wählt die Injektionszeit zu hoch oder verwendet im Reaktionsansatz zu viel Template-DNA, dann können die Kapillaren regelrecht „verstopfen“. Der folgende Sequenzansatz wurde pipettiert, wobei die DNA vorher gereinigt vorlag.

10 μl Sequenzier-Ansatz	
DNA	400 ng
Primer (5 pmol)	1 μl
DYEnamic ET Dye Terminator Kit	4 μl
H_2O_{HPLC} ad 10 μl	

Anschließend wird das „MegaBace cycle sequencing“-Programm in den entsprechenden Thermocyclern ausgeführt. Zur Aufreinigung der Sequenzansätze wurden zwei Fällungen mit 24 μl Isopropanol (80 % v/v) und 40 μl Ethanol (80 % v/v) durchgeführt.

MegaBace cycle sequencing		
LIDHT	110 °C	ON
PAUSE	95 °C	FOREVER
TEMP	95 °C	5 min
LOOP [	30 X	
TEMP	95 °C	20 Sek
TEMP	50 °C	15 Sek
TEMP	60 °C	1 min
LOOP]		
LIDHT	OFF	
TEMP	4 °C	FOREVER
END		

2.8.2. Sequenzierung mit ABI377

Bei den ABI377-Geräten handelt es sich um Plattensequenzierer. Hier wird zwischen zwei Glasplatten ein Polyacrylamid-Gel gegossen, in dem die Elektrophorese stattfindet. 96 Proben werden manuell mit Hilfe einer Mehrkanal-Hamilton-Pipette auf das Gel aufgetragen. Der Vorteil dieser Geräte ist, dass sie relativ tolerant gegenüber der aufgetragenen DNA-Menge sind. So ist es auch möglich, bei schwachen Signalen (z. B. bei der Cosmid Sequenzierung) noch relativ gute Sequenzen zu erhalten. Als Nachteile sind der hohe Arbeitsaufwand beim Gießen der Gele und manuellen Auftragen der Proben zu nennen, sowie die Möglichkeit, dass es durch Ausfall einzelner Spuren oder Überstrahlen einzelner Spuren (bei zu viel DNA im Ansatz) zu Verwechslungen einzelner Sequenzen kommen kann.

Durchführung

Für die Sequenzierung am ABI wurden die Gele wie in Tabelle 2.13 beschrieben angesetzt und die Elektrophorese durchgeführt. Die Zusammensetzung eines ABI-Sequenzieransatzes für Plasmide ist in Tabelle 2.11 zu sehen. Den entsprechenden Ansatz für die Sequenzierung von Cosmiden zeigt Tabelle 2.12. Die jeweiligen Cycling-Programme sind in Tabelle 2.14 sowie in Tabelle 2.15 dargestellt. Nach dem Cycle-Schritt erfolgte eine Fällung mit Ethanol (99 % v/v) und 3 M Natriumacetat (pH 4,6), wodurch überschüssige Dye-Terminatoren und Primer entfernt wurden. Vor dem Auftragen der Proben auf das Sequenziergel wurden diese mit 1 μl Stop-Mix

(s. Tabelle 2.10) versetzt.

Tabelle 2.9.: *loading-buffer* für ABI-Sequenzansätze

Bestandteil	Menge
EDTA	25 mM
Blue Dextran	50 mg/ml
$H_2O_{bidest.}$	ad 500 ml

Tabelle 2.10.: Stop-Mix für ABI-Sequenzansätze

Bestandteil	Menge
Formamid	5 Teile
loading buffer	1 Teil

Tabelle 2.11.: „1/2-half BD-Ansatz“ ABI-Plasmid-Sequenzieransatz

	Menge
Premix	2 μl
half BD	2 μl
Betain (5 M)	1 μl
Primer (5 pmol/μl)	1 μl
DNA Template	400 ng
$H_2O_{bidest.}$	ad 20 μl

2.8.3. Sequenzabbrüche

Es kann vorkommen, dass sich bestimmte Bereiche eines Genoms schlecht sequenzieren lassen. In diesen Regionen bricht die Sequenzreaktion ab, ohne dass die maximale Leselänge erreicht worden ist. Solche Abbrüche sind in den meisten Fällen unabhängig von der Template-DNA (Plasmid, Cosmid oder PCR-Produkt). Ursachen für Probleme an diesen Stellen können sowohl ein sehr hoher lokaler G+C-Gehalt als auch repetitive Sequenzen sein, die unter Umständen komplexe Sekundärstrukturen in der DNA hervorrufen. Für diese Bereiche wurde ein spezieller Sequenzieransatz verwendet, der die Leseweise meist nur wenig, aber dennoch ausreichend, verlängert hat. Der Ansatz und das Cycleprogramm sind in Tabelle 2.16 bzw. Tabelle 2.17 gezeigt.

Tabelle 2.12.: ABI-Cosmid-Sequenzieransatz

	Menge
Big Dye Mix	4 μl
Primer	10..30 pmol
DNA Template	500 ng
$H_2O_{bidest.}$	ad 10 μl

Tabelle 2.13.: ABI-Sequenziergel

Zusammensetzung eines ABI-Sequenziergels	
Harnstoff 7 M	18 g
Acrylamid (30 %, 29:1)	5,3 ml
10 X TBE-Puffer$_{long\ run}$	6 ml
$H_2O_{bidest.}$	21,5 ml
10 % APS	250 μl
TEMED	25 μl
Laufparameter	
Geldicke	0,2 mm
Gellänge	48 cm
Spannung	3.500 V
Stromstärke	60 mA
Leistung	40 W
Temperatur	45 °C
CCD offset	0
CCD gain	4
Vorlaufzeit	90 Sek
Laufzeit	1,5 Stunden + 14,5 Stunden
Laufpuffer	2 x TTE oben, 1 x TTE unten
Minimale Anzahl von Scans	22.000

Tabelle 2.14.: ABI-Plasmid-Cycleprogramm

ABI cycle sequencing		
LIDHT	110 °C	ON
PAUSE	98 °C	FOREVER
TEMP	98 °C	2 min
LOOP [	25 X	
TEMP	96 °C	20 Sek
TEMP	50 °C	15 Sek
TEMP	60 °C	4 min
LOOP]		
LIDHT	OFF	
TEMP	4 °C	FOREVER
END		

Tabelle 2.15.: ABI-Cosmid-Cycleprogramm

forward			reverse		
LIDTH	110 °C	ON	LIDTH	110 °C	ON
TEMP	95 °C	5 min	TEMP	95 °C	5 Min
LOOP [	60 X		LOOP [	60 X	
RAMP	95 °C	3 °C/Sek	RAMP	95 °C	3 °C/Sek
TEMP	95 °C	30 Sek	RAMP	95 °C	30 Sek
RAMP	55 °C	3 °C/Sek	RAMP	51 °C	3 °C/Sek
TEMP	55 °C	20 Sek	TEMP	51 °C	20 Sek
RAMP	60 °C	3 °C/Sek	RAMP	60 °C	3 °C/Sek
TEMP	60 °C	4 min	TEMP	60 °C	4 min
LOOP]			LOOP]		
LIDHT	OFF		LIDHT	OFF	
RAMP	4 °C	3 °C/Sek	RAMP	4 °C	3 °C/Sek
END			END		

Tabelle 2.16.: Sequenzieransatz für „Abbruch"-Regionen

Stoff	Menge
Polyoxyethylensorbitan monolaurate (1 % v/v)	1 μl
IGPAL (1 % v/v)	1 μl
Big Dye Mix (V3.0)	1 μl
Primer	200 pmol
DNA Template (100 ng / μl)	4 μl
$H_2O_{bidest.}$	ad 9 μl

Tabelle 2.17.: ABI-Plasmid-Cycleprogramm für „Abbruch"-Regionen

ABI cycle sequencing		
LIDHT	110 °C	ON
PAUSE	95 °C	FOREVER
TEMP	95 °C	5 min
LOOP [	60..99 X	
TEMP	95 °C	30 Sek
TEMP	50 °C	20 Sek
TEMP	60 °C	4 min
LOOP]		
LIDHT	OFF	
TEMP	4 °C	FOREVER
END		

2.9. Sequenzierungs-Software

Die Verarbeitung der Daten aus den Sequenzierungseinheiten erfolgt zu großen Teilen vollautomatisch. Hier soll nicht ausführlich auf die den Programmen zugrundeliegende Theorie eingegangen werden, sondern nur ein allgemeiner Überblick über die Abläufe gegeben werden. Die Verarbeitung erfolgt in folgenden Schritten, nachdem die Sequenzdateien die Sequenzierungseinheiten verlassen haben:

- *Basecalling* mit `phred` [24]: Zuweisen von Qualitätsparametern zu jeder Base (Fehlerwahrscheinlichkeit).
- *Vector-screening* und Umwandlung der Sequenzdateien mit `pregap` [8]: Maskieren von Sequenzen, die dem Sequenzierungsvektor zuzuordnen sind.
- *Assembly*, d. h. Zusammensetzen der Contigs durch Finden von überlappenden Sequenzen mit `gcphrap` [8].
- Erstellen einer Datenbank zur Bearbeitung der Sequenzen in `GAP4` [8].

2.10. Lückenschluss - *gap closure*

Nach dem Ende der Rohsequenzierungsphase des *E. blattae* Genomsequenzierungsprojekts bestand die Sequenz aus 220 *contigs*. Dabei handelt es sich um zusammenhängende Sequenzabschnitte, die durch das computergestützte Zusammenlagern (*assembly*) viele kleiner Sequenzstücke aus der Ansequenzierung von Plasmiden und Cosmiden entstehen. Zwischen diesen Stücken befinden sich die Lücken, die es in der *gap closure*-Phase zu schließen gilt. In diesem Abschnitt sollen verschiedene Methoden vorgestellt werden, mit deren Hilfe man die verbleibenden Lücken in einer Genomsequenz schließen kann.

2.10.1. „Plasmid walking“

Bei einer durchschnittlichen Insertgröße von 2..3 kb in den Plasmidgenbanken von Genomprojekten (s. 2.7.1) kann mit der Ansequenzierung der Inserts von beiden Seiten nicht die komplette Insertsequenz erfasst werden. Das führt dazu, dass Contigenden dadurch entstehen, dass die Sequenzinformation über die Inserts abbricht, weil nicht weiter sequenziert wurde. In GAP4-Programm wird ein solcher Fall wie in Abbildung 2.2 dargestellt.

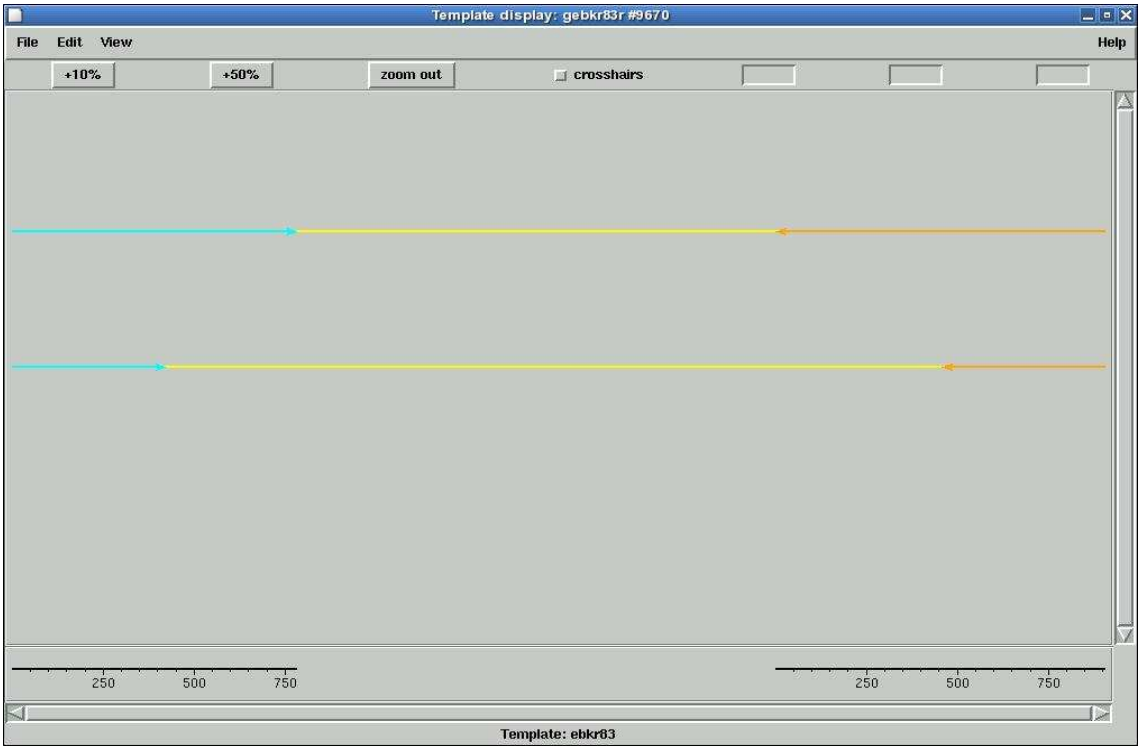

Abbildung 2.2.: Darstellung zweier, durch Plasmidbrücken (gelb) verbunden, Contigs (schwarz).

Beim *plasmid walking* wurde an den Contigenden jeweils ein Primer abgeleitet und mit diesem Primer auf dem jeweiligen Plasmid in die Lücke hineinsequenziert. Dabei erreicht man die jeweils schon bekannte Sequenz am anderen Ende der Lücke. Diese Methode hat dann Grenzen, wenn keine Lücken überspannenden Plasmide mehr vorhanden sind. Zu diesem Zeitpunkt wird in einem Sequenzierungsprojekt ein Stratgiewechsel vom *Plasmid walking* zu den PCR-basierten Methoden vollzogen.

2.10.2. Cosmid-Brücken zum Schließen von Lücken

Das Vorhandensein einer Cosmidgenbank ist in Genomsequenzierungsprojekten sehr vorteilhaft. Sie ermöglichen es, die Reihenfolge von Contigs zu bestimmen (d. h. man bildet s. g. *supercontigs*). Außerdem erleichtern sie das Auflösen von *repeats*, die teilweise größer als die durchschnittliche Insertgröße von Plasmiden sind. Cosmide werden wie Plasmide von den Enden her ansequenziert (s. 2.8.2) und enthalten Inserts in der Größe von ca. 20..50 kb. Cosmide, die Lücken überspannen, werden in GAP4 ähnlich wie Plasmidbrücken dargestellt (s. Abbildung 2.2), nur die Insertgröße wird in der Darstellung von 2..4 kb auf 20..50 kb angepasst. Nachdem man

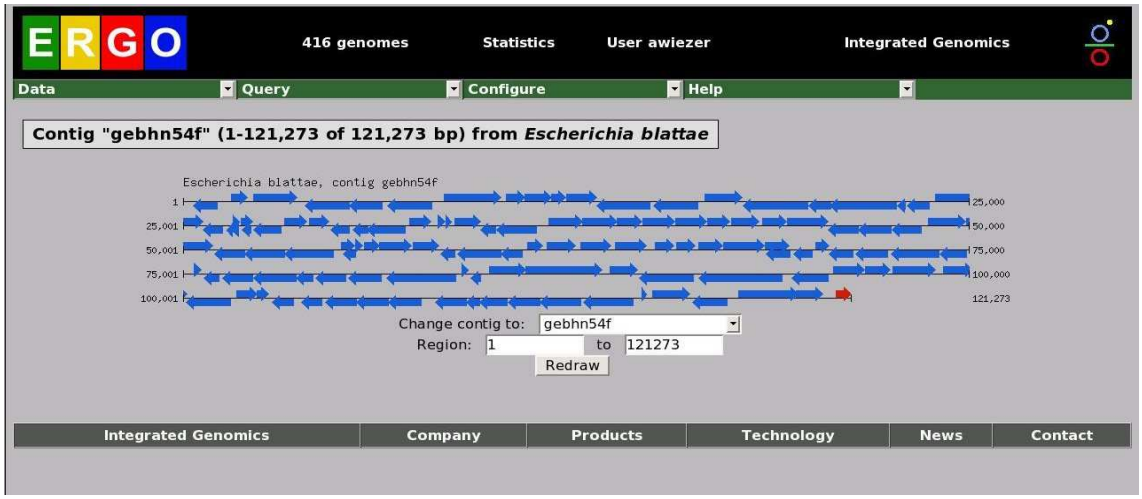

Abbildung 2.3.: Darstellung der Gene auf einem Contig in ERGO.

supercontigs gebildet hat, ist es mittels (*long range*)PCR möglich, den Bereich der Lücke mit chromosomaler DNA als Template zu amplifizieren. Das PCR Produkt kann nach der Aufreinigung direkt zur Sequenzierung eingesetzt werden. Die Primer für die PCR werden dabei an den Contigenden abgeleitet und ebenfalls für die Sequenzierung des PCR-Produkts verwendet.

2.10.3. ERGO zum Schließen von Lücken

Zusätzlich zu den bereits erwähnten Methoden werden oft s. g. *peptide links* benutzt, um Lücken zu schließen, über die man keine Information durch Plasmid- oder Cosmidbrücken erhalten kann. Dabei braucht man mindestens einen, bereits vollständig sequenzierten, Vergleichsorganismus. Je näher der zu sequenzierende Organismus mit dem Vergleichsorganismus verwandt ist (d. h. je ähnlicher sich die abgeleiteten Proteinsequenzen sind), desto größer ist die Wahrscheinlichkeit, dass man mit *peptide links* arbeiten kann. Doch eine hohe Proteinähnlichkeit als einziger Faktor reicht nicht aus. Wichtiger ist eine hohe Synthenie (konservierte Anordnung großer DNA-Bereiche) zwischen den zu vergleichenden Organismen. In dieser Arbeit wurde das ERGO Programmpaket [74] verwendet, um mit *peptide links* zu arbeiten. Das Prinzip beruht darauf, dass man versucht, über die konservierte Anordnung der Gene das Ende eines Contigs an einen Bereich zu alignieren, der im Vergleichsorganismus schon bekannt ist. Dieses Vorgehen wird in ERGO unter der Funktion *contig regions* zusammengefasst. Abbildung 2.3 zeigt ein solches Contigende.

Die Gene sind in blau und das aktuell ausgewählte Gen in rot dargestellt. Nachdem

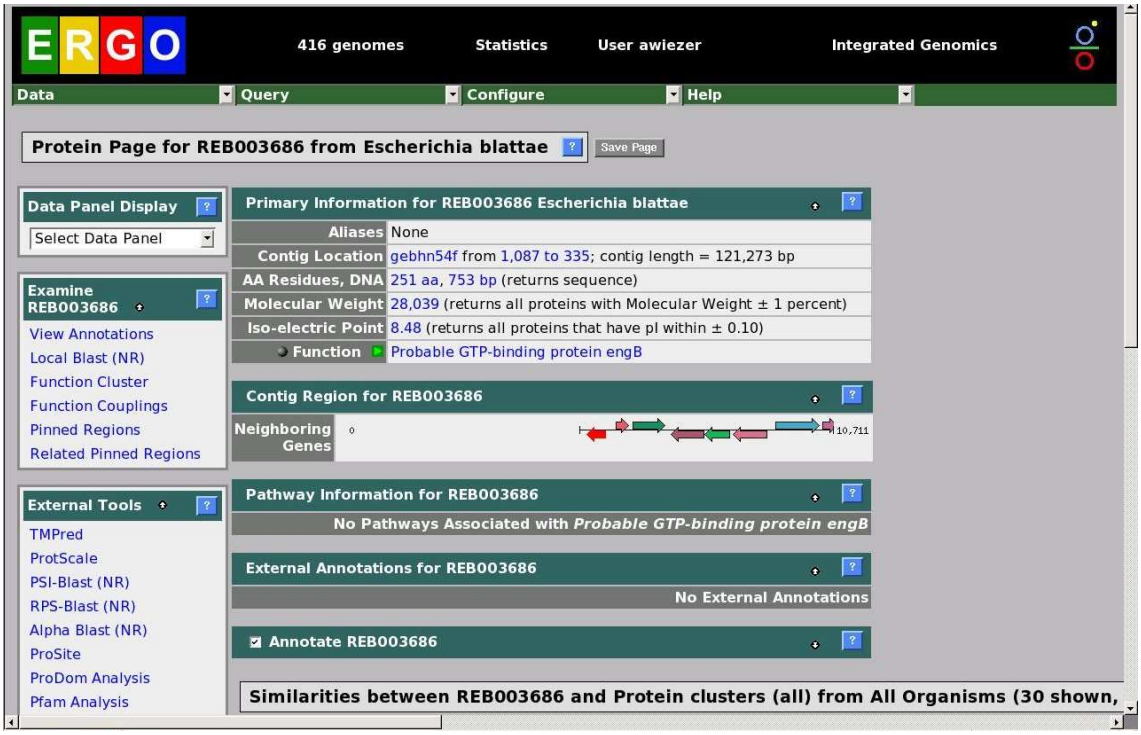

Abbildung 2.4.: Darstellung des in Abbildung 2.3 ausgewählten Proteins in ERGO.

man sich für ein Gen an einem Contigende entschieden hat, wählt man das Gen aus und gelangt zu der entsprechenden Proteinseite (s. Abbildung 2.4).

Hier erhält man eine Übersicht über die nächsten homologen Proteine aus anderen Organismen (die Vergleichsorganismen, s. Abbildung 2.5). Dann wählt man Proteine aus den Organismen aus, die eine hohe Ähnlichkeit aufweisen (d. h. oben in der Liste stehen) und von denen man annimmt, dass die Anordnung der Gene konserviert ist (s. Abbildung 2.5, Haken in den weißen Kästen).

Anschließend wählt man die *contig regions*-Funktion an und erhält eine Darstellung der betreffenden Region der ausgewählten Organismen (s. Abbildung 2.6). Hier werden homologe Proteine in derselben Farbe dargestellt. Man erkennt im Beispiel eine klare Konservierung der Genanordnung und das Ende des Contigs in *E. blattae*. Um nun den Contig in *E. blattae* zu identifizieren, der die blau bzw. grün dargestellten Gene aus den Vergleichsorganismen enthält, wählt man diese Proteine aus und sucht in den entsprechenden Listen homologer Proteine nach Einträgen aus *E. blattae*. Danach identifiziert man die Contigs, auf denen die Proteine liegen (höchstwahrscheinlich am Contigende, wenn es sich nicht um paraloge Gene handelt) und leitet zum Schluss an den beiden Contigenden aus *E. blattae* Primer für

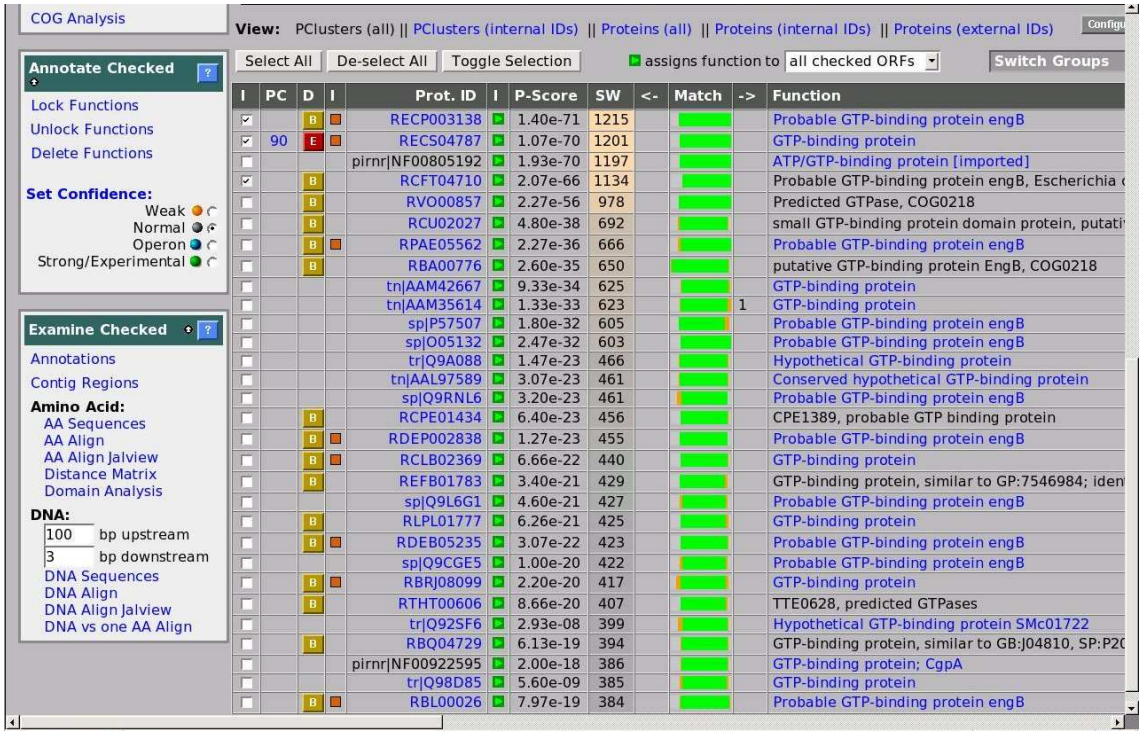

Abbildung 2.5.: Liste homologer Proteine zum Protein aus Abbildung 2.4.

eine PCR ab.

2.10.4. Lückenschluss mit BLAST

Es ist nicht immer möglich, Contigs durch die *peptide link*-Methode zu arrangieren, da bestimmte Regionen keine ausreichende Homologie auf Proteinebene aufweisen. In diesen Fällen bietet es sich an, BLAST [1] zu verwenden. Das Prinzip beruht in diesem Fall darauf, von der DNA-Sequenz eines Vergleichsorganismus eine BLAST-Datenbank zu erstellen und die DNA-Sequenz eines Contigendes gegen diese Datenbank zu blasten. Der Vorteil dabei ist, dass diese Methode deutliche weniger stringent ist als *peptide links* und somit auch schwach konservierte Regionen im Vergleichsorganismus wiederfindet. Der Nachteil ist, dass man einiges mehr an Handarbeit investieren muss, um einem Treffer im Vergleichsorganismus das entsprechende andere Contigende der Lücke zuzuordnen. Man kann sich diese Arbeit durch Programme erleichtern, allerdings ist auch dabei noch viel „Augenmaß“ erforderlich, um falsch positive Ergebnisse durch repetitive Sequenzen zu eliminieren.

Dieses Verfahren wurde in der vorliegenden Arbeit selten verwendet. Die wenigen PCRs, die abgeleitet wurden, haben mit ca. 50 % zum Erfolg geführt.

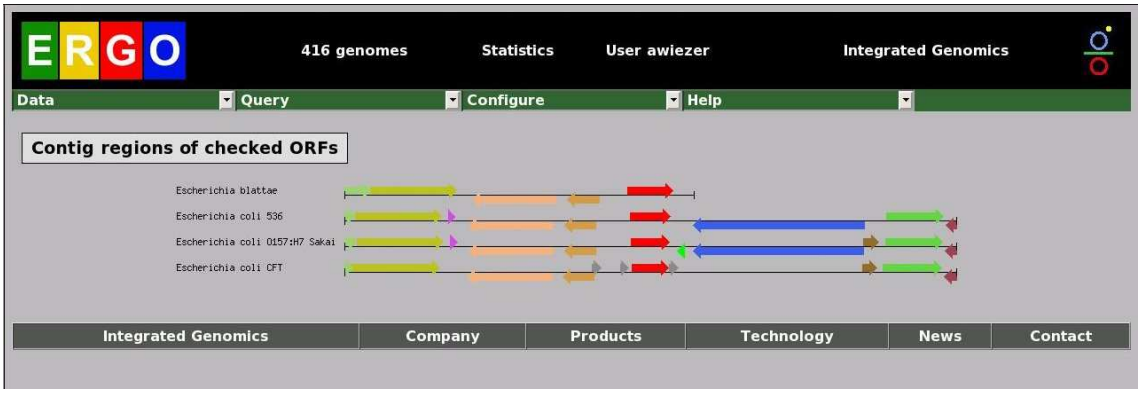

Abbildung 2.6.: Konservierte Genanordnung in ERGO. Von oben nach unten: *E. blattae*, *E. coli* 536, *E. coli* O157:H7, *E. coli* CFT

2.10.5. Multiplex-PCR

Im *E. blattae* Genomsequenzierungsprojekt lagen keinerlei Information über die verbleibenden Lücken mehr vor, nachdem die Lückenschlussmethoden (s. 2.10.3 und 2.10.4) angewendet wurden. Das lag unter anderem daran, dass bestimmte genomische Bereiche in den Genbanken nicht kloniert wurden, da diese Bereiche für *E. coli* toxisch sind.

In diesem Stadium bietet sich die Multiplex-PCR an. Bei diesem Verfahren werden Primer (ca. 22 Mere) an allen Contigenden abgeleitet. Diese Primer werden dann zu *pools* zusammengefasst und in eine PCR-Reaktion eingesetzt. Somit erreicht man, dass alle Primer untereinander als mögliche Paare getestet werden und verringert den Arbeitsaufwand erheblich, da man nicht alle möglichen Primerkombinationen einzeln ansetzen muss.

Durchführung

Das hier beschriebene Verfahren ist eine Abwandlung aus Tettelin et al. [92]. Im *E. blattae* Genomsequenzierungsprojekt wurden 90 Primer an den Contigenden abgeleitet. Jeweils 5 diese Primer wurden zu einem pool zusammengefasst. Der folgende Ansatz wurde verwendet

20 μl Multiplex-PCR-Ansatz	
Primer-Pool 1	1 μl
Primer-Pool 2	1 μl
Tuning Buffer	2 μl
dNTPs	1 μl
TripleMaster	0,16 μl
DNA	120 ng
$H_2O_{bidest.}$	ad 1000 ml

Aus 90 Primern und 17 Pools (A..R) ergaben sich 128 PCRs ($16 \cdot 16/2 = 128$). Wenn eine Pool-Kombination ein oder mehrere Produkte ergab, so versucht man durch ein Ausschlussverfahren die beiden Primer zu ermitteln, die ein bestimmtes Produkt ergeben haben. Im *E. blattae* Genomprojekt konnten mit dieser Methode PCR-Produkte 28 Lücken zugeordnet werden.

2.11. BLAST

BLAST [1] wird in der Bioinformatik häufig als Werkzeug zum Vergleichen von Sequenzen unter anderem wegen seiner Geschwindigkeit und Sensitivität verwendet. In dieser Arbeit wurde BLAST eingesetzt, um Sequenzen nach Vektorkontaminationen zu durchsuchen (s. 2.9), um überlappende Contigs zu finden und um ganze Genome miteinader zu vergleichen (s. 3.2.2 sowie 3.3.2). Außerdem wurde mit secureBLAST [97] ein Werkzeug geschaffen, um den kontrollierten Zugang zur Sequenzsuche zu erleichtern (s. 3.4.3). In diesem Abschnitt soll BLAST als Werkzeug zum Vergleich kompletter Genome vorgestellt werden.

2.11.1. Bi-direktionales BLAST

Beim Vergleich ganzer Genome wird das Programm `blastp` verwendet, was bei genauer Betrachtung nicht Genome sondern Proteome vergleicht. Man verwendet deshalb `blastp`, weil so eine Aussage über das Vorhandensein von Proteinen gemacht werden kann. Ein Vergleich auf DNA-Niveau würde, bedingt durch das BLAST-Prinzip des „local alignment“, viele kurze Treffer liefern und sich nicht auf die, in diesem Zusammenhang, wesentlichen Bestandteile konzentrieren.

Das Vorgehen beim Vergleich ganzer Genome (hier A und B) besteht aus folgenden

Schritten:

- Erstellen zweier *multiple fasta files* der Proteinsequenzen der Genome
- Erstellen zweier BLAST-Datenbanken:
  `formatdb -i a.fas -p T -o T`
  `formatdb -i b.fas -p T -o T`
- BLAST der Genome gegeneinander:
  `blastall -p blastp -i a.fas -d b.fas -e 1.0e-5`
  `blastall -p blastp -i b.fas -d a.fas -e 1.0e-5`
- Skriptgesteuerte Auswertung der Treffer (s. Anhang A.1)

Das Skript liest dabei zuerst die Ausgaben beider `blastall`-Aufrufe ein und baut eine Datenstruktur auf, die Angaben darüber enthält, welches Protein (aus A) (*query*) den besten Treffer (*hit*) zu welchem anderen Protein (aus B) hat und umgekehrt. Falls kein Treffer gefunden wurde, wird *NOHIT* vermerkt. Dabei wird keine Rücksicht auf die Qualität der Treffer genommen, weil schon bei Aufruf von BLAST ein Schwellenwert des Erwartungswertes e (hier $1 \cdot 10^{-5}$) definiert wurde. In einem zweiten Schritt wird die Datenstruktur auf solche Proteine durchsucht, die sowohl in die eine Richtung A→B als auch in die andere Richtung B→A denselben Treffer liefern (bi-direktionaler Treffer). Mit diesem Verfahren schließt man den falsch positiven Treffer zu paralogen Proteinen aus. Das Ergebnis dieser Untersuchung ist eine Liste orthologer Proteine.

2.12. Annotation

Nachdem ein Großteil der *E. blattae* Sequenz bekannt war, wurde diese Sequenz dem Annotationsprozess unterzogen. Unter Annotation versteht man das Finden aller Gene. Soweit möglich soll den abgeleiteten Proteinen dabei eine Funktion zugewiesen werden. Das Zuweisen der Funktion beruht auf Sequenzähnlichkeit und folgt dem Paradigma, dass ab einer Ähnlichkeit von 30 % mit hoher statistischer Sicherheit auf übereinstimmende räumliche Strukturen für Proteine geschlossen werden kann [83].

Für die Annotation im Genomprojekt *E. blattae* wurden zwei Systeme verwendet. Zum einen Quann (s. 3.4.1) für eine vollautomatische Annotation in einem frühen

Stadium (ca. 100 Lücken), um einen schnellen Überblick über die vorhandenen Proteine zu erhalten, und zum anderen ERGO für eine automatische und eine detailierte manuelle Annotation.

2.12.1. ORF Vorhersage

Bevor mit der Annotation begonnen werden konnte, mussten zuerst alle potentiellen offenen Leserahmen (ORFs, *open reading frames*) identifiziert werden. Diese Aufgabe wurde durch *ORF-finder* bewerkstelligt, Programme, die Sequenzen im einfachsten Fall nach Stop-Codons (TAA, TGA, TAG) durchsuchen und damit zuerst das Ende eines ORFs festlegen. Im nächsten Schritt wird dann im selben Leserahmen in 5'-Richtung nach einem Start-Codon gesucht. Erreicht dieser ORF eine vorher festgelegte Mindestlänge, werden die Koordination (Start und Stop) des Gens für die weitere Verarbeitung gespeichert. Für diese bioinformatische Aufgabenstellung gibt es mehrere Programme, die alle bestimmte Vor- und Nachteile mit sich bringen. Für eine optimale ORF-Vorhersage ist also eine Kombination mehrere Programme und Methoden vorteilhaft und wünschenswert. Das Programm YACOP [91] erfüllt diesen Anspruch und wurde im vorliegenden Projekt verwendet.

2.12.2. ERGO

Für eine abschließende Annotation des *E. blattae* Proteoms wurde das ERGO Programmpaket verwendet. Nach einer automatisch durchgeführten Annotation wurde diese für jedes Protein zusätzlich manuell validiert.

Die automatische Annotation durch ERGO beinhaltet die Übersetzung der, von YACOP in Form von Start- und Stoppositionen vorhergesagten, Gene in Proteinsequenzen und deren Abgleich mittels FASTA [75] gegen eine nicht-redundante Proteindatenbank bestehend aus:

- Swiss-Prot (`uniprot_sprot.*.gz`)
- TrEMBL (`uniprot_trembl.*.gz`
- TrEMBL new (`new/uniprot_trembl_new.*.gz`)

Die Dateien befinden sich unter:
`ftp.expasy.org/databases/uniprot/knowledgebase/`. Um die manuelle Annotation zu erleichtern, wurde außerdem eine BLAST-Suche gegen die jeweils aktuelle

Proteindatenbank (`ftp.ncbi.nih.gov/blast/db/nr.tar.gz`) des NCBI durchgeführt und das Ergebnis auf der jeweiligen Protein-Seite in ERGO abgebildet (s. Abbildung 2.4). Die Suche nach tRNAs wird nicht von ERGO selbest sondern von tRNAscan [54] durchgeführt. Während der manuellen Annotation kommen zusätzliche, externe Werkzeuge zum Einsatz, die teilweise sehr hilfreich sind, wenn Annotationen allein nach Treffern in den Datenbanken nicht möglich sind. Bei den Werkzeugen handelt es sich um:

- TMPred zur Vorhersage membranspannender Regionen (`www.ch.embnet.org/software/TMPRED_form.html`)
- Pfam zur Einordnung von Proteinen in Proteinfamilien [4] (`www.sanger.ac.uk/Software/Pfam/`)
- COG zur Einordnung von Proteinen in *cluster of orthologous groups* [90] (`www.ncbi.nlm.nih.gov/COG/`)

3. Ergebnisse

In diesem Kapitel werden die Ergebnisse der bioinformatischen Auswertung dreier verschiedener Genome dargelegt. Begonnen wird mit der Sequenzierung und Auswertung des Enterobakteriums *Escherichia blattae*. Danach wird auf die Analyse der Genomsequenzen von *Methanosarcina mazei* Gö1 und *Thermus thermophilus* HB27 eingegangen. Diese beiden Organismen werden außerdem mit ihren nächsten Verwandten verglichen.

3.1. Genomprojekt *Escherichia blattae*

Ein Teil der vorliegende Arbeit befasst sich mit der Genomsequenzierung von *Escherichia blattae* [12]. Bei diesem Projekt wurden bis heute 4,11 Mbp sequenziert, was bei einer geschätzten Genomgröße von 4,3 Mbp 95 % der Gesamtsequenz entspricht. Zum jetzigen Zeitpunkt sind noch 43 Bereiche im Genom unbekannt, die einen Teil der fehlenden 5 % Sequenz enthalten. Der andere Teil der unbekannten Sequenz ist partiell in kleineren Contigs zu finden, deren Anordnung in der Gesamtsequenz bisher nicht bekannt ist. In den folgenden Abschnitten sollen die aktuellen Erkenntnisse aus dem Genomprojekt beschrieben werden. Weiterhin wird *E. blattae* mit anderen Bakterien aus der Familie der *Enterobacteriaceae* verglichen und auf mögliche Fremdgene hin untersucht.

3.1.1. Der Organismus

Escherichia blattae zählt zur Familie der *Enterobacteriaceae* und wurde aus dem Enddarm einer Küchenschabe isoliert [12]. Der Organismus ist als nicht pathogen eingestuft (S1-Organismus, s. Seite 421 [56]), eine Voraussetzung für die Verwendung in der biotechnologischen Industrie. Als Labor- bzw. Produktionsstamm für biotechnologische Anwendungen eignet er sich deshalb, weil alle bekannten genetischen Werkzeuge wie z. B. Expressionsvektoren, die für *E. coli* hergestellt wurden, auch in *E. blattae* funktionieren. Des Weiteren wächst *E. blattae* aerob bei 37 °C und

kann somit relativ kostengünstig im Großmaßstab angezogen werden. Die möglichen biotechnologischen Anwendungen waren auch ein Grund für die Sequenzierung des Genoms von *E. blattae*, denn die Entschlüsselung des Genoms würde den Grundstein für den gezielten Einsatz des Organismus als Produktionsstamm legen.

3.1.2. Sequenzierung

Bei der Sequenzierung des *E. blattae* Genoms wurde mit der *whole-shotgun* Methode gearbeitet. Bei diesem Verfahren wird die komplette genomische DNA eines Organismus in Fragmente von ca. 2 bis 4 kbp zerkleinert (geschert oder partiell geschnitten, s. 2.7) und in Plasmidvektoren kloniert. Im vorliegenden Genomsequenzierungsprojekt wurde die chromosomale DNA partiell mit dem Restriktionsenzym *Sau*3A geschnitten und in den Plasmidvektor pBluescript SK^+ kloniert. Für eine weitere Genbank wurde chromosomale DNA mechanisch geschert, die Enden wurden zu *blunt ends* aufgefüllt (s. 2.5.3) und ebenfalls in den Plasmidvektor pBluescript SK^+ kloniert. Die Transformation der Plasmide erfolgte durch Elektroporation in *E. coli* DH5α (s. 2.5.8). Nach Isolierung der rekombinanten Plasmide wurden diese von den Insert-Enden her sequenziert. Die daraus resultierenden Sequenzen wurden mit Hilfe von bioinformatischen Methoden anhand von Sequenzübereinstimmungen zu s. g. *Contigs* zusammengesetzt. Die dabei verwendete Software und die einzelnen Schritte sind unter 2.9 beschrieben. In diesem Projekt kam außer den Plasmidgenbanken (s. 2.7.2) noch eine Cosmid-Genbank zum Einsatz, damit die Contigs besser arrangiert werden können.

Das Genomsequenzierungsprojekt *E. blattae* kann in die drei Phasen Rohsequenzierung, Lückenschluss und Qualitätsverbesserung unterteilt werden, die im Folgenden näher beschrieben werden sollen.

3.1.3. Rohsequenzierung

In der Phase der Rohsequenzierung wurden die rekombinanten Plasmide aus den Genbanken von beiden Seiten her mit Vektor-spezifischen Primern ansequenziert, d. h. es wurde nicht das komplette Insert sequenziert, sondern nur in das Insert des Plasmids hineingelesen. Die resultierenden Sequenzen wurden gesammelt und assembliert, d. h. anhand von Sequenzübereinstimmungen mit bioinformatischen Methoden zu größeren *Contigs* zusammengesetzt (s. 2.9).

„Zwischenassemblierungen“ zeigten die jeweils aktuelle Anzahl von Contigs während der Rohsequenzierung und konnten dazu verwendet werden zu kontrollieren, ob die Sequenzierung dem Verlauf der von LANDER und WATERMAN [48, 95] vorhergesagten Rohsequenzierungsphase folgt. Für eine Abschätzung, der für die Rohsequenzierung benötigten Sequenzläufe, wird folgende Formel verwendet

$$N = \frac{A \cdot L}{\omega}$$

Dabei ist N die Anzahl der einzelnen Sequenzläufe, ω die durchschnittliche Sequenzlänge, A die angestrebte Abdeckung und L die Größe des Genoms. Unter Abdeckung versteht man in diesem Fall die durchschnittliche Anzahl von Sequenzläufen, durch die eine Base repräsentiert wird. Daraus ergaben sich für *E. blattae* bei ca. 4,3 Mbp Genomgröße und einer achtfachen Abdeckung bei einer Sequenzlänge von ca. 600 Basen etwa 57.000 Rohsequenzen.

LANDER und WATERMAN [48, 95] haben zur Kontrolle der Rohsequenzierungsphase die folgende Formel vorgeschlagen

$$P = e^{\frac{-N \cdot \omega}{L}}$$

Mit dieser Formel soll die Anzahl von Contigs vorhergesagt werden, die unter idealen Bedingungen (konstante Sequenzlänge und Genbanken mit statistischer Verteilung der Inserts) einer gegebenen Anzahl von Sequenzen entspricht. Dabei ist P die Wahrscheinlichkeit dafür, dass eine beliebige Base bei einer gegebenen Anzahl von Sequenzen noch nicht sequenziert wurde. Beispielsweise ist $P = 0,03\,\%$ bei einer Sequenzlänge ω von 600 Basen, einer Anzahl N von 57.000 Sequenzen und einer Genomgröße von 4,3 Mbp. Mit diesem Wert kann dann errechnet ($= L \cdot P$) werden, dass in diesem Beispiel 141.900 Basen noch nicht sequenziert sind. Bei einer achtfachen Abdeckung ($A = N \cdot \frac{\omega}{L}$) ergibt sich eine Zahl von 1.749 ($= N \cdot P$) Contigs. Trägt man diese theoretische Contiganzahl in einer Grafik und parallel dazu die aus den Assemblierungen gewonnenen Zahlen der Contigs gegen die Anzahl der Sequenzläufe auf, kann man die tatsächliche Anzahl von Lücken (entspricht der Anzahl von Contigs) mit der theoretisch erwarteten Anzahl vergleichen. Wenn man während der Rohsequenzierung den Verlauf der Kurve verfolgt, so kann man schnell erkennen, ob die Abschätzung der Genomgröße korrekt war. Ist dieses nicht der Fall, nimmt die Anzahl der Lücken mit steigender Anzahl von Sequenzen nicht stark ge-

nug, d. h. der theoretischen Vorhersage entsprechend, ab. Wenn die Qualität, d. h. die statistische Verteilung und die Insertgröße der Genbanken nicht gut genug ist, dann kann es ebenfalls zu Abweichungen von der Idealline kommen. Sollte ein bestimmter Bereich des Genoms in den Genbanken unterrepräsentiert sein, könnten auch weitere Sequenzen keine Information über diese Bereich liefern, wenn er nicht kloniert ist. Ist ein Bereich überrepräsentiert, so liefern mehr neue Sequenzen immer dieselbe Information über schon sequenzierte Bereiche und tragen so nicht zu einer abnehmenden Anzahl von Lücken bei. Die Rohsequenzierung im *E. blattae* Genomsequenzierungsprojekt wurde nach 51.486 Sequenzen, die zu 220 Contigs assembliert werden konnten, beendet. Im Anschluss wurden ungefähr 700 Cosmide von beiden Enden ansequenziert und manuell der Genomdatenbank hinzugefügt. Sie sollten im weiteren Verlauf der Genomsequenzierung bei der Anordnung der Contigs helfen.

3.1.4. Lückenschluss

Bevor mit der eigentlichen „Lückenschlussphase“ begonnen werden konnte, mussten zunächst noch misassemblierte Regionen manuell bereinigt werden. Da das *E. blattae* Genom, wie viele bakterielle Genome, repetitive Sequenzen enthält (z. B. Transposasen), ist es dem verwendeten phrap-Assembler (`www.phrap.org`, Green nicht publiziert) nicht möglich gewesen, alle Sequenzläufe korrekt zusammenzulagern. Nachdem diese Bereiche korrekt assembliert waren, begann die Lückenschlussphase, wie unter 2.10 beschrieben. Dabei wurden zuerst über Plasmidbrücken (s. 2.10.1) verbundene Contigs durch vollständige Sequenzierung der verbindenden Plasmide (*plasmid walking*) miteinander verbunden und diese Lücken dadurch geschlossen. Diese Methode setzt aber voraus, dass die Plasmide wirklich nur ein Insert enthalten und nicht 2 oder womöglich mehrere aus unterschiedlichen Bereichen des Genoms. Solche durch Misligationen entstandenen Plasmide liefern falsche Informationen über Contigzusammenhänge und verursachen ein erhebliches Maß an Arbeit. Misligierte Plasmide verhindern teilweise auch das logische Zusammensetzen ganzer Genombereiche, da die Sequenzinformation dort nicht ausreicht, um diese korrekt anzuordnen. Man ist in dieser Phase außerdem darauf angewiesen, Vorwärts- und Rückwärtslauf einer Sequenz genau zuordnen zu können. Bei misligierten Plasmiden ist dieses nicht möglich.

Als nächster Schritt im Lückenschlussverfahren wurden die Cosmidsequenzen ausgenutzt, um größere Contigs zu ordnen. Hierzu sind Cosmide besonders geeignet,

da sie mit ihren Inserts von bis zu 50 kbp auch repetitive Bereiche überspannen, die von Plasmidinserts nicht vollständig erfasst werden und so die Anordnung der Plasmidsequenzen erschweren. Die Methode des *plasmid walking*, d. h. direktes Sequenzieren auf Cosmiden, ist relativ schwierig und ineffektiv, da große Mengen an Cosmid-DNA für den jeweiligen Sequenzansatz benötigt werden. Deshalb wurde im *E. blattae* Genomsequenzierungprojekt nicht direkt auf Cosmiden sequenziert. Statt dessen wurden mit Hilfe der Cosmide Contigs angeordnet und an den Contigenden Primer abgeleitet, die dann in PCR-Reaktionen mit chromosomaler DNA als Matrize eingesetzt wurden (s. 2.10.2). Die entstandenen PCR-Produkte konnten nach Aufreinigung direkt sequenziert werden. Auf diese Weise wurden bisher die meisten Lücken geschlossen. Aber auch dieser Methode sind Grenzen gesetzt, denn im ungünstigsten Fall kann eine Lücke bis zu 50 kbp groß sein, wenn keinerlei Informationen über den Bereich der Lücke vorliegen. Eine PCR-Reaktion stößt aber bei ungefähr 10 kbp an ihre Grenzen, so dass solche Lücken nicht auf Anhieb geschlossen werden können.

Die Subklonierung von Cosmiden (s. 2.5.6) ist eine Methode, die eingesetzt wird, um die vollständige Sequenz eines Cosmids zu erschließen. Die entstandenen Plasmide werden von den Enden her ansequenziert, assembliert und anschließend in das Gesamtprojekt integriert. Diese Methode ist allerdings sowohl arbeitsintensiv als auch teuer. So ist dieses Verfahren nur dann gefragt, wenn die Lücke sehr groß ist und noch nichts von der Sequenz in der Lücke bekannt ist, da man sonst diesen Bereiche sozusagen doppelt sequenzieren würde. Im vorliegenden Projekt wurde diese Methode bisher bei vier großen Lücken angewandt.

Nachdem die Cosmid- und Plasmidinformationen ausgenutzt worden waren, wurde die Multiplex-PCR (s. 2.10.5) und der Lückenschluss mit dem Annotationsprogramm ERGO durch *peptide links* (s. Materialien und Methoden 2.10.3) verwendet. Durch das „alle gegen alle“ Verfahren der Multiplex-PCR hat man die Möglichkeit, PCR-Produkte zu erhalten, auch wenn man keine Brückeninformation (Plasmide oder Cosmide) zwischen zwei Contigs hat. Sehr große Lücken können mit der Multiplex-PCR meist verkleinert werden, indem große Contigs mit kleineren erweitert werden. Dadurch wird auch die Lücke verkleinert. Im vorliegenden Genomsequenzierungsprojekt wurden mit dieser Methode bisher 35 PCR-Produkte erstellt und vollständig bzw. teilweise sequenziert. Mittels *peptide links* erreicht man im Prinzip das gleiche, betreibt aber einen deutlich geringeren Aufwand und kann gezielt einzelne Contigenden untersuchen. Diese Methode wurde im *E. blattae*-Projekt

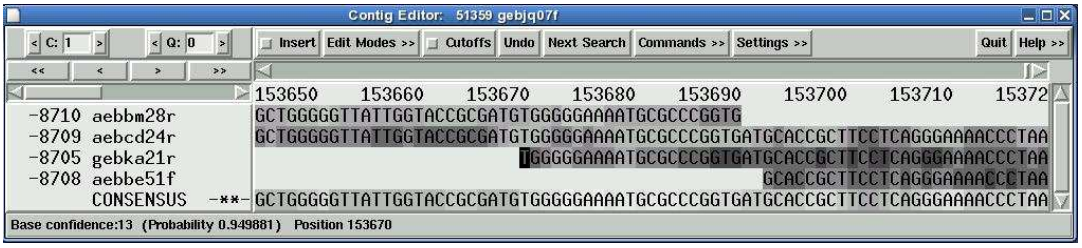

Abbildung 3.1.: Contigeditor in GAP4 mit vier Sequenzläufen und Konsensussequenz.

bisher bei ca. 50 Lücken benutzt.

3.1.5. Qualitätsverbesserung

Die Qualität, d. h. die Verlässlichkeit, einer sequenzierten Base wird durch den *basecaller* PHRED [24] berechnet und als s. g. *quality value* (QV) im Editierprogramm GAP4 [8] als Graustufe in der Hintergrundfarbe dargestellt (s. Abbildung 3.1). Die Qualität der Konsensussequenz ist dabei abhängig von der Qualität der Sequenzläufe. Je heller die Graustufe, ist desto besser ist auch der *quality value*, der sich wie folgt berechnet

$$QV = -10 \cdot log_{10}(P_e)$$

Woraus sich ergibt, dass ein QV von 40 bedeutet, dass die Wahrscheinlichkeit P_e für einen Fehler pro Base kleiner gleich 1/10.000 ist, d. h. bei 10.000 sequenzierten Basen ist im Schnitt mit einem Fehler zu rechnen. Um letztendlich eine gute Sequenzqualität gewährleisten zu können, ist schon während der Lückenschlussphase eine Qualitätsverbesserung in den größeren Contigs durchgeführt worden. Dabei sucht man im Contigeditor nach solchen Stellen in der Konsensussequenz, die einen QV kleiner 40 aufweisen. Hier wird dann mit zusätzlichen Sequenzen auf der Basis von PCRs oder *plasmid walking* eine durchschnittlich höhere Qualität erreicht und somit der *quality value* verbessert.

3.1.6. Annotation

In einem frühen Stadium des Genomprojekts, bei ca. 100 Lücken, wurde einen vollautomatische Annotation mit dem Programm Quann (s. 3.4.1) durchgeführt, das im Rahmen dieser Arbeit entwickelt wurde. Mit dieser ersten Annotation konnte man schon in diesem Stadium schnell einen Überblick über das vorhandene gene-

Tabelle 3.1.: Eigenschaften der *E. blattae* Genomsequenz. [a] Bei diesen Angaben ist eine gewisse, noch vorhandene, Redundanz zum jetzigen Zeitpunkt nicht auszuschließen.

Contigs	43
Anzahl Sequenzen	52.902
Coverage	7
sequenzierte Basen	4.115.744
kodierende Basen (%)	88,00
GC-Gehalt (%)	56,54
ORFs	3.980
Proteine in COG-Gruppen	3.140
„hypothetische“ Proteine	446
rRNA-Gene	31[a]
tRNAs	71[a]

tische Potential des Organismus erhalten. Für eine endgültige Annotation ist das komplexere Programm ERGO (s. 2.12) besser geeignet, da in Quann keine manuelle Annotationsfunktion implementiert ist. Dieses würde auch dem Konzept von Quann widersprechen, das eine schnelle automatische Annotation vorsieht. Aus diesem Grund wurde die endgültige Annotation mit ERGO [74] durchgeführt (s.2.12.2). Die Vorhersage möglicher offener Leserahmen (ORFs) wurde im Programm YACOP (s. 2.12.1) durch die Kombination verschiedener ORF-Vorhersageprogramme erreicht [91]. Insgesamt konnten für *E. blattae* bisher 3.980 offenen Leserahmen vorhergesagt werden.

Im folgenden soll ein Überblick über den aktuellen Stand des Genomprojekts, der in Tabelle 3.1 dargestellt ist, und die Annotation desselben gegeben werden. Die Zahlen entsprechen dem Stand bei 43 Contigs und sind somit nicht endgültig, da die Sequenzierung noch nicht abgeschlossen ist. Von den 3.980 vorhergesagten ORFs konnte 3.140 eine Funktion zugewiesen werden.

Einteilung in COG-Gruppen

Die Klassifikation von Proteinen und ihre Einordnung in Funktionsgruppen erleichtert die Auswertung großer Datenmengen, weil dadurch nicht mehr jede einzelne Annotation ausgewertet werden muss. Die *Cluster of Orthologous Groups* (COGs) [90] eignen sich sehr gut, um Proteine in solche Gruppen (genaue Erklärung s. Anhang A.3.1) einzuteilen. Proteine werden im COG-System in 4.873 *Cluster of Orthologous Groups* eingeteilt, die wiederum zu 25 COG-Klassen zusammengefasst werden

(s. Anhang A.3.1 Seite 143). Ein weiterer Vorteil der Einordnung von Proteinen in COGs besteht darin, dass man die Ergebnisse sehr gut bioinformatisch auswerten kann. Man ist durch die Nummerierung der COGs nicht auf die Auswertung von Annotationen angewiesen, weil diese in der Praxis nicht einheitlich bezüglich der Funktionsnamen sind.

Das in dieser Arbeit entwickelte und verwendete Programm Cogger dient der automatischen Einordnung von Proteinen in ihre COGs und wird unter 3.4.2 beschrieben. Für *E. blattae* konnte diese Einordnung in COG-Gruppen bei 3.140 Proteinen vorgenommen werden, was in Tabelle 3.2 und in Abbildung 3.2 dargestellt ist. Zusätzlich zur relativen Einteilung der Proteine aus *E. blattae* in die COG-Gruppen sind in Abbildung 3.2 dieselben Zahlen für ausgewählte Enterobakterien zu sehen. Vergleicht man die relativen Anteile an Proteinen in den einzelnen COG-Gruppen, so wird deutlich, dass in *E. blattae* keine der Gruppen einen auffällig hohen oder niedrigen Anteil an Proteinen enthält. Im Gegensatz dazu zeigen *Yersinia pestis* und *Shigella flexneri* einen hohen Anteil an Proteinen in der Gruppe L (*Replikation, Rekombination und Reparatur*).

Zusammenfassung

Das Genomsequenzierungsprojekt *E. blattae* befindet sich in seiner Endphase. Die verbleibenden Lücken sind durch eine Kombination aus PCR, Multiplex-PCR und Plasmidwalking zu schließen und werden letztendlich eine Komplettsequenz des Organismus ermöglichen. Die Aufteilung der 3.980 Proteine auf die COG-Gruppen zeigt für *E. blattae* ein ähnliches Protein-Muster wie für andere Enterobakterien.

Tabelle 3.2.: Aufteilung der *E. blattae* Proteine auf die einzelnen COG-Gruppen. Die Summe der Proteine in den Gruppen ergibt nicht die Gesamtzahl von 3.140 Proteinen, da einzelne Proteine in mehrere Gruppen eingeordnet werden konnten.

	Gruppe	**Proteine**	**/ %**
	INFORMATIONSSPEICHERUNG UND -VERARBEITUNG		
J	Translation, ribosomale Struktur und Biogenese	152	4,14
A	RNA-Verarbeitung und -Modifikation	1	0,03
K	Transkription	277	7,55
L	Replikation, Rekombination und Reparatur	203	5,53
B	Chromatin Struktur und Dynamik	0	0
	ZELLULÄRE PROZESSE UND SIGNALVERARBEITUNG		
D	Zelluläre Kontrolle, Zellteilung, Chromosomenaufteilung	33	0,9
Y	Kernstrukturen	0	0
V	Verteidigungsmechanismen	41	1,12
T	Signaltransduktionsmechanismen	119	3,24
M	Zellwand-/Membran-/Hüllenbiogenese	212	5,78
N	Zellbeweglichkeit	18	0,49
Z	Cytoskelett	0	0
W	Extrazelluläre Strukturen	0	0
U	Intrazellulärer Verkehr, Sekretion und vesikulärer Transport	59	1,61
O	Posttranslationelle Modifikation, Proteinumsatz, Chaperone	136	3,71
	METABOLISMUS		
C	Energieproduktion und -umwandlung	275	7,50
G	Kohlenhydratmetabolismus	276	7,52
E	Aminosäuretransport und -metabolismus	378	10,31
F	Nukleotidtransport und -metabolismus	75	2,04
H	Coenzymtransport und -metabolismus	175	4,77
I	Lipidtransport und -metabolismus	100	2,73
P	Transport und Metabolismus anorganischer Ionen	283	7,72
Q	Biosynthese von Sekundärmetaboliten, Transport und Katabolismus	92	2,51
	DÜRFTIG CHARAKTERISIERT		
R	Generelle Funktionsvorhersage	474	12,92
S	Funktion unbekannt	289	7,88

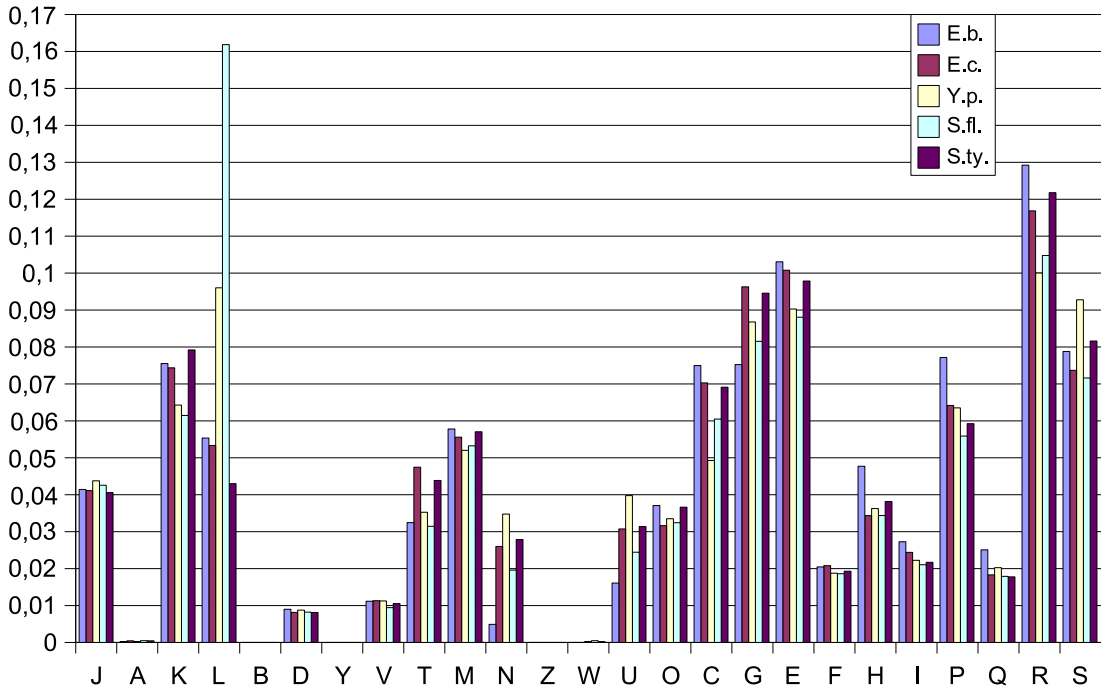

Abbildung 3.2.: Relative Aufteilung der Proteine aus *E. blattae* in die einzelnen COG-Gruppen im Vergleich zu anderen Enterobakterien. Abkürzungen: E.b. *Escherichia blattae*, E.c. *Escherichia coli*, Y.p. *Yersinia pestis*, S.fl *Shigella flexneri*, S.ty. *Salmonella typhimurium* COG-Gruppen (s. Tabelle 3.2): **J..B** *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signale* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert.*

3.1.7. Horizontaler Gentransfer in *Escherichia blattae*

Der Transfer von Genen zwischen verschiedenen Spezies wird als horizontaler Gentransfer (HGT) bezeichnet. Dieses Phänomen wird als eine der treibenden Kräfte in der mikrobiellen Evolution angesehen [44]. Innerhalb der Enterobakterien ist horizontaler Gentransfer unter anderem für *Salmonella* beschrieben worden [76] und auch außerhalb dieser Familie wird horizontaler Gentransfer häufig beobachtet. In dieser Arbeit soll *E. blattae* auf putativ horizontal erworbene Gene hin untersucht werden. Dazu wird das Programm SIGI [62] genutzt, das einen konservativen Ansatz zur Fremdgenerkennung verwendet und auf die Erkennung von Fremdgen-Inseln spezialisiert ist.

Die Auswertung der bisherigen 3.980 ORFs mit SIGI ergab 337 mögliche Fremdgene, die detailiert im Anhang A.13 auf Seite 176 aufgelistet sind. Unter diesen Fremdgenen konnten drei Fremdgen-Inseln identifiziert werden. Bei näherer Betrachtung stellte sich heraus, dass es sich bei diese Inseln um Bakteriophagen handelt (s. Tabelle 3.3).

Um einen Überblick über die Proteine zu bekommen, die als putative Fremdgene identifiziert wurden, untersuchte man diese Proteine auf ihre COG-Gruppen-Einteilung. Das normierte Ergebnis ist in Abbildung 3.3 im Vergleich zur Verteilung aller Proteinen dargestellt. Ein hoher Anteil der erkannten Fremdgene wurde den Gruppen L, M, R und S zugeordnet.

Es stellt sich die Frage, ob Fremdgene in bestimmten COG-Gruppen unter- oder überrepräsentiert sind. Abbildung 3.4 zeigt das Verhältnis zwischen den normierten Werten für die Anzahl der Fremdgene und aller Nicht-Fremdgene in den einzelnen Gruppen. In den Gruppen L und M liegen Verhältnisse von 5,3 und 2,2 vor, d. h. in diesen Gruppen liegen überproportional viele Fremdgene, währen in J, O und C nur wenige Fremdgene vorkommen. Die Einteilung der putativen Fremdgene in *E. blattae* (s. Abbildung 3.4) lässt erkennen, dass die COG-Gruppen L (*Replikation,*

Tabelle 3.3.: Durch Fremdgenerkennung mit SIGI und durch Annotation bestätigte Bakteriophagen in *E. blattae*

Name	Proteine		Anzahl Proteine	Größe in bp
	Anfang	Ende		
Eb1	REB003928	REB002868	33	34.190
Eb2	REB002811	REB003138	56	49.531
MuEb	REB004375	REB003517	28	30.924

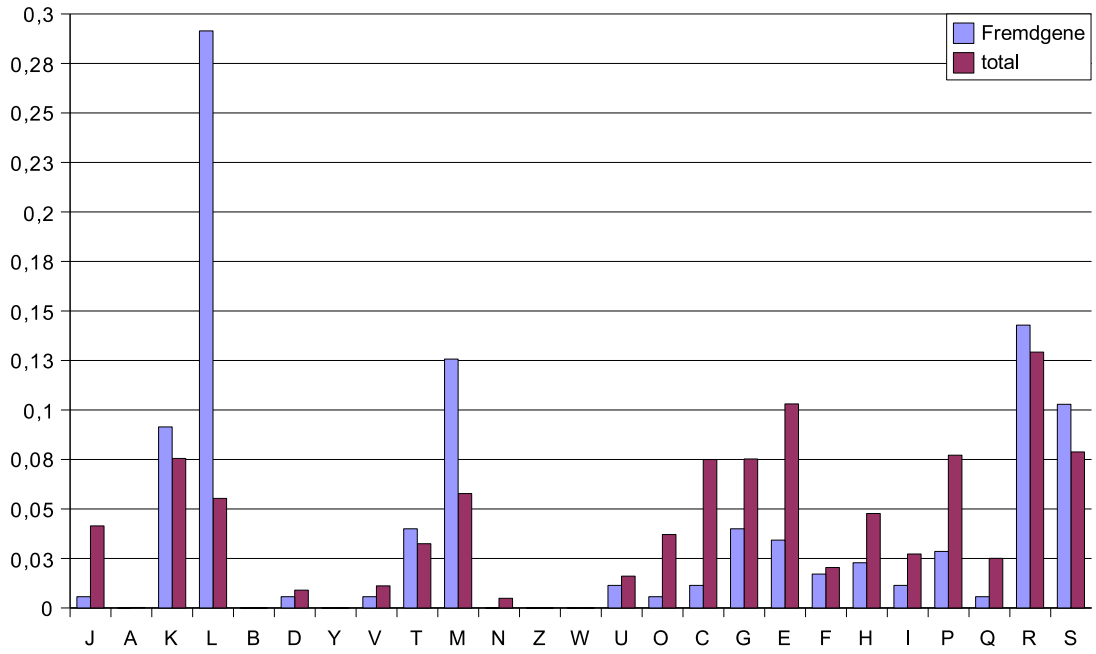

Abbildung 3.3.: Anteilige Zugehörigkeit der *E. blattae* Fremdgene zu COG-Gruppen im Vergleich zur Verteilung bei allen Proteinen (genaue Erklärung s. Anhang A.3.1): **J..B** *Informationsspeicherung- und verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *Dürftig charakterisiert.*

Rekombination und Reparatur) und M (*Zellwand-/Membran-/Hüllenbiogenese*) um den Faktor 5 bzw. 2 häufiger in Fremdgenen auftreten als in solchen, die nicht als fremd eingestuft werden. Die hohe Anzahl von Proteinen aus der Gruppe L ist dadurch zu erklären, dass zu den 51 Fremdgenen in dieser Gruppe 31 Transposasen und 4 Integrasen aus Phagen zählen. Die Gruppen K (*Transkription*), T (*Signaltransduktionsmechanismen*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) zeigen nur schwache aber ebenfalls überdurchschnittlich viele Fremdgene.

Zusammenfassung

In *E. blattae* konnten 337 mögliche Fremdgene identifiziert werden. Unter diesen Genen befinden sich die Gene von drei Prophagen. Die putativen Fremdgene sind nicht gleichmäßig auf alle COG-Gruppen verteilt. Die Gruppen K, L, M, R und S sind statistisch auffällig häufig in den Fremdgenen zu finden.

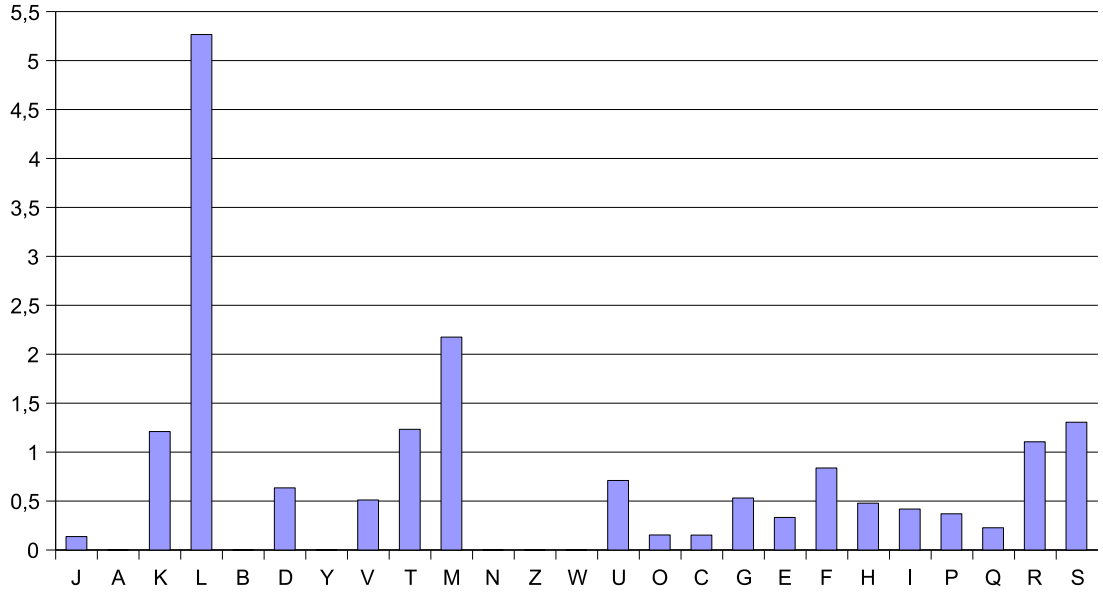

Abbildung 3.4.: Verhältnis von Fremdgenen in *E. blattae* zu Nicht-Fremdgenen in den einzelnen COG-Gruppen.

3.1.8. Vergleich mit anderen Enterobakterien

Die Annotation der bisher abgeleiteten 3.980 Proteine aus *E. blattae* liefert Hinweise darauf, dass die größten Ähnlichkeiten dieser Proteine nur zum Teil im Genus *Escherichia* zu finden sind. Einige Proteine weisen auf eine nahe Verwandtschaft mit dem Genus *Salmonella* hin. Um die Ähnlichkeit von *E. blattae* mit den anderen sequenzierten Organismen der *Enterobacteriaceae* zu untersuchen, wurde das *E. blattae* Proteom mit denen von *Salmonella typhimurium* LT2 [60], *Salmonella enterica* [17], *Yersinia pestis* KIM [16], *E. coli* K12 [7], *Shigella flexneri* 2a [39] und *Buchnera aphidicola* [93] durch bi-direktionale BLAST-Analysen verglichen (s. 2.11.1). In Tabelle 3.4 sind die Ergebnisse zusammengefasst. Es wird ersichtlich, dass *E. blattae* zwischen 2.630 (66 %) und 2.168 (54 %) Proteine mit den anderen Enterobakterien teilt. Eine Ausnahme ist *Buchnera aphidicola* mit 435 ähnlichen Proteinen, dessen Genom mit nur 616 kbp sehr klein ist. Die mittlere Übereinstimmung in der Proteinsequenz beim Vergleich der Proteine aus *E. blattae* mit denen anderer Enterobakterien beträgt zwischen 74 und 68 Prozent. Diese Zahlen belegen, dass es sich bei *E. blattae* offensichtlich um einen Vertreter der Enterobakterien handelt, der aber einen relativ hohen Anteil an eigenen, spezifischen Proteinen besitzt.

Es ergibt sich, dass 650 Proteine zwar in *E. blattae* aber nicht in den anderen Enterobakterien zu finden sind. Unter diesen 650 Proteinen sind 535 (82 %) als

Tabelle 3.4.: Vergleich von *E. blattae* mit anderen Enterobakterien. Gemeinsam: bi-direktionale BLAST Treffer und Anteil am *E. blattae* Proteom. **Identität**: durchschnittliche identische Residuen im Protein-Alignment. **Hypothetisch**: Anteil hypothetischer Proteine an den „eigenen". **Abkürzungen**:*E. Escherichia*, *S. Salmonella*, *Sh. Shigella*, *Y. Yersinia*, *B. Buchnera*.

Organismus	**gemeinsame Proteine (%)**	**Identität in %**	**nur in *E. blattae***	**hypothetisch (%)**
E. coli	2630 (66)	74	897	649 (72)
S. typhimurium	2619 (66)	73	891	649 (73)
S. enterica	2545 (64)	73	962	637 (66)
Sh. flexneri	2436 (61)	74	1015	701 (69)
Y. pestis	2168 (54)	68	1183	725 (62)
B. aphidicola	435 (11)	52	3241	1158 (35)

hypothetical oder *hypothetical conserved* annotiert. Eine vollständige Liste dieser Proteine findet sich im Anhang A.14. In Abbildung 3.5 ist dargestellt, welche COG-Gruppe in den spezifischen Proteinen aus *E. blattae* über- oder unterrepräsentiert ist. Spezifisch sind dabei die 650 Proteine, die keinen Treffer in der bi-direktionalen BLAST-Suche ergeben haben. Überrepräsentiert in dieser Menge von Proteinen sind die COG-Gruppen K (*Transkription*), L (*Replikation, Rekombination und Reparatur*), N (*Zellbeweglichkeit*), Q (*Biosynthese von Sekundärmetaboliten, Transport und Katabolismus*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*). Die Annotationen der einzelnen spezifischen Proteine machen klar, dass die Gruppe K hauptsächlich Transkriptionsregulatoren und die Gruppe L Phagenproteine und Transposasen enthält. In den Gruppen N und S sind hauptsächlich hypothetische Proteine zu finden, währen die Gruppe R hypothetische und einige Phagenproteine enthält. Isomerasen und Oxygenasen sind in der Gruppe Q anzutreffen.

Aus dem Vergleich von *E. blattae* mit den anderen Enterobakterien zeigt sich, dass 417 Proteine in allen wiederzufinden sind, von denen nur 28 (7 %) als *hypothetical* annotiert sind (s. Anhang A.16).

Zusammenfassung

Der Proteom-Vergleich von *E. blattae* mit anderen sequenzierten Enterobakterien hat ergeben, dass dieser Organismus die meisten homologen Proteine mit *E. coli* und *S. typhimurium* teilt. Hier ist mit jeweils 74 % bzw. 73 % auch die größte durchschnittliche prozentuale Identität der Proteinsequenz zu finden. Bei diesem Vergleich

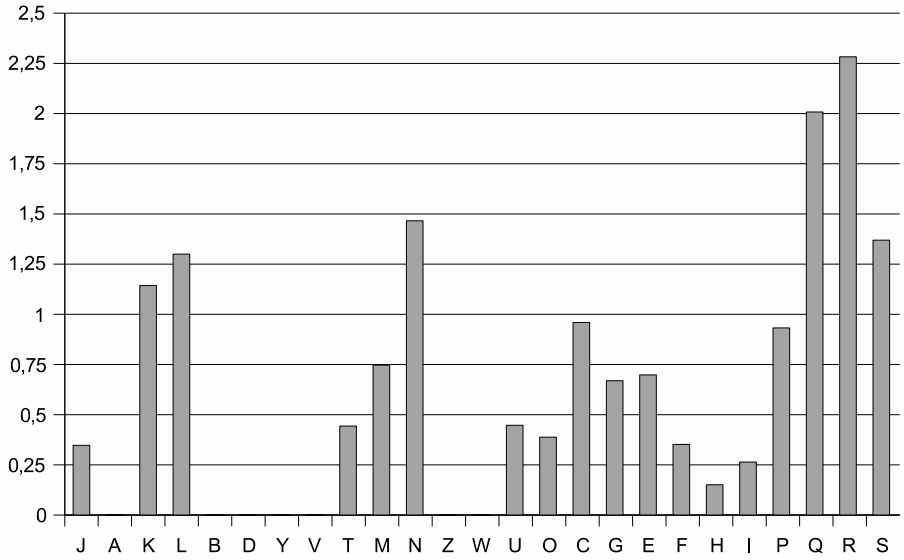

Abbildung 3.5.: Relative Präsenz in COG-Gruppen von Proteinen aus *E. blattae*, die kein homologes Protein in den anderen Enterobakterien haben. Verglichen mit allen Proteinen aus *E. blattae*.

ist außerdem deutlich geworden, dass *E. blattae* 650 Proteine besitzt, die keine ausreichende Ähnlichkeit zu den anderen Enterobakterien zeigten. Von den *E. blattae*-spezifischen Proteinen, sind 46,6 % in der Liste der putativen Fremdgene wiederzufinden (s. Anhang A.15).

3.2. Untersuchung des *Methanosarcina mazei* Gö1 Genoms

Zu den bereits vollständigen Genomensequenzen, die in dieser Arbeit betrachtet werden, gehört die von *Methanosarcina mazei*. Die Sequenz des methanogenen Archaeons *Ms. mazei* Stamm Gö1 wurde im Juni 2001 entschlüsselt [18]. Der Organismus ist biochemisch gut untersucht, so dass die nun vorhandene vollständige Sequenz des Organismus weitere Untersuchungen geradezu herausfordert. In dieser Arbeit sollen in den folgenden Abschnitten mehrere Aspekte der bioinformatischen Analyse des *Ms. mazei* Genoms aufgezeigt werden.

3.2.1. Lokalisation des Replikationsursprungs

Nachdem die Sequenzierung eines Genoms abgeschlossen ist, stellt sich oft als erstes die Frage nach dem Replikationsursprung, weil es sich anbietet, die endgültige Sequenz in den Sequenzdatenbanken mit dem Ursprung „beginnen“ zu lassen. Die Replikation bakterieller Genome startet in einem zwischen 100 und 1.000 bp langen Bereich auf dem Genom, der *origin of Replication* oder auch kurz *Ori* genannt wird [55].

Die *in silico* Bestimmung des Replikationsursprungs ist bei Bakterien relativ einfach [27]. Hier führt der GC-Skew (ein Maß für die Unverteilung des GC-Gehalts, siehe weiter unten) normalerweise zum Erfolg, weil die meisten *Bacteria* im *leading strand* mehr Guanin- als Cytosin-Basen aufweisen [66]. Bei Archaeen treten allerdings oft Probleme auf, weil genau diese Präferenz für eine G- vor einer C-Base nicht zu finden ist [66]. Hier sollen am Beispiel von *Ms. mazei* einige Methoden vorgestellt werden, mit deren Hilfe man Hinweise auf den Bereich finden kann, an dem die DNA-Replikation startet. Das wohl bekannteste Verfahren, der GC-Skew [52]

$$GCSkew = \frac{G_i - C_i}{G_i + C_i}$$

wurde unter anderem im *E. coli* K12 Genomsequenzierungsprojekt angewandt [7]. Bei dieser Formel wird ein Fenster, das immer eine bestimmte Anzahl von Basen enthält (ein s. g. *sliding window*), über die Sequenz verschoben und dabei der GC-Skew berechnet. Ein solcher GC-Skew stellt eine Funktion dar, die im Idealfall um eine Achse oszilliert und am *Origin of Replication* das Vorzeichen wechselt. Dieses Verfahren nutzt eine Ungleichverteilung von G- und C-Basen im *leading-* bzw. *lagging-strand* aus. Die Ursachen für diese Verteilung liegen vermutlich im unterschiedlichen Selektionsdruck, der sowohl auf dem kodierenden als auch auf dem nicht-kodierenden Strang liegt [52]. Daraus folgt, dass dieser Effekt umso stärker auftritt, je mehr Gene in einem Organismus in derselben Richtung wie die fortschreitende Replikationsgabel transkribiert werden. Dieser Effekt ist z. B. bei *Bacilli* und *Clostridien* sehr ausgeprägt [10].

Mittlerweise gibt es mehrere Ansätze zur Berechnung von Skews, die in [35] und [27] beschrieben werden. Grigoriev [35] beschreibt den kumulativen GC-Skew

$$GCSkew_{kum.} = \frac{G_{(i,i+n)} - C_{(i,i+n)}}{G_{(i,i+n)} + C_{(i,i+n)}}$$

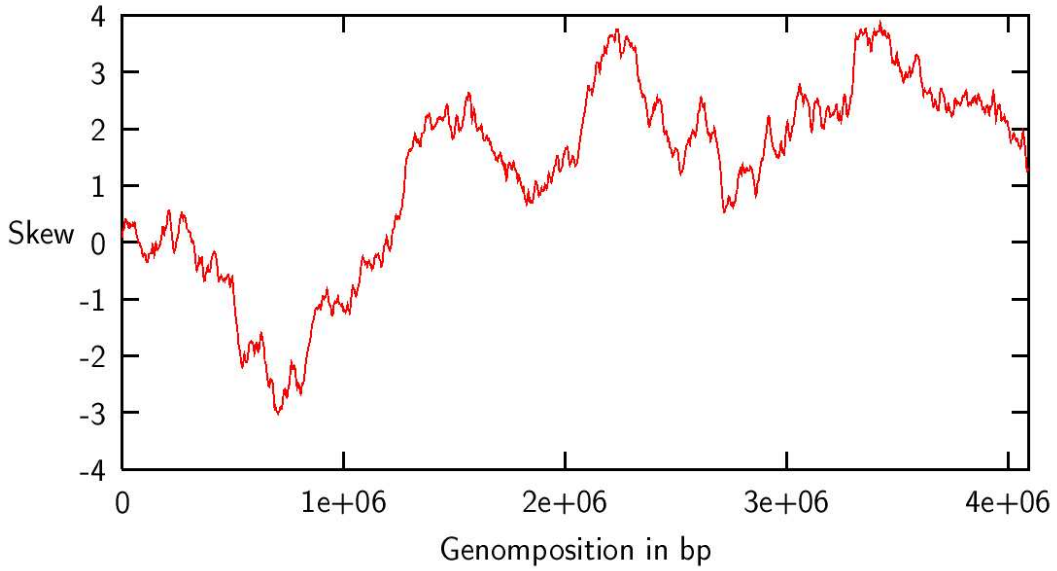

Abbildung 3.6.: Kumulativer GC-Skew von *Ms. mazei*.

wobei das Ergebnis eine Kurve mit einem Minimum am Origin und einem Maximum am Terminus liefern sollte. Die Ergebnisse dieser Methode sind erheblich einfacher zu interpretieren und außerdem sehr viel genauer als eine oszillierende Funktion wie der klassische GC-Skew. Der GC-Skew zeigt nicht für alle Genome, insbesondere nicht für Archaeen, aussagekräftige Ergebnisse (s. Abbildung 3.6). Ein klares Minimum bzw. Maximum, wie es für *E. coli* [27] zu sehen ist, konnte bei *Ms. mazei* nicht gezeigt werden.

Freeman stellt in [27] drei weitere Berechnungsmethoden vor, die in dieser Arbeit bei *Ms. mazei* angewandt wurden, um eine bessere Auflösung bei der Lokalisation des Replikationsursprungs zu erhalten. Der *purine-excess* wird durch

$$PE_i = (A_{(0,i)} + G_{(0,i)}) - (T_{(0,i)} + C_{(0,i)})$$

beschrieben und ermittelt das Verhältnis von Purin- zu Pyrimidin-Basen entlang einer Sequenz. Das Minimum einer Auftragung des *purin-excess* zeigt in den meisten Fällen auf den Replikationsursprung. Ähnlich verhält sich der *keto-excess*, berechnet durch

$$KE_i = (T_{(0,i)} + G_{(0,i)}) - (A_{(0,i))} + C_{(0,i)})$$

Hier wird das Verhältnis von Keto-Gruppen (GT) zu Amino-Basen (AC) beschrieben. Wie in Abbildung 3.7 dargestellt, wies die Kurve des *keto-excess* im Bereich

von ca. 1,55 Mbp ein Minimum und der *purin-excess* dort ein Maximum auf. Der kumulative GC-Skew war für *Ms. mazei* in dieser Region unauffällig.

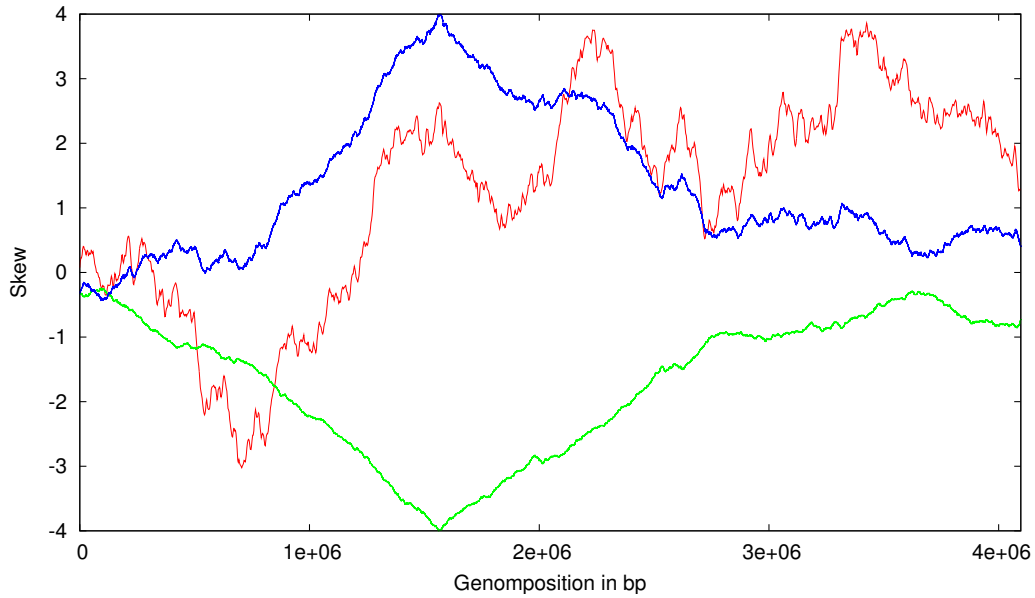

Abbildung 3.7.: Verschiedene Skews des *Ms. mazei* Genoms. Rot: kumulativer GC-Skew, blau: *purine-excess*, grün: *keto-excess*.

Der *coding-strand excess* spiegelt die Summe aller kodierenden Nukleotide entlang einer Sequenz wider ($cdns_{(0,i+n)}$). Ihre Komplemente auf dem anderen Strang, beschrieben durch $cdns'_{(0,i+n)}$, werden für die Berechnung des *coding-strand excess* subtrahiert

$$CE = 3(cdns_{(0,i)} - cdns'_{(0,i)})$$

Abbildung 3.8 gibt eine Übersicht über die bisher vorgestellten Skews. Dabei fällt auf, dass der *coding-strand excess* der Kurve des kumulativen GC-Skews folgt. Bis zu diesem Zeitpunkt lässt sich festhalten, dass alle Kurven einen unklaren Verlauf nehmen, der sich so bei den *Bacteria* nicht wiederfinden lässt (s. [27]). Die Kurven des *purin-* und *keto-excess* haben ihre Maxima und Minima bei ca. 1,55 Mbp. Dass das Minimum des *purin-excess* auf den Replikationsursprung zeigt [27], war bei *Ms. mazei* nicht zu beobachten. Die verschiedenen Skewauswertungen deuten eher darauf hin, dass das Maximum des *purin-excess* einen Hinweis auf den *Ori* gibt.

Um die Frage nach dem Ort des Replikationsstarts letztendlich zu klären, wurde eine weitere Methode zu Hilfe genommen. Der Oligomer Skew beschreibt das Vorkommen bestimmter Oligomere - in diesem Fall Tetramere [68]. Ein Tetramer ($a_1..a_4$) in einer Sequenz ($i..n$) wird durch $O_{(i,i+n)}(a_1a_2a_3a_4)$ dargestellt. Der kumu-

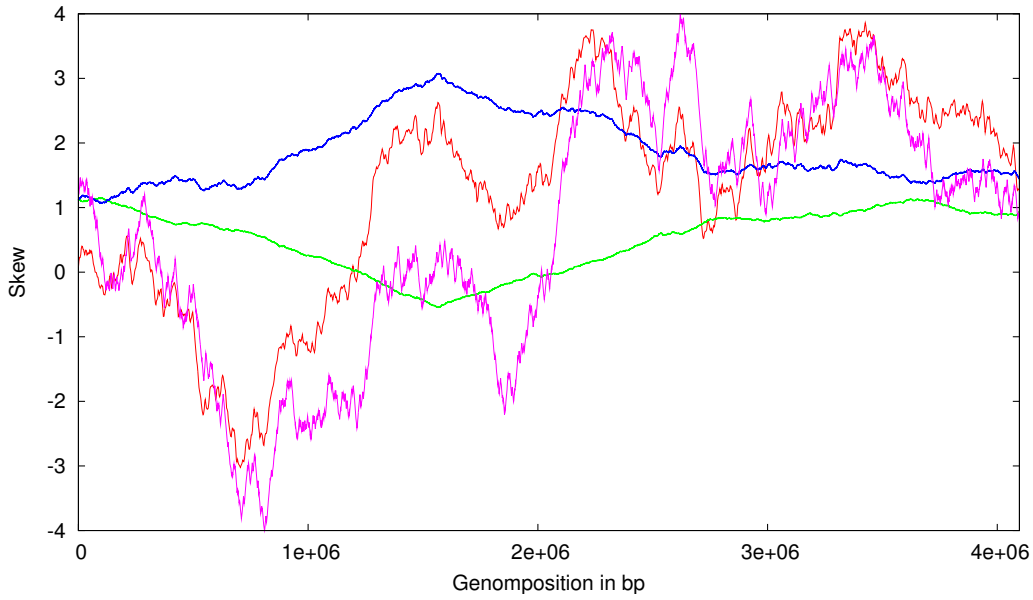

Abbildung 3.8.: Verschiedene Skews des *Ms. mazei* Genoms. Rot: kumulativer GC-Skew, blau: *purine-excess*, grün: *keto-excess*, violett: *coding-strand excess*.

lative Tetramerskew für das definierte Tetramer GGGT ergibt sich dann aus

$$TS_{(i)} = \frac{O_{(0,i)}(GGGT)}{O_{(0,i)}(ACCC)}$$

Abbildung 3.9 zeigt den kumulativen Oligomerskew für das Oligomer GGGT in *Ms. mazei*. Das Oligomer ist gewählt worden, weil es für *Pyrococcus horikoshii* und *Pyrococcus abyssi* verwendet wurde [68]. Andere Oligomere gaben ein schwächeres Signal, hatten ihr Minimum aber alle an derselben Position (Daten nicht gezeigt). Dieses Ergebnis ist der dritte Hinweis auf einen möglichen Replikationsursprung im Bereich bei 1,55 Mbp im *Ms. mazei* Genom.

Verlässt man die Skew-Ebene und betrachtet die Protein-Ebene, so muss erwähnt werden, dass in Archaeen kein Homolog zum bakteriellen Replikationsprotein DnaA zu finden ist. Statt dessen findet man ein Homolog zum eukaryotischen Orc1/Cdc6 Protein [33]. Dieses Protein beginnt in *Ms. mazei* an der Postion 1.563.415 bp und endet an Position 1.564.656 bp (RMMZ01787 oder MM1314). Stromaufwärts davon befindet sich eine 1.586 bp große intergene Region, die mit 25,74 % einen auffallend niedrigen GC-Gehalt aufweist, während der GC-Gehalt des Genoms bei 41,48 % liegt (siehe dazu auch[33]).

Die Lokalisation des Proteins in direkter Nachbarschaft zum Replikationsursprung

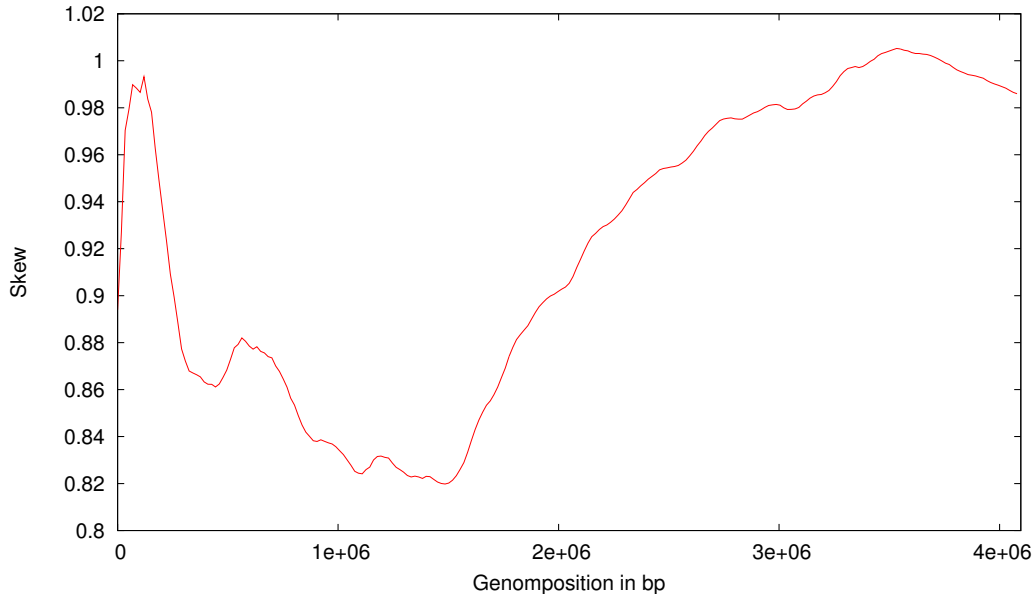

Abbildung 3.9.: Oligomer Skew für das Tetramer GGGT in *Ms. mazei*.

geht einher mit den Beobachtungen von MYLLYKALLIO [68]. Ebenfalls typisch für die Sequenz des Replikationsursprungs sind lange *inverted repeats* an den Enden mit vielen kürzeren Repeats dazwischen [43]. Dieses Ergebnis konnte ebenfalls für die putative Origin-Region in *Ms. mazei* gezeigt werden (Daten nicht gezeigt). In Abbildung 3.10 ist die Origin-Region in *Ms. mazei* dargestellt. Es fällt auf, dass in dem gesamten Bereich eine Reihe, für diese Regionen typische, Proteine [33] zu finden sind (s. Tab. 3.5). In *Ms. mazei* ist ein zweites Cdc6-Protein (MM1387 bei 1.564.656 bp) zu finden. Dieses Protein ist aber zu keiner größeren intergenen Region benachbart, so dass ein Replikationsstart hier eher unwahrscheinlich ist.

Zusammenfassung

Unter Berücksichtigung der Informationen aus den verschiedenen Skews, der relativ großen intergenen Region mit einem GC-Gehalt von nur 25 % neben einem Cdc6-Protein und den gefundenen Replikationsproteinen ist anzunehmen, dass die Replikation des Genoms von *Ms. mazei* im Bereich zwischen Position 1.564.656 und 1.566.242 beginnt. Es wird außerdem klar, dass die Hinweise auf den Replikationsursprung nur durch die Kombination der Information mehrerer Methoden gewonnen werden konnte.

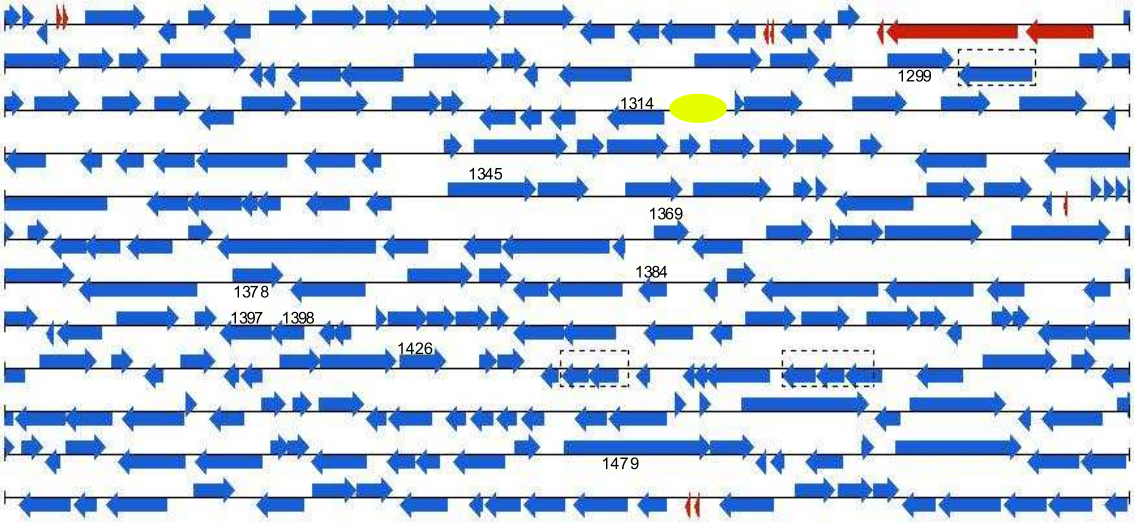

Abbildung 3.10.: Darstellung der Gene (blau) in der Region (1,5..1,8 Mbp) um den *Origin of Replication* (gelbes Oval) in *Ms. mazei*. Rot: (t)RNA-Gene. Die Annotation der abgeleiteten Proteine ist in Tabelle 3.5 dargestellt.

Tabelle 3.5.: Proteine, die mit der Replikation in Zusammenhang stehen, in der Umgebung des Replikationsursprungs von *Ms. mazei*.

Proteinnummer	Annotation
MM1295	DNA primase
MM1299	Replication factor-A protein
MM1314	origin recognition complex subunit
MM1345	DNA polymerase II small subunit
MM1369	DNA repair protein
MM1378	origin recognition complex subunit
MM1384	Superfamily I DNA and RNA helicase
MM1397	DNA polymerase sliding clamp
MM1398	DNA-directed RNA polymerase subunit M
MM1426	replication factor C subunit
MM1479	SWF/SNF family helicase

3.2.2. Vergleich mit *Methanosarcina acetivorans* C2A

Mit der Genomsequenz von *Methanosarcina acetivorans* [31] ist die Sequenz eines mit *Ms. mazei* sehr nah verwandten Organismus verfügbar. Beide Archaeen gehören zur Familie der *Methanosarcinaceae*. *Ms. acetivorans* zeichnet sich im Gegensatz zu *Ms. mazei* durch die Fähigkeit aus, bei tieferen Temperaturen (9..30 °C) wachsen zu können. Bei *Ms. mazei* ist im Gegensatz dazu Wachstum im Bereich von 20 bis 40 °C nachgewiesen. Gravierende physiologische Unterschiede zwischen diesen beiden *Methanosarcina*-Spezies sind allerdings nicht bekannt [56].

Vor diesem Hintergrund erstaunt es, dass *Ms. acetivorans* ein deutlich größeres Genom besitzt als *Ms. mazei* (s. Tab. 3.6). In den folgenden Abschnitten soll auf den Vergleich der beiden Organismen genauer eingegangen werden.

Tabelle 3.6.: Übersicht über die Genome von *Ms. mazei* und *Ms. acetivorans*.

	Ms. mazei	***Ms. acetivorans***	**Differenz**
Genomgröße (bp)	5.751.492	4.096.345	1.655.147
Proteine	3.371	4.528	1.169
Kodierende Sequenz (%)	75,13	73,96	

Genom-Alignment

Needleman und Wunsch [70] sowie Smith und Waterman [89] haben die ersten Algorithmen zum Sequenzvergleich beschrieben. Ihre Verfahren sind zum Standard in der Bioinformatik geworden. Mit diesen Methoden kann man allerdings, bei der heutigen Computerausstattung, keine ganzen Genome alignieren, da der Aufwand für diese Verfahren mit $O(n^2)$ sehr groß ist. Die Rechenzeit nimmt hierbei mit dem Quadrat der Sequenzlänge zu. Es wurden für den Zweck des *whole genome alignment* neue, schnellere und effizientere Verfahren entwickelt. Ein Beispiel ist MUMmer [15], das auf *Suffix-Trees* basiert und somit außerordentlich schnell ist (etwa $O(n)$).

Mit dem Alignment ganzer Genome kann man zwei grundsätzliche Ziele verfolgen: zum einen ist es möglich, Sequenzen unterschiedlicher Stämme einer Spezies (z. B. pathogene gegen apathogene) zu vergleichen und somit Unterschiede zu finden, die womöglich nur im Austausch weniger Basen in einem Genom bestehen. Zum anderen verfolgen *whole genome alignments* das Ziel, sich einen allgemeinen Überblick über die Ähnlichkeit zweier Genomsequenzen zu verschaffen. In dieser Arbeit wurden zum Vergleich von *Ms. mazei* mit *Ms. acetivorans* zwei sich komplementierende

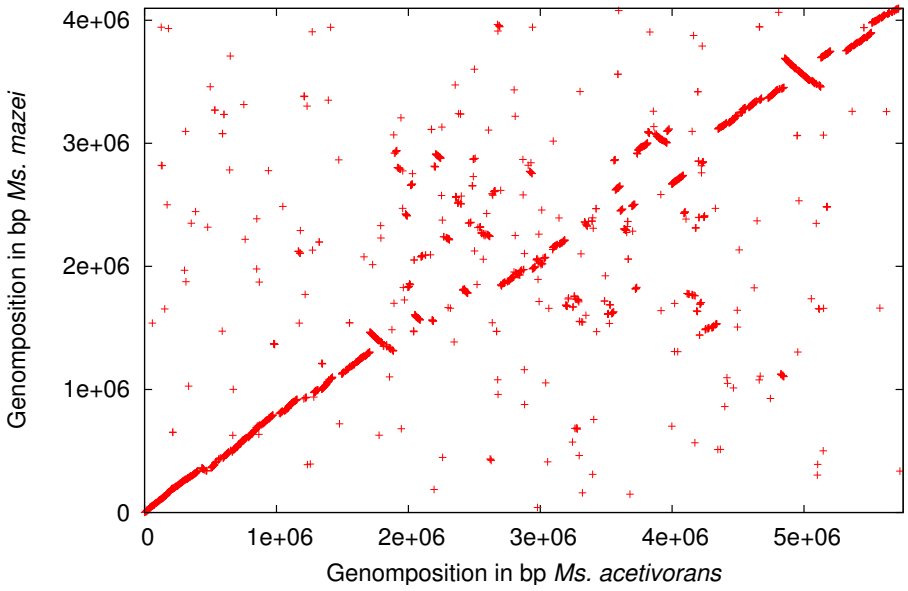

Abbildung 3.11.: Genomplot *Ms. mazei* gegen *Ms. acetivorans*. Ein Kreuz in der Grafik stellt einen bi-direktionalen BLAST-Treffer an dieser Position dar.

Alignment-Verfahren angewandt: Alignments auf DNA- und Protein-Ebene. Das Alignment auf DNA-Ebene mit MUMmer (Mindestlänge des Alignments 20 bp) [15] zeigte viele kleine, nicht zusammenhängende Bereiche (Daten nicht gezeigt), woraus man schließen kann, dass die Ähnlichkeit der beiden Organismen auf DNA-Ebene zwar vorhanden aber nicht durchgängig groß ist. Um die Ähnlichkeit der beiden Archaeen besser darzustellen, wurde ein anderes Vergleichsverfahren gewählt. Dabei werden die Organismen auf Proteinebene verglichen. Das Vorgehen besteht aus einem bi-direktionalen BLAST (s. 2.11.1) mit grafischer Darstellung und anschließender visuellen Auswertung.

Das Ergebnis dieses Vergleichs ist in Abbildung 3.11 als Auftragung der BLAST-Treffer zu sehen. Das Genom von *Ms. mazei* ist dabei auf der Y-Achse und das von *Ms. acetivorans* auf der X-Achse dargestellt. Ein Kreuz in der Abbildung entspricht, einem `blastp`-Treffer mit einem *e-value* von $\leq 1{\cdot}10^{-5}$ zwischen den beiden Proteinen an der jeweiligen Position im Genom. Im Nullpunkt des Koordinatenkreuzes sind die Replikationsursprünge der beiden Sequenzen angesiedelt, wobei sich in der Mitte einer gedachten Diagonalen der Terminus befindet. Die Sequenz von *Ms. mazei* wurde für diese Auswertung zur besseren Darstellung um 1,5 Mbp verschoben. Vergliche man ein Genom mit sich selbst, so träfen in einem BLAST-Vergleich alle Proteine auf ihre eigene Position im „Vergleichsgenom“. In diesem Fall bestünde die Grafik

aus einer einzigen durchgängigen Diagonalen, neben der keine anderen Treffer sichtbar wären. Die Synthenie, das heißt die konservierte Genanordnung, wäre dann 100 prozentig. Die Realität sieht aber etwas anders aus, wie der Fall von *Ms. mazei* und *Ms. acetivorans* zeigt. Bei 1,8 Mbp und 5 Mbp auf der X-Achse sind Inversionen daran zu erkennen, dass die Bereiche orthogonal zur Diagonalen stehen. Auffällig ist weiterhin, dass der Bereich um den Replikationsursprung höher konserviert ist als die Region um den Terminus bei 3 Mbp, so dass hier die meisten Unterschiede der beiden Genome zu erwarten sind.

Wirft man einen Blick in die Liste der Gene in *Ms. acetivorans*, die kein homologes Protein in *Ms. mazei* haben, so fällt auf, dass sich in eben dieser Region zwischen 2 und 4,5 Mbp 550 Gene befinden, die spezifisch für *Ms. acetivorans* sind. Von diesen 550 Genen sind 492 als *predicted protein* oder *hypothetical protein* annotiert. Unter den restlichen 58 Proteinen sind 13 Transposasen, diverse andere Proteine und eine H^+ transportierende ATP-Synthase (6 Untereinheiten) zu finden. Zu dieser Beobachtung passen die in Tabelle 3.7 dargestellten Ergebnisse, die zeigen, dass 87 % aller Proteine aus *Ms. acetivorans* hypothetischen Charakter besitzen.

Vergleich der Proteome

Globale Genom-Alignments zeigen Inversionen, Deletionen oder Insertionen auf, geben einen Einblick in eventuell vorhandene Synthenie oder geben einen Überblick darüber, ob zusammenhängende Bereiche zweier Genome ähnlich sind. Sie liefern jedoch keine genaue Information über vorhandene bzw. nicht vorhandene Proteine in den Organismen. Das bedeutet, dass erst eine Gegenüberstellung der Proteome genaue Erkenntnisse über das versteckte Potential der Organismen im Vergleich aufzeigt.

Die Aufgabe des Proteomvergleichs wird in dieser Arbeit mit dem schon erwähnten bi-direktionalen BLAST (s. 2.11.1) auf Proteinsequenzebene bewältigt. Tabelle 3.7 gibt einen Überblick über die Ergebnisse des Vergleichs (`blastp` mit *cut-off e-value* von $\leq 1 \cdot 10^{-5}$). Die letzte Spalte repräsentiert die Anzahl von Proteinen, die im jeweils anderen Organismus vom *ORF-finder* GLIMMER nicht als Proteine vorhergesagt waren. Sie konnten nur durch eine `tblastn`-Suche (Proteinsequenz verglichen mit der Übersetzung aller sechs Leserahmen einer DNA-Sequenz) gefunden werden. Wenn die Treffer einen *e-value* von $1 \cdot 10^{-5}$ aufwiesen und Homologie über den gesamten Bereich des Proteins vorhanden war, dann galt das Protein als homo-

Tabelle 3.7.: *Ms. mazei* und *Ms. acetivorans* Proteome im Vergleich. Die zweite Spalte gibt die Anzahl der Proteine wider, die durch bi-direktionalen BLAST als Homologe bezeichnet werden konnten. Spalte 3 zeigt an, wie viele Proteine für den jeweiligen Organismus spezifisch sind, also kein Homolog besitzen. Die letzte Spalte enthält die Proteine, die nicht durch die ORF-Vorhersage-Programme vorhergesagt wurden (s. Diskussion). Die Addition von Spalte 1 und 2, ergibt nicht die Gesamtzahl der Proteine in dem entsprechenden Organismus, weil die Proteine, die einen nicht bi-direktionalen BLAST-Treffer haben, nicht mitgezählt werden.

	Homologe Proteine	**Organismus-spezifische Proteine („hypothetisch“ %)**	**Regionen von `tblastn` gefunden**
Ms. mazei	2.449	425 (227/53%)	61
Ms. acetivorans	2.449	1.051 (912/87%)	10

loges. Im Anhang A.2 und A.3 ist eine detaillierte Liste der organismenspezifischen oder „eigenen“ Proteine abgedruckt, die nur in *Ms. mazei* oder *Ms. acetivorans* zu finden sind (d. h. keinen BLAST-Treffer im Vergleichsgenom haben). Der hohe Anteil hypothetischer (in *Ms. acetivorans* „predicted“) Proteine in der Menge derer, die kein homologes Protein im anderen Genom haben, ist besonders bei *Ms. acetivorans* auffällig.

Zur Klassifikation von Proteinen und ihrer Einteilung in Gruppen wird auch hier wieder das COG-System [90] verwendet, das schon unter 3.1.6 beschrieben wurde. Die „eigenen“ Proteine der beiden Organismen sind, soweit möglich, in die jeweiligen COG-Gruppen [90] eingeteilt worden, um einen raschen Überblick über die Protein-Funktionsklassen zu geben. Die Abbildungen 3.12 und 3.13 zeigen auf die einzelnen COG-Gruppen bezogen das normierte Verhältnis von „eigenen“ Proteinen zur Gesamtanzahl im Organismus. Hier wird auf den ersten Blick deutlich, dass in beiden Organismen bestimmte Gruppen häufiger in den spezifischen Proteinen als in allen Proteinen vertreten sind. In den Diagrammen 3.14 und 3.15 wird für jede COG-Gruppe gezeigt, ob sie in der Gruppe der „eigenen“ Gene über- oder unterrepräsentiert ist. Die überproportionale Repräsentation bestimmter COG-Gruppen in den spezifischen Proteinen konnte durch einen χ^2-Test statistisch nachgewiesen werden (Daten nicht gezeigt).

Die Abbildungen 3.14 und 3.15 zeigen deutlich, dass in den spezifischen Proteinen beider Archaeen die Gruppen L (*Replikation, Rekombination und Reparatur*),

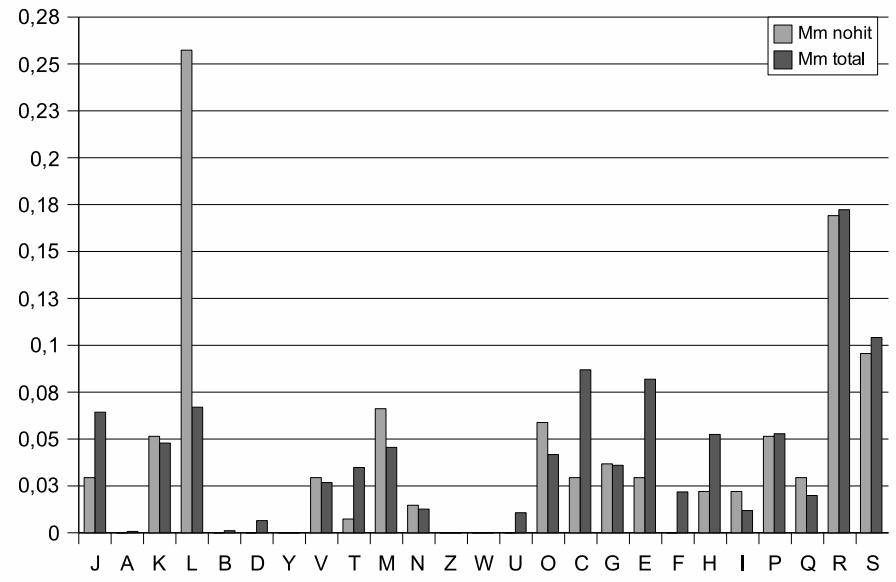

Abbildung 3.12.: *Ms. mazei* „eigene" Proteine (Mm nohit) und alle Proteine (Mm total) normiert. COG-Gruppen (genaue Erklärung s. Anhang A.3.1): **J..B** *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert.*

I (*Lipidtransport und -metabolismus*) und Q (*Biosynthese von Sekundärmetaboliten, Transport und Katabolismus*) übermäßig häufig vertreten sind. Bei *Ms. mazei* kommen noch die Gruppen M (*Zellwand-/Membran-/Hüllenbiogenese*) und O (*Posttranslationelle Modifikation, Proteinumsatz, Chaperone*) hinzu, die hier mindestens um den Faktor 1,25 häufiger auftreten als im Rest des Genoms. Dasselbe trifft auf die Gruppen G (*Kohlenhydratmetabolismus*), P (*Transport und Metabolismus anorganischer Ionen*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) im *Ms. acetivorans* Genom zu.

Betrachtet man nur *Ms. acetivorans*, so stellt sich die Frage, welche Funktion die 1.051 *Ms. acetivorans*-spezifischen Proteine haben (s. Tabelle 3.7). Die Abbildungen 3.14 und 3.15 zeigen, dass die COG-Gruppen L (*Replikation, Rekombination und Reparatur*), G (*Kohlenhydratmetabolismus*), I (*Lipidtransport und -metabolismus*), P (*Transport und Metabolismus anorganischer Ionen*), Q (*Biosynthese von Sekundärmetaboliten, Transport und Katabolismus*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) in den *Ms. acetivorans*-spezifischen Proteinen überrepräsentiert sind.

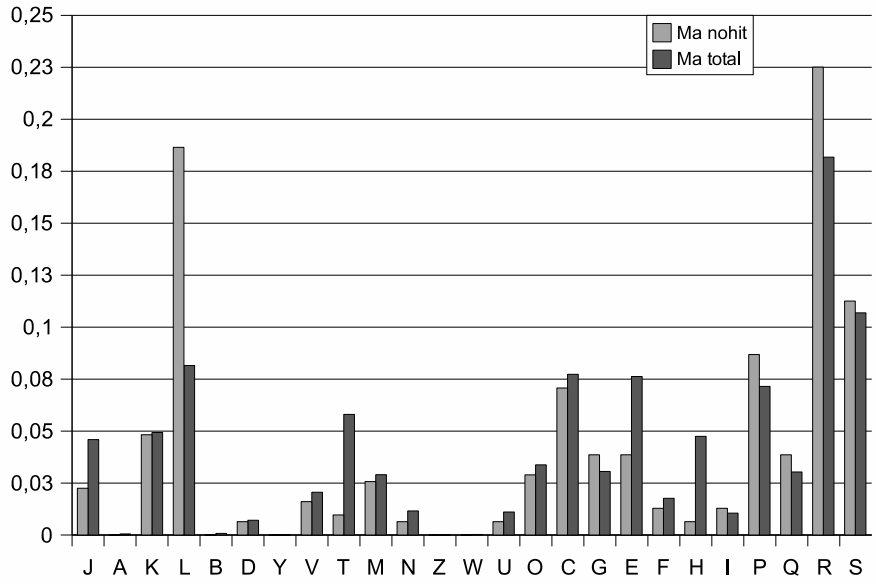

Abbildung 3.13.: *Ms. acetivorans* „eigene“ Proteine (Ma nohit) und alle Proteine (Ma total) normiert. COG-Gruppen (genaue Erklärung s. Anhang A.3.1): **J..B** *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert.*

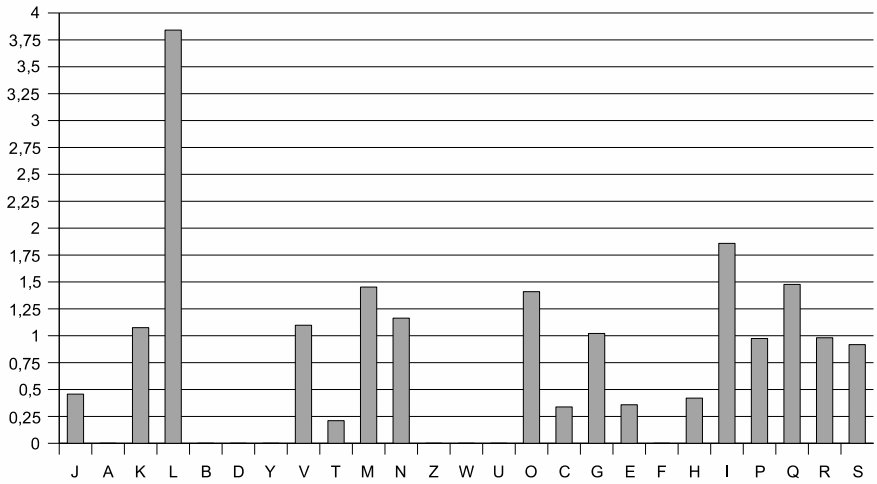

Abbildung 3.14.: Repräsentation der COG-Gruppen in *Ms. mazei*-spezifischen Proteinen.

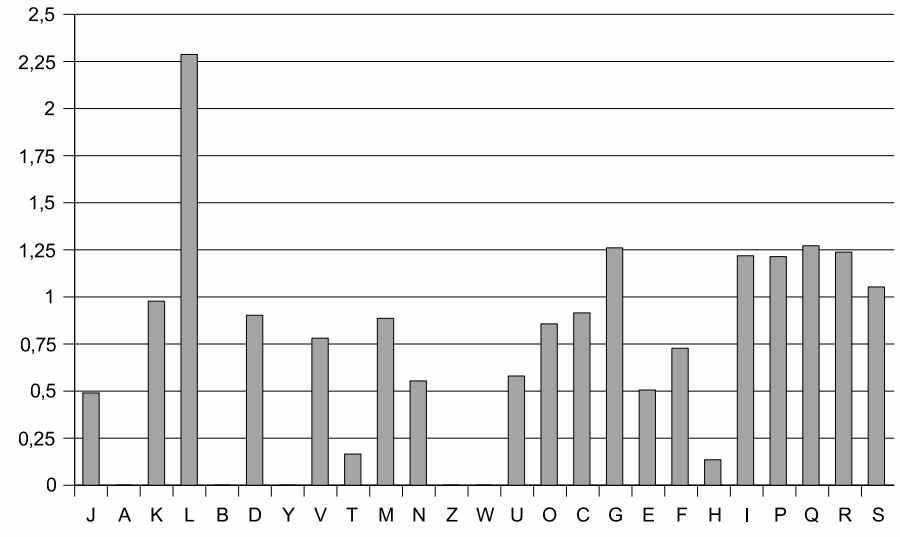

Abbildung 3.15.: Repräsentation der COG-Gruppen in *Ms. acetivorans*-spezifischen Proteinen.

Die Gruppe L gehört dabei zur Übergruppe „Informationsspeicherung und Verarbeitung“ und beinhaltet im Fall von Genominseln oder anderen auffälligen Bereichen Transposasen. Während die Gruppen I, P und Q zur Übergruppe „Metabolismus“ gehören, zählen die Gruppen R und S zur Übergruppe der „wenig charakterisierten“ Proteine, was die meisten hypothetischen und *predicted* Proteine einschließt. Ausschließlich hypothetische Proteine sind im vorliegenden Fall in der Gruppe Q zu finden. Es bleiben also I und P als interessante Gruppen übrig. In der Gruppe I sind 3 *Ms. acetivorans*-spezifische Proteine anzutreffen: eine Dehydrogenase, eine Phosphatase und eine Lipase. Die Gruppe P enthält 24 Proteine, von denen 8 hypothetisch oder *predicted* sind. Das grenzt die Anzahl auf 16 ein. Diese 16 Proteine enthalten regulatorische Proteine, Proteine eines Kaliumkanals, eine Superoxid Dismutase und eine Cytochrom-c Peroxidase. Zusammenfassend kann man sagen, dass nach der Auswertung der Proteine aus *Ms. acetivorans*, die kein homologes Protein in *Ms. mazei* haben, keine Aussagen über eine globale Funktion derselbigen gemacht werden können.

Die Gegenüberstellung der Proteome durch BLAST liefert umfassende Informationen über das proteinkodierte Potential eines Organismus oder einer Gruppe von Organismen. Die Zahlen des Vergleichs sind aber nur begrenzt aussagekräftig, da sie keinen Eindruck von der Qualität der Ähnlichkeit zwischen Proteinen vermitteln. Um einen Überblick über die Qualität der Protein-Alignments zu geben, wurden die beobachteten Angaben über identische Residuen im Proteinsequenz-Alignment

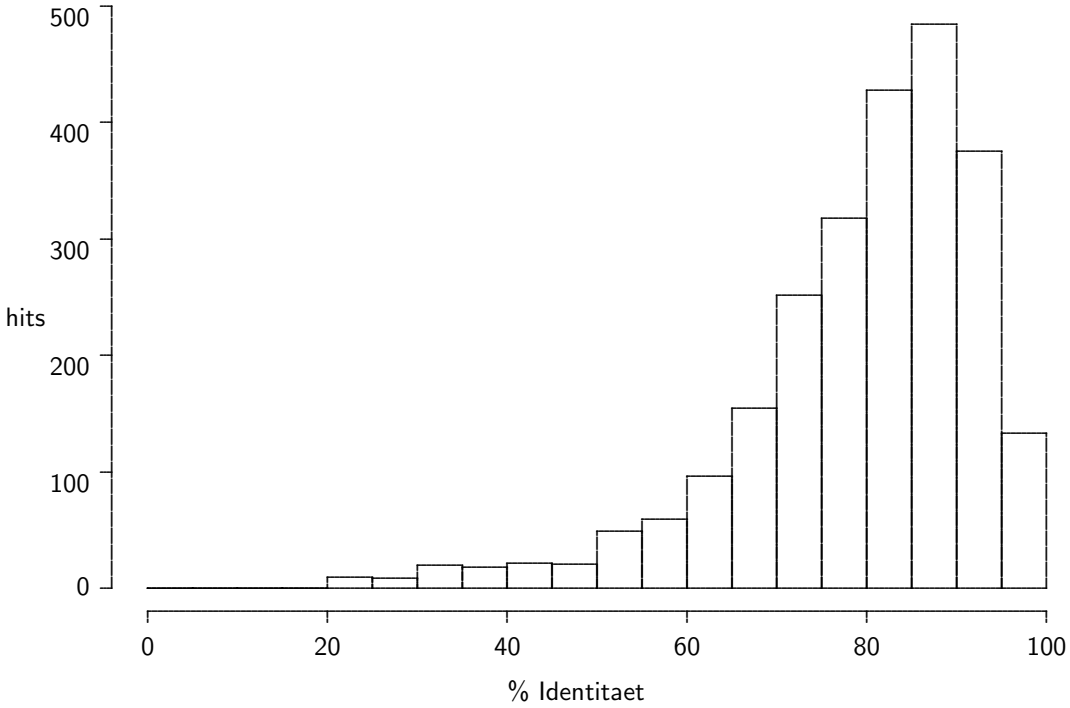

Abbildung 3.16.: Histogramm der Identitätswerte zwischen Proteinen aus *Ms. mazei* und *Ms. acetivorans*. Dargestellt ist der Anteil identischer Residuen der Proteinsequenz-Alignments in Prozent.

aus den BLAST-Suchen in einem Histogramm (s. Abbildung 3.16) abgebildet. Die Histogramm-Darstellung der Proteinidentitäten zeigt eine mittlere Proteinidentität von 82,42%. Die hohe durchschnittliche Identität kommt durch eine 100 prozentige Übereinstimmung großer Teile der Proteinsequenzen (letzter Balken) zustande.

Zusammenfassung

Der Vergleich der beiden Archaeen *Ms. mazei* und *Ms. acetivorans* verdeutlicht, dass die beiden Genome trotz ihrer offensichtlichen Unterschiede (s. Tab. 3.6) große bis sehr große Ähnlichkeit aufzeigen, sobald man das DNA-Niveau verlässt. Neben großen Bereichen von Synthenie (s. Abbildung 3.11) zeigen die beiden Genome auch Bereiche, in denen *Ms. acetivorans* offensichtlich zusätzliche DNA enthält. Die Proteome der beiden *Methanosarcinaceae* sind sich sehr ähnlich (s. Abbildung 3.16). Diese Ähnlichkeit drückt sich unter anderem in ähnlichen metabolischen Fähigkeiten aus.

3.2.3. Horizontaler Gentransfer in *Ms. mazei* - „first hit"-Methode

Der Annotationsprozess des *Ms. mazei* Genoms zeigte, dass ein großer Teil der Proteine dieses Organismus ihre höchsten Ähnlichkeiten nicht wie erwartet zu der Domäne der *Archaea*, sondern zu der der *Bacteria* aufweisen (Treffer zum Genus *Methanosarcina* selber wurden dabei außer Acht gelassen).

Im Genomsequenzierungsprojekt *Ms. mazei* wurde das Programm ERGO [74] zur Annotation der Proteine verwendet. In diesem Programm findet sich zu jedem einzelnen Protein eine Liste mit Proteinen aus anderen Organismen (Ähnlichkeitslisten), die eine gewisse Ähnlichkeit zu dem zu untersuchenden Protein aufweisen.

Die Beobachtung, dass Proteine aus der Domäne der *Bacteria* sehr häufig in den Listen ähnlicher Proteine zu finden waren, führte zu einer genaueren Untersuchung dieser Ähnlichkeitslisten im Annotationsprogramm ERGO. Zu diesem Zeitpunkt (Mitte 2001) waren insgesamt 258 Genome verfügbar (17 Archaeen, 83 Eukaryoten und 158 Bakterien). Von 3.371 Proteinen, die in *Ms. mazei* identifiziert werden konnten, zeigten 1.043 (ca. 31%) ihre höchsten Ähnlichkeiten (*e-value* Schwellenwert von $1 \cdot 10^{-5}$) in der Domäne der *Bacteria* [18]. Bei 544 (51 %) dieser 1.043 Proteine finden sich die ähnlichen Proteine ausschließlich im Reich der *Bacteria*. Diese Informationen über die Ähnlichkeitsverhältnisse eines Teils der Proteine aus *Ms. mazei* legten die Vermutung nahe, dass es sich um horizontal übertragene Gene handelt. Die Proteine, die die nachfolgenden Kriterien erfüllen, werden im Folgenden durch eine Art *first hit*-Methode als „bakteriell" bezeichnet:

- BLAST *e-value* $\leq 1 \cdot 10^{-5}$
- Unter den ersten 200 BLAST-Treffern sind nur solche in den *Bacteria* aufgeführt.
- Falls ein archaeelles Protein unter den ersten 200 Treffern ist, muss die Differenz in den Exponenten der *e-values* zwischen dem *Ms. mazei* Protein und dem ersten archaeellen Treffer $\leq$ 14 sein. Dieser Wert wurde während der Annotation empirisch ermittelt.

In Abbildung 3.17 sind die absoluten Zahlen *bakterieller* Gene in einigen archaeellen Genomen dargestellt (*first hit*). Das Genom von *Ms. acetivorans* war zum Zeitpunkt dieser Auswertung noch nicht veröffentlicht. Der prozentuale Anteil schwankt dabei zwischen 10 % in *T. acidophilum* und *Pb. aeorphilum* und 30 % in *Ms. mazei*.

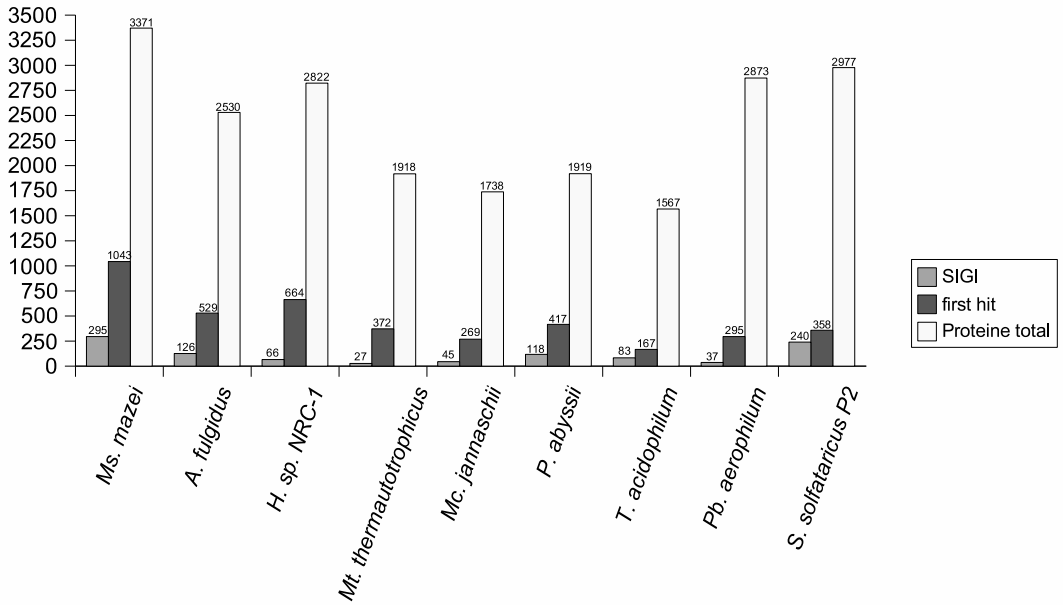

Abbildung 3.17.: Absolute Anzahl *bakterieller* Gene (*first hit*-Methode) und Fremdgene (Vorhersagen von SIGI) in ausgewählten archaeellen Genomen. M. *Methanosarcina*, A. *Archaeoglobus*, H. *Halobacterium*, Mt. *Methanothermobacter*, Mc. *Methanococcus*, P. *Pyrococcus*, T. *Thermoplasma*, Pb. *Pyrobaculum*, S. *Sulfolobus*.

3.2.4. Horizontaler Gentransfer II - Vorhersagen mit SIGI

In diesem Abschnitt wird der Gehalt an putativ horizontal transferierten Genen im Genom von *Ms. mazei* zusätzlich zur *first hit*-Methode mit einer weiteren Methode SIGI [62] untersucht.

SIGI basiert auf den Auswertungen der *codon usage* einzelner Gene und einer anschließenden Suche nach auffälligen Gen-Clustern. Die *codon usage* wird mit *log-odds-scores* bewertet, die aus dem Vergleich von ca. 400 mittleren Codon Häufigkeiten mikrobieller Genome abgeleitet wurden. In einem multiplen paarweisen Vergleich wird festgestellt, ob die *codon usage* des betrachteten Gens auffällig von der mittleren des Genoms abweicht. Folgen mindestens fünf solcher auffälliger Gene direkt aufeinander, wird diese Folge als Genominsel interpretiert. Diese Beschränkung in der Vorhersage auf Genominseln reduziert aus statistischen Gründen die Anzahl falsch positiver Treffer. Schließlich wird für jedes der putativen Fremdgene versucht, den möglichen Donor zu identifizieren. Diejenigen drei Spezies, deren *codon usage* im Rahmen der *log-odds-scores* zu den auffälligsten Bewertungen des betrachteten Gens führten, werden auf ihre taxonomische Verwandtschaft hin untersucht. Derjeni-

ge taxonomische Knoten, der die drei Spezies als „Kinder“ besitzt, wird als mögliche Herkunft des Gens vorhergesagt.

Eine Analyse des *Ms. mazei* Genoms ergab, dass 294 Gene vermutlich fremden Ursprungs sind und wahrscheinlich durch horizontalen Gentransfer erworben wurden. Für 52 dieser 294 putativen Fremdgenen sagt das Programm eine höchste Abstammungswahrscheinlichkeit aus die Domäne der *Bacteria* vorher. In Abbildung 3.17 wird eine Zusammenstellung der bisherigen Untersuchungen zu Fremdgenen in Archaeen gezeigt.

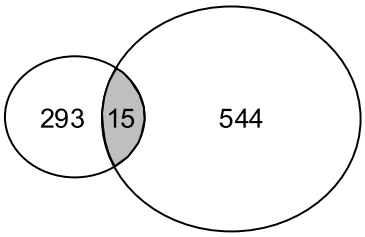

Abbildung 3.18.: Fremdgenvorhersagen bei *Ms. mazei* im Vergleich: links SIGI 293 Proteine, rechts *first hit* 544 rein bakterielle Proteine. 15 Proteine wurden von beiden Methoden als Fremdgene vorhergesagt.

Zusammenfassung

Die beiden verwendeten Verfahren zur Fremdgenvorhersage *first hit* und SIGI liefern sehr unterschiedliche Ergebnisse und haben nur eine kleine Schnittmenge von 15 Genen (s. Abbildung 3.18). Die Ursache hierfür ist die unterschiedliche Herangehensweise an die Fragestellung, ob ein Gen als fremd bezeichnet werden kann oder nicht. SIGI wählt hier eher einen konservativen Ansatz und beschränkt sich auf die Vorhersage von Genominseln.

3.3. Untersuchung des *Thermus thermophilus* HB27 Genoms

Thermus thermophilus HB27 [9] ist ein extrem thermophiles, Gram-negatives Bakterium, das eine hohe natürliche Kompetenz aufweist. Die Arbeiten mit *T. thermophilus* sind bis heute in über 1.000 Publikationen erschienen, was das Interesse an diesem Organismus verdeutlicht. *Thermus thermophilus* wird aufgrund von 16S-rRNA-Analysen mit *Deinococcus radiodurans* R1 zusammen phylogenetisch in die

Gruppe der *Deinococci* eingeordnet [98, 77]. Die beiden Bakterien *T. thermophilus* und *D. radiodurans* bewohnen extreme Lebensräume. *T. thermophilus* HB27 ist extrem thermophil und *D. radiodurans* R1 toleriert hohe Dosen radioaktiver Strahlung [37, 96]. Die Genomsequenzen der beiden Organismen sind verfügbar und machen so bioinformatische Untersuchungen möglich. Der Fokus soll dabei auf zwei Bereiche gelegt werden: auf den Proteomvergleich und die Suche nach möglicherweise durch horizontalen Gentransfer erworbenen Genen.

3.3.1. Bestimmung des Replikationsursprungs

In der Sequenz des Archaeons *Ms. mazei* gestaltete sich die Suche nach dem Replikationsursprung etwas schwierig (s. 3.2.1). Im folgenden wird am Beispiel *T. thermophilus* aus den *Bacteria* gezeigt, dass man mit der einfachsten Skew-Methode den Replikationsursprung ermitteln kann. Bei *Thermus thermophilus* HB27 liefert der schon vorgestellte kumulative GC-Skew gute Ergebnisse, die grafisch in Abbildung 3.19 dargestellt sind und verweist direkt auf den *Origin of Replication*. Das Minimum der GC-Skew Funktion weist auf den Replikationsursprung bei ca. 1.525 Mbp und das Maximum demzufolge auf den Terminus. Zwischen den beiden Punkten liegen ca. 50 % der Genomsequenz, was zeigt, dass sich *Origin of Replication* und der Terminus im Genom genau gegenüber befinden. Auf den ersten Blick sieht man, dass hier eine weitere Untersuchung mit anderen Skews nicht notwendig ist. Die Gene im Bereich um den vermuteten Replikationsursprung in *T. thermophilus* sind in Abbildung 3.20 zu sehen. Die Zahlen bezeichnen Proteine (s. Tabelle 3.8), die in diesem Bereich auffällig sind und deren Auftreten in dieser Region eventuell mit dem Replikationsursprung im Zusammenhang steht. Hierbei seien auch Proteine wie Pili erwähnt, die bei natürlicher Kompetenz eine Rolle spielen, da sie offensichtlich mit der DNA-Aufnahme im Zusammenhang stehen.

Möglicherweise ist ihre Translation oder ihre Translationskontrolle mit der Replikation der DNA gekoppelt. Der vermutete Replikationsursprung liegt zwischen den Genen TTC1608 und TTC1609 (Bereich 1.524.340 .. 1.524.642 bp). Hier befindet sich ein 300 bp langer intergener Bereich, der einen auffallend niedrigen GC-Gehalt von 50 % aufweist.

In dieser intergenen Region sind elf typische Bindemotive des bakteriellen Replikationsproteins DnaA [85] zu finden (s. Abbildung 3.21).

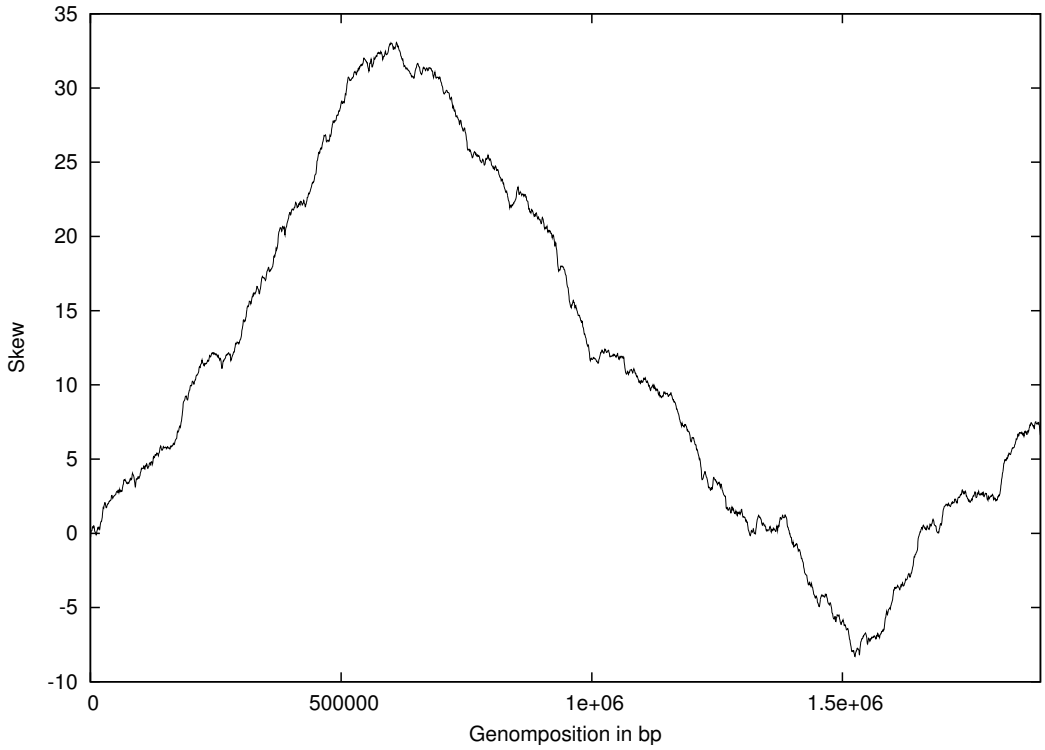

Abbildung 3.19.: Kumulativer GC-Skew von *T. thermophilus*.

Zusammenfassung

Der Replikationsursprung des *Thermus thermophilus* HB27 Genoms ist mit dem kumulativen GC-Skew eindeutig nachzuweisen und zeigt auf einen Bereich zwischen den Genen TTC1608 und TTC1609. Hier sind klassische DnaA-Bindemotive zu finden, die den bakteriellen Replikationsursprung charakterisieren. Außerdem ist der GC-Gehalt dieses intergenen Bereichs mit 50 % auffällig niedriger als der mittlere GC-Gehalt von 69,4 %.

3.3.2. Vergleich mit *Deinococcus radiodurans* R1

Die schon erwähnte gemeinsame phylogenetische Einordnung von *T. thermophilus* und *D. radiodurans* lässt vermuten, dass auch ihre Sequenzen ein gewisses Maß an Ähnlichkeit aufweisen. Es bietet sich an, wie im Fall von *Ms. mazei* und *Ms. acetivorans*, einen Genomvergleich vorzunehmen (s. 3.2.2), da die Komplettsequenzen beider Organismen verfügbar sind. Dieser Vergleich soll aus zwei Teilen bestehen, zum einen aus einem Genom-Alignment und zum anderen aus einer bidirektionalen BLAST-Analyse. In Tabelle 3.9 sind die Eckpunkte der Genome von

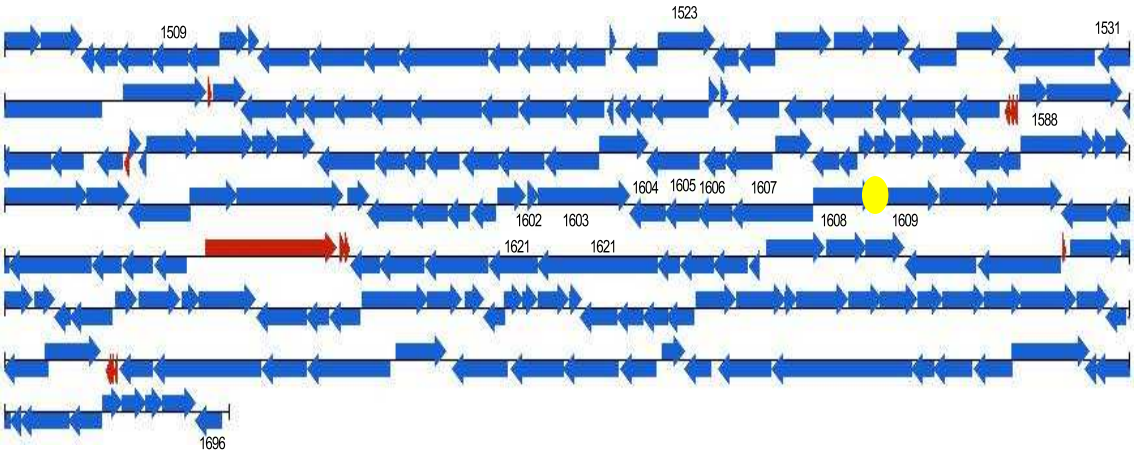

Abbildung 3.20.: Darstellung der Gene (blau) im Bereich um den vermuteten Replikationsursprung (gelbes Oval) im Genom von *T. thermophilus*. Rot: (t)RNA-Gene. Die Funktion der einzelnen Gene mit ihrer Genbank-Nummer können Tabelle 3.8 entnommen werden.

Tabelle 3.8.: Auswahl an Proteinen in der Umgebung des vermuteten Replikationsursprungs in *T. thermophilus*.

Proteinnummer	Annotation
TTC1523	ATPase associated with chromosome architecture/replication
TTC1531	excinuclease ABC subunit B
TTC1588	DNA polymerase III holoenzyme tau subunit
TTC1602	competence protein/comEA-related protein
TTC1603	competence factor comEC
TTC1604	chromosome partitioning protein parB
TTC1605	putative partitioning protein
TTC1606	glucose inhibited division protein B
TTC1607	glucose inhibited division protein A
TTC1608	chromosomal replication initiator protein dnaA
TTC1609	DNA polymerase III, beta chain
TTC1621	pili retraction protein pilT
TTC1622	ATP-binding motif-containing protein pilF
TTC1696	holliday junction DNA helicase ruvA

CCCCTGTGGATAACCTGTGGATAACCCTGTGGATAACCGGGCTCAAGCTGTGGATAACCC
TGTGGATAACCTGTGGAAAACTCCGGGGGGGCCGACCCCTGTGGATAACCCGCACTTTAT
CCACAGGTTATCCACAGGCCGGGGCCAGTTATCCACAGCAGTTATCCACAGCCCGTTTCT
CGTCCTGGAGGGCGTTTTCCCCCGCTTATCCACATTTCCACAGGCCCTACTACTACGACT
ACGGATCTTTTAAAAGTTTTAAAAAGAGAGTCGTAATAGCCGTTAAAGGAGGCGTCG

Abbildung 3.21.: Sequenz des Replikationsursprungs aus *T. thermophilus* HB27. Die DnaA-Bindemotive sind unterstrichen [85].

T. thermophilus HB27 und *D. radiodurans* R1 gegenübergestellt. Bei diesem Vergleich fällt sofort auf, dass *D. radiodurans* insgesamt ca. 1 Mbp mehr Sequenzinformation besitzt, was gleichzeitig bedeutet, dass *D. radiodurans* ca. 1.000 Proteine mehr als *T. thermophilus* enthält. Der Vergleich von *T. thermophilus* HB27 und *D. radiodurans* soll einen Überblick darüber geben, wie nah verwandt die beiden Bakterien sind.

Tabelle 3.9.: Vergleichender Überblick über die Genome von *T. thermophilus* HB27 und *D. radiodurans* R1.

	T. thermophilus	***D. radiodurans***	**Differenz**
Genomgröße (bp)	2.217.482	3.284.156	1.066.674
Chromosomen (bp)	1.894.877	2.648.638 und 412.348	
Megaplasmide (bp)	232.605	177.466	
Plasmide (bp)		45.704	
Proteine	2.218	3.182	964

Genom-Alignment

Das Genom-Alignment von *T. thermophilus* und *D. radiodurans* wurde zuerst mit MUMmer [15] auf DNA-Ebene durchgeführt. Dabei werden die Sequenzen der *Thermus thermophilus* Replikons mit den Chromosomen und Plasmiden aus *D. radiodurans* aligniert. Im Ergebnis sind die Übereinstimmungen auf DNA-Ebene sehr gering, und es kann keinerlei Synthenie festgestellt werden (Daten nicht gezeigt). Aus diesem Grund wird wie schon beim Vergleich der beiden *Methanosarcinaceae* (s. 3.2.2) eine Methode verwendet, die auf bi-direktionalen BLAST-Treffern basiert (s. 2.11.1). Das Ergebnis dieses BLAST-Vergleichs der *T. thermopilus*-Genoms mit den *D. radiodurans* ist in den Abbildungen 3.22 und 3.23 zu sehen, wobei ein Kreuz in der Darstellung einen bi-direktionalen BLAST-Treffer mit einem *e-value* $\leq 1 \cdot 10^{-5}$ zwischen den beiden Proteinen an dieser Position entspricht. Hier sieht man deutliche Unterschiede zum Vergleich der *Methanosarcina* Spezies (s. 3.11). Eine genomweite Synthenie oder eine konservierte Genanordnung in größeren Bereichen ist nicht erkennbar.

Die Alignments zwischen dem *T. thermophilus*-Chromosom und den *D. radiodurans*-Plasmiden ergeben nur vereinzelte Treffer. Ähnliches gilt für die Vergleiche des *T. thermophilus* Megaplasmids, hier gibt es mit keinem der *D. radiodurans*-Replikons mehr als vereinzelte Treffer. Auffällig ist allerdings, dass es zwischen

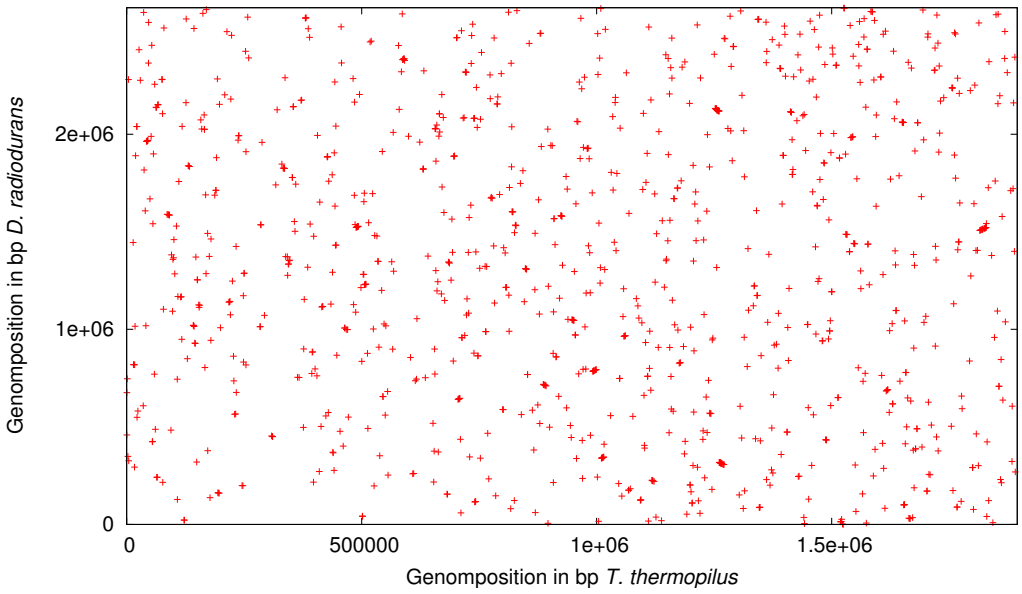

Abbildung 3.22.: Genomplot *T. thermophilus* gegen *D. radiodurans* Chromosom 1. Ein Kreuz in der Grafik entspricht einem bi-direktionalen BLAST-Treffer an dieser Position.

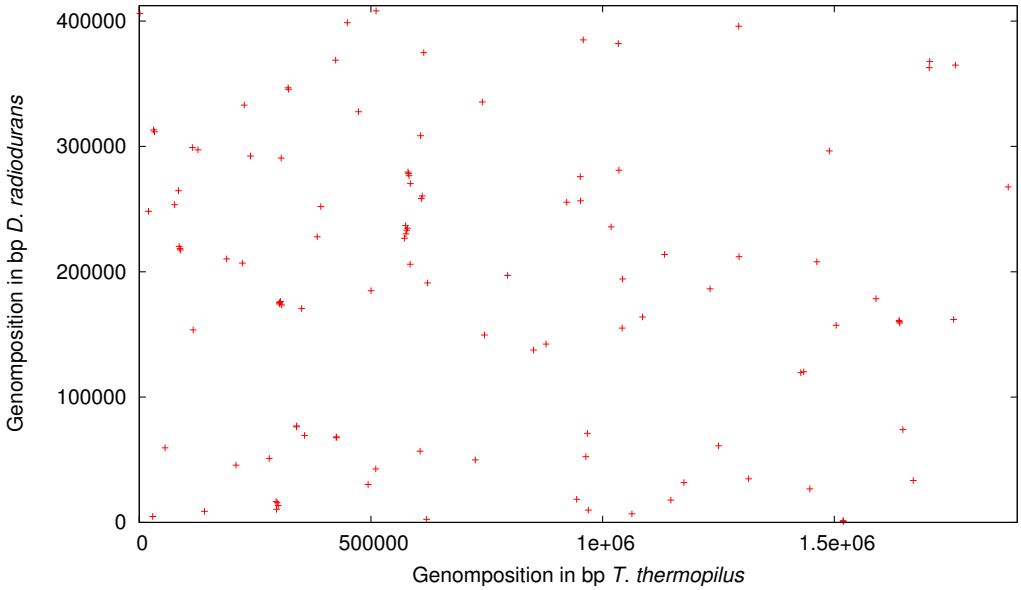

Abbildung 3.23.: Genomplot *T. thermophilus* gegen *D. radiodurans* Chromosom 2. Ein Kreuz in der Grafik entspricht einem bi-direktionalen BLAST-Treffer an dieser Position.

dem *D. radiodurans* Plasmid CP1 und dem *T. thermophilus*-Megaplasmid pTT27 keinen einzigen BLAST-Treffer mit einem *e-value* $\leq 1 \cdot 10^{-5}$ gibt. Zwischen dem *D. radiodurans*-Plasmid CP1 und *T. thermophilus* gibt es nur 6 Treffer auf dem Chromosom von *T. thermophilus*.

Eventuell erfüllt das Plasmid CP1 in *D. radiodurans* einen speziellen Zweck, denn 19 von 39 Proteinen auf dem Plasmid sind hypothetischer Natur und 9 Proteine sind Transposasen.

Vergleich der Proteome

Das Genom-Alignment von *T. thermophilus* und *D. radiodurans* (s. Abbildung 3.22) hat keine Synthenie zwischen den beiden Genomen gezeigt. Die in den Abbildungen 3.22 und 3.23 grafisch dargestellte Auswertung der bi-direktionalen BLAST-Analyse zeigt, dass der gemeinsame Gen-Pool von *T. thermophilus* und *D. radiodurans* nur 1.196 Proteine beträgt (s. Tabelle 3.10). Als *T. thermophilus*-spezifisch oder „eigen“ kann man nach dem bi-direktionalen BLAST 780 Protein bewerten. Dasselbe gilt für 1.508 *D. radiodurans*-Proteine.

Tabelle 3.10.: *T. thermophilus* und *D. radiodurans* Proteome im Vergleich. Die zweite Spalte gibt die Anzahl der Proteine wider, die sich durch bi-direktionalen BLAST als Homologe herausgestellt haben. Spalte 3 zeigt an, wie viele Proteine für den jeweiligen Organismus spezifisch sind, also kein Homolog besitzen. Die letzte Spalte enthält die Proteine, die nicht als solche vorhergesagt wurden (s. Diskussion). Die Addition von Spalte 1 und 2, ergibt nicht die Gesamtzahl der Proteine in dem entsprechenden Organismus, weil die Proteine, die zwar einen BLAST-Treffer haben, der aber nicht bi-direktional ist, nicht mitgezählt werden.

	Homologe Proteine (%)	**Organismus-spezifische Proteine („hypothetisch“ %)**	**Regionen von `tblastn` gefunden**
T. thermophilus	1.196 (54)	780 (390/50 %)	19
D. radiodurans	1.196 (38)	1.508 (881/58 %)	4

Mit `tblastn` können 19 *T. thermophilus*- bzw. 4 *D. radiodurans*- Proteine identifiziert werden, die nicht als solche vorhergesagt wurden und somit mittels `blastp` nicht gefunden werden können. Die Kriterien, die bei der `tblastn`-Suche angelegt werden, sind unter 3.2.2 beschrieben. Die Tatsache, dass in dem einen Genom bestimmte Sequenzen als ORFs vorhergesagt werden und dem anderen nicht, lässt

sich mit unterschiedlichen *ORF-findern* oder unterschiedlichen Annotationskriterien erklären. So kann es sein, dass ein Protein im einen Organismus wieder aus der Liste der Proteine entfernt wird, weil sich keine Ähnlichkeiten zu bereits bekannten Organismen ergeben. Wenn neue Genome veröffentlicht werden, können sich aber Ähnlichkeiten zu dem entsprechenden Protein neu ergeben. In diesem Fall wäre das in Frage kommende Protein wohl nie gelöscht worden.

Eine Aufstellung über die jeweils spezies-spezifischen Proteine ist im Anhang A.6 und A.9 zu finden. Bemerkenswert ist bei diesem Vergleich der relativ hohe Anteil an hypothetischen Genen in den jeweils spezifischen Proteinen der beiden Bakterien. Die gemeinsamen Proteine von *T. thermophilus* und *D. radiodurans* sind mit 1.088 Treffern zum größten Teil auf das *T. thermophilus*-Chromosom und das *D. radiodurans*-Chromosom 1 beschränkt. Chromosom 2 von *D. radiodurans* weist nur 113 bi-direktionale BLAST-Treffer, das Plasmid MP1 28 und das kleine Plasmid CP1 nur sechs Treffer auf. Ähnlich sieht die Situation des Megaplasmides pTT27 aus *T. thermophilus* aus. Hier finden sich die meisten Gemeinsamkeiten mit Chromosom 1 (46 Treffer) gefolgt von Chromosom 2 (15 Treffer) und Plasmid MP1 mit 24 Treffern. Zwischen dem kleinen *D. radiodurans*-Plasmid CP1 und dem *T. thermophilus* Megaplasmid konnten keine Gemeinsamkeiten gefunden werden.

Zur besseren Klassifizierung dieser Gruppe von spezifischen Proteinen wird eine Einordnung in die COG-Gruppen vorgenommen (s. Abbildungen 3.24 und 3.25). Die Balken der jeweiligen Gruppe stellen dabei das Verhältnis von „eigenen“ Proteinen zur Gesamtzahl von Proteinen dar.

Abbildungen 3.26 und 3.27 lieferen Informationen darüber, ob die jeweilige Gruppe in den „eigenen“ Proteinen über- oder unterrepräsentiert ist. Hier sieht man, dass in beiden Organismen die Gruppen L (*Replikation, Rekombination und Reparatur*), N (*Zellbeweglichkeit*), G (*Kohlenhydratmetabolismus*), P (*Transport und Metabolismus anorganischer Ionen*), Q (*Sekundärmetabolit Biosynthese, Transport und Katabolismus*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) überproportional vertreten sind. In *T. thermophilus* ist die COG-Gruppe C (*Energieproduktion und -umwandlung*) und in *D. radiodurans* die Gruppen K (*Transkription*) und V (*Verteidigungsmechanismen*) in den Spezies-spezifischen Proteinen häufiger vertreten bezogen auf alle Proteine.

Die bisherigen Daten über nicht vorhandene Synthenie, die Zahl der gemeinsamen oder eigenen Proteine der beiden Organismen und ebenfalls die Einteilung in die COG-Gruppen sagen nichts aus über die Qualität der Alignments beim Vergleich der

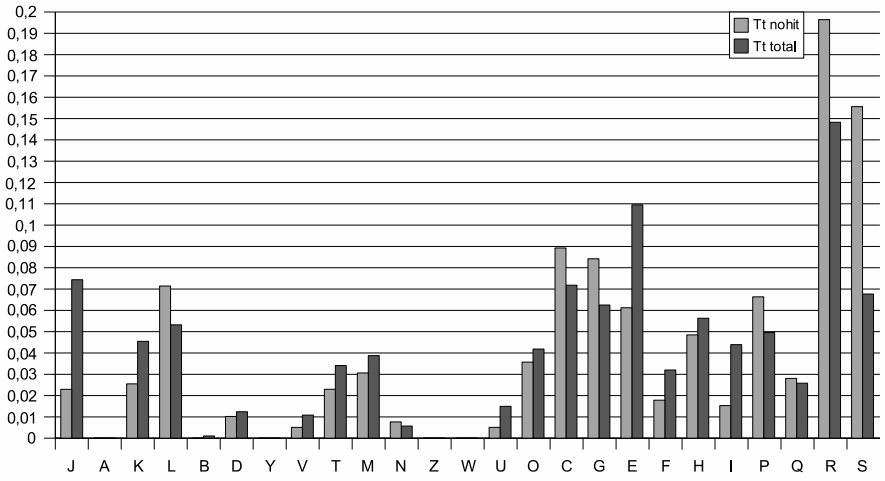

Abbildung 3.24.: ***T. thermophilus*** **„eigene" Proteine (Tt nohit) und alle Proteine (Tt total) normiert.** COG-Gruppen (genaue Erklärung s. Anhang A.3.1): **J..B** *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert*

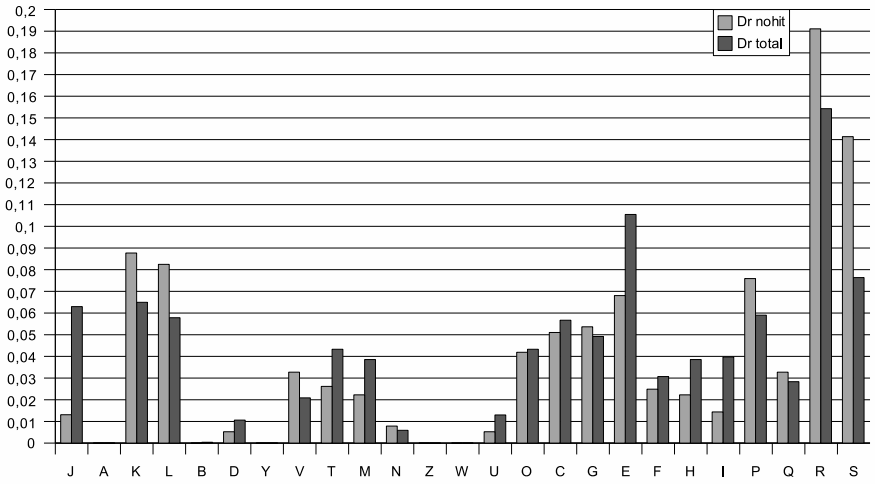

Abbildung 3.25.: ***D. radiodurans*** **„eigene" Proteine (Dr nohit) und alle Proteine (Dr total) normiert.** COG-Gruppen (genaue Erklärung s. Anhang A.3.1): **J..B** *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert*

gemeinsamen Proteine. Anders ausgedrückt wissen wir nichts über die Ähnlichkeit in Prozent der Proteome. Wie schon bei *Ms. mazei* und *Ms. acetivorans* (s. Abbildung 3.16) wird eine Histogramm Darstellung gewählt, um einen Einblick in die Qualität der Protein-Alignments zu geben (s. Abbildung 3.28). Abbildung 3.28 zeigt eine mittlere Proteinähnlichkeit zwischen *T. thermophilus* und *D. radiodurans* von 46 %. Betrachtet man andere thermophile Bakterien, so ergibt sich in einem Vergleich mit *T. thermophilus*, dass *Thermosynechococcus elongatus* BP-1 [69] mit 40 % Identität bei 714 gemeinsamen Proteinen nach *D. radiodurans* der nächste Verwandte auf Proteinebene ist. Danach folgen *Thermoanaerobacter tengcongensis* [3] mit 40 % Identität und 755 gemeinsamen Proteinen und *Thermotoga maritima* [71] ebenfalls mit 40 % Identität und 681 gemeinsamen Proteinen.

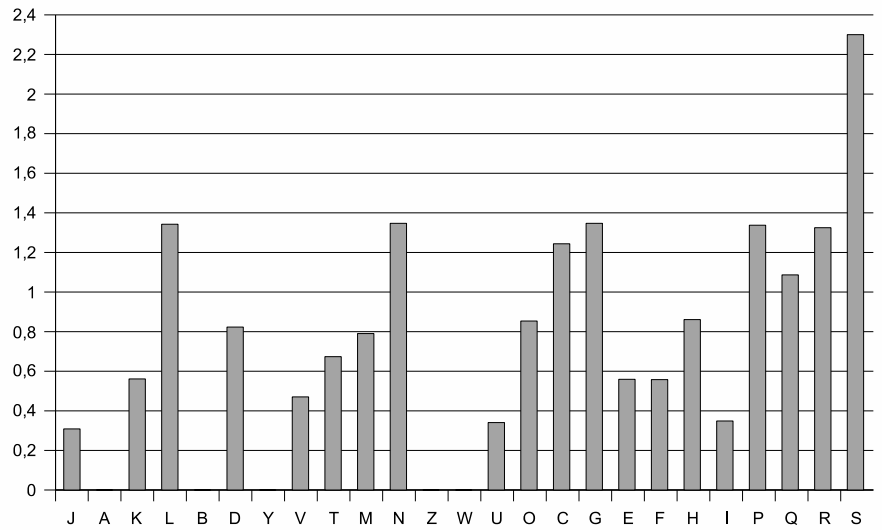

Abbildung 3.26.: Repräsentation der COG-Gruppen in *T. thermophilus*- spezifischen Proteinen

Zusammenfassung

Der Vergleich von *Thermus thermophilus* HB27 und *Deinococcus radiodurans* R1 hat gezeigt, dass die beiden Organismen zwar Gemeinsamkeiten auf Proteinebene besitzen, aber genauso viele Unterschiede. Die Alignments der Genome haben keine Synthenie gezeigt. Der gemeinsame Genpool beträgt dennoch 1.196 Proteine, was 54 % des *T. thermophilus* Proteoms entspricht und 38 % von *D. radiodurans*. Die mittlere Proteinidentität beträgt dabei nur 46 %. In den beiden Organismen *T. thermophilus* und *D. radiodurans* sind bestimmte COG-Gruppen innerhalb der

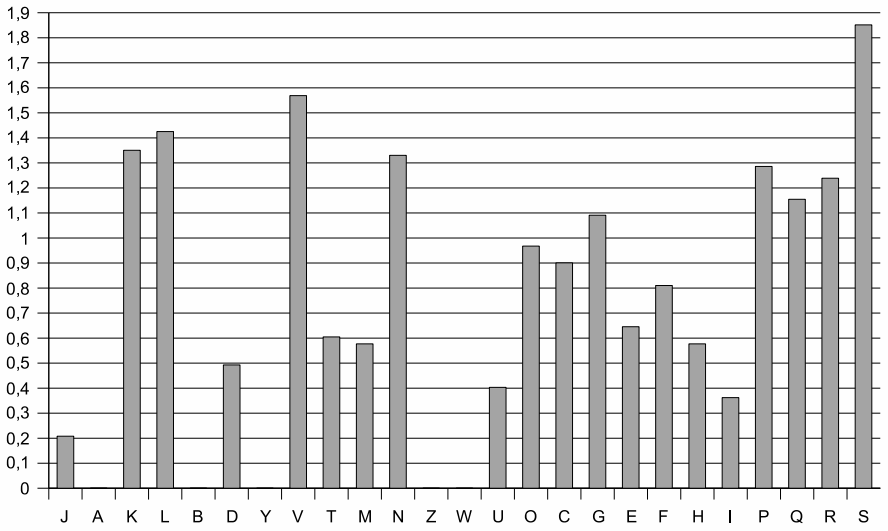

Abbildung 3.27.: Repräsentation der COG-Gruppen in *D. radiodurans*- spezifischen Proteinen

spezies-spezifischen Proteine überproportional vertreten und andere wiederum nur in einem der beiden Organismen.

3.3.3. Horizontaler Gentransfer in *Thermus thermophilus*

Thermus thermophilus zeichnet sich unter anderem durch seine natürliche Kompetenz aus. Das Bakterium ist somit in der Lage freie DNA aufzunehmen und in sein Genom zu integrieren [46]. Es stellt sich somit die Frage, ob über lange Zeiträume bzw. in der Evolution dieses Organismus aufgenommene DNA im Genom durch bioinformatische Mechanismen der Fremdgenerkennung zu detektieren ist. Des Weiteren ist interessant zu wissen, ob es sich um relativ viele Gene handelt, die als fremd erkannt werden, da Mechanismen zur DNA-Aufnahme existieren und *T. thermophilus* in dieser Hinsicht einen Vorteil gegenüber anderen Organismen hat. In der Literatur ist horizontaler Gentransfer in der Familie der *Deinococcaceae*, zu denen zum Beispiel *T. thermophilus*, *Meiothermus* und *D. radiodurans* gehören, beschrieben [72]. Die ATPase dieser Organismen ist vom A/V-Typ (TTC0905 .. 0913 in *T. thermophilus*) wie man ihn in den Archaea (A-Typ) oder in Eukaryoten (V-Typ) findet und nicht vom F-Typ wie es für die *Bacteria* typisch wäre. Dadurch, dass die Strukturen der beiden Hauptuntereinheiten A und B (TTC0906 .. 0907) der *T. thermophilus* ATPase archaeellen Charakters sind [72, 32], wird für diesen Fall ein horizontaler Gentransfer aus dem Reich der Archaea angenommen [38].

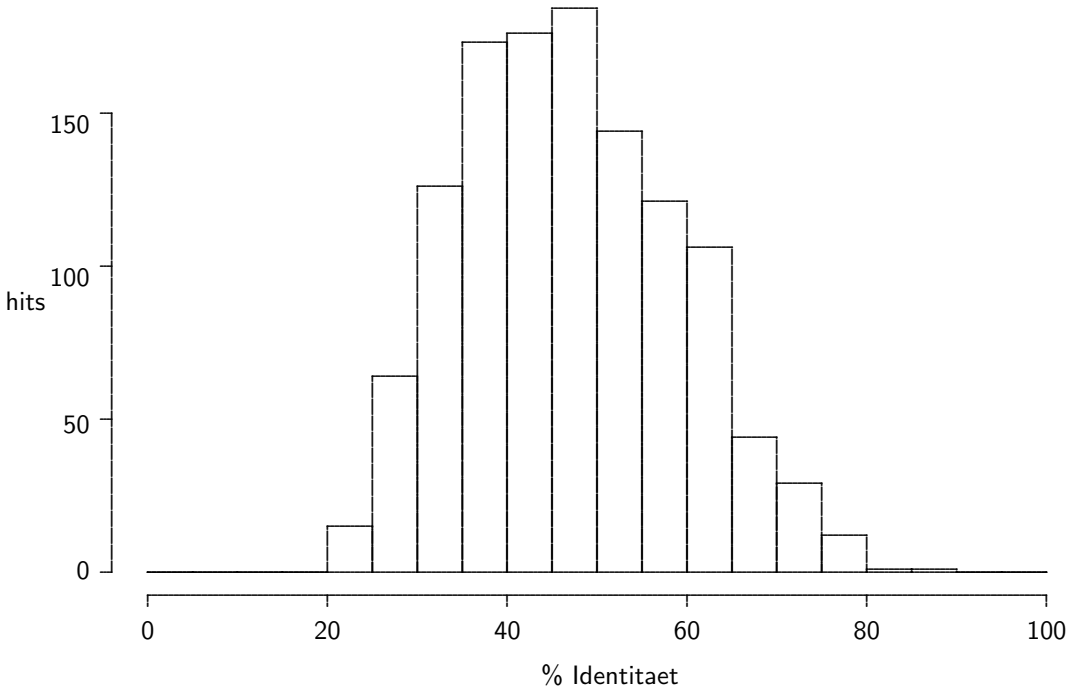

Abbildung 3.28.: Histogramm der Identität zwischen Proteinen aus *T. thermophilus* und *D. radiodurans*. Dargestellt ist die Identität der Aminosäure-Alignments in Prozent. Der Durchschnitt beträgt 46 %.

In diesem Abschnitt soll mit Hilfe des Fremdgenerkennungsprogramms SIGI untersucht werden, wieviele putative Fremdgene es in *T. thermophilus* gibt und welche Funktion diese ausüben. Die Suche nach Fremdgenen mittels SIGI hat ergeben, dass vermutlich 122 Gene (5,5 % aller ORFs) durch horizontalen Gentransfer erworben wurden. Von diesen 122 Genen sind 92 (75 %) als *hypothetical protein* oder *conserved hypothetical* annotiert. Von 92 hypothetischen Genen, die als fremd vorhegesagt wurden, sind außerdem 83 in der Gruppe der *T. thermophilus*-spezifischen Proteine zu finden, die kein homologes Protein in *D. radiodurans* besitzen. Insgesamt sind 99 putative Fremdgene in dieser Gruppe anzutreffen.

Um einen Überblick über diese Proteine und ihre Funktion zu bekommen, wurden sie mit dem COG-System in ihre jeweiligen Gruppen eingeordnet. 48 der 122 putativen Fremdgene konnten so funktionell gruppiert werden. Das Ergebnis ist in den Abbildungen 3.29 und 3.30 zu sehen. Letztere zeigt, welche COG-Gruppen in der Menge der Fremdgene überrepräsentiert sind. Dabei fallen einige Gruppen besonders auf: L (*Replikation, Rekombination und Reparatur*) und V (*Verteidigungsmechanis-*

men). Sie sind um den Faktor 6 häufiger in der Menge der Fremdgene vertreten als in allen Genen. Um den Faktor 1 bis 2 sind die Gruppen K (*Transkription*), D (*Zelluläre Kontrolle, Zellteilung Chromosomenaufteilung*), T (*Signaltransduktionsmechanismen*), M (*Zellwand- / Membran- Hüllenbiogenese*), U (*Intrazellulärer Verkehr, Sekretion und vesikulärer Transport*) und R (*Generelle Funktionsvorhersage*) überrepräsentiert. Eine Liste der als Fremdgene charakterisierten Proteine ist im Anhang A.7 zu finden.

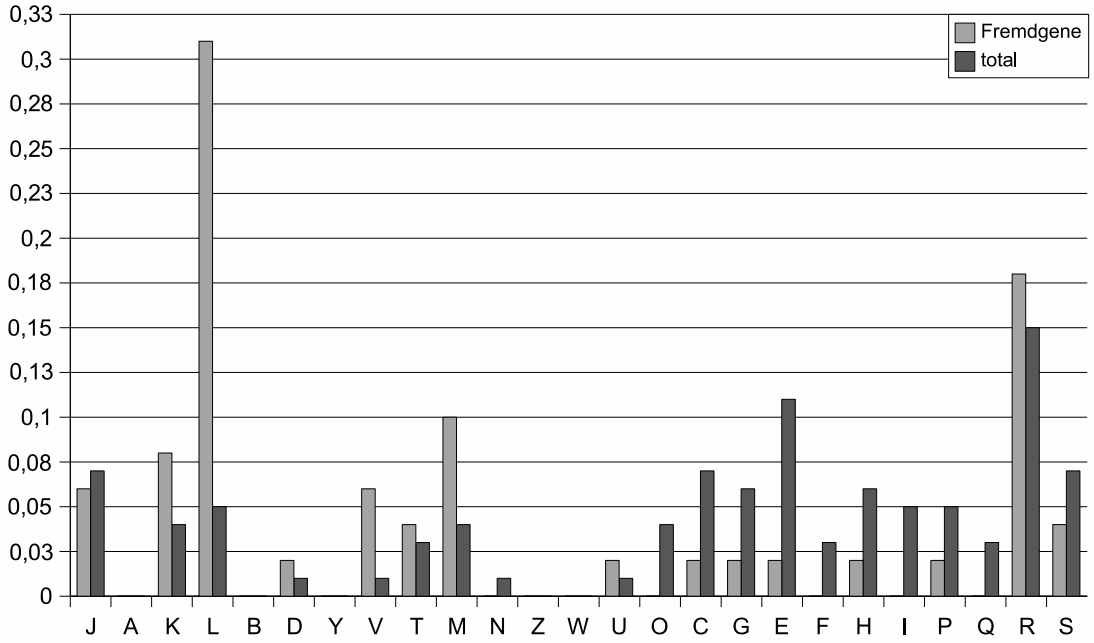

Abbildung 3.29.: *T. thermophilus* Fremdgene in die COG-Gruppen eingeteilt (s. Tabelle 3.2): J..B *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert.*

Zusammenfassung

SIGI findet 122 putative Fremdgene, was 5,5 % der *T. thermophilus* ORFs entspricht. Dieser Wert erscheint im Vergleich zu *Ms. mazei* (8,7%) sehr gering, was eigentlich wegen der Kompetenz von *T. thermophilus* nicht zu erwarten war. Die COG-Gruppen, die in den putativen Fremdgenen überrepräsentiert sind, waren für den Fall der Gruppe L zu erwarten, da L viele Transposasen und einige hypothetische Gene enthält, die gewöhnlich als fremd erkannt werden.

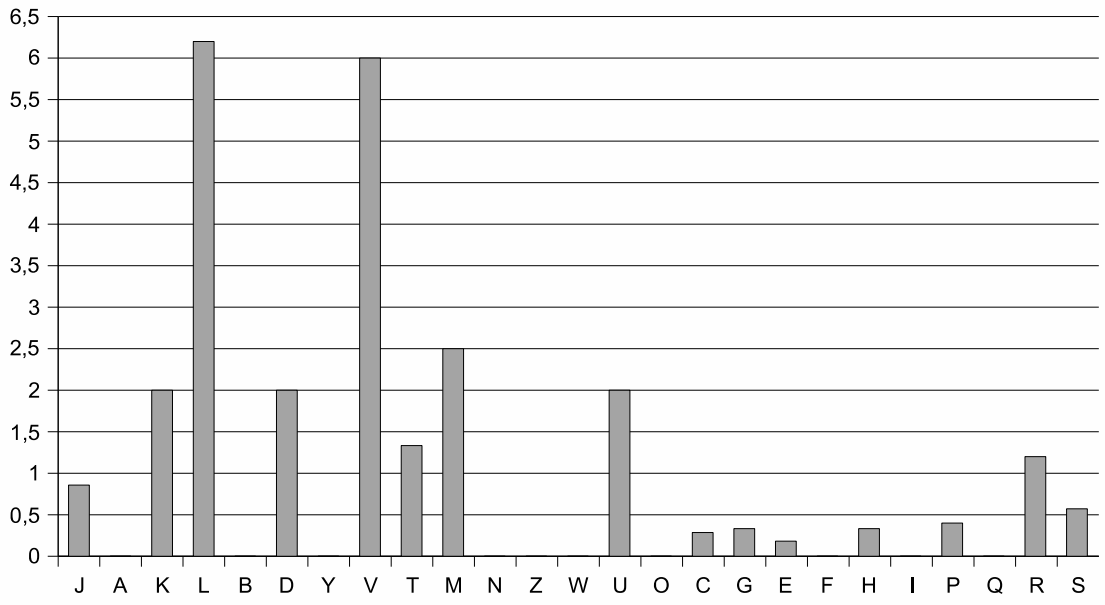

Abbildung 3.30.: Repräsentation der COG-Gruppen in putativen *T. thermophilus* Fremdgenen (s. Tabelle 3.2): J..B *Informationsspeicherung und -verarbeitung* **D..O** *Zelluläre Prozesse und Signalverarbeitung* **C..Q** *Metabolismus* **R..S** *dürftig charakterisiert.*

3.4. Programme

3.4.1. Quann

Nachdem die ersten größeren Contigs erstellt wurden, stellte sich die Frage nach dem „Inhalt“ der Sequenzen. Das bedeutet, dass die vorhandenen offenen Leserahmen vorhergesagt, die Proteinsequenzen abgeleitet und annotiert werden müssen. Für die Annotation kleiner Contigs (d. h. unter 200 kb) ist das Aufwand/Nutzen- Verhältnis einer Annotation mit einem komplexen Annotationsprogramm wie ERGO ungünstig. Für die Annotation in den frühen Stadien eines Genomsequenzierungsprojekts wurde im Rahmen dieser Arbeit Quann entwickelt. Quann steht dabei für *Quick annotation*. Das Programm führt eine vollautomatische Annotation von Sequenzen durch und beinhaltet folgende Arbeitsschritte:

- Finden von ORFs mit dem Programm GLIMMER [14]
- Übersetzen der DNA-Gensequenzen in Proteinsequenzen
- BLAST der Proteine gegen die jeweils aktuelle NCBI-Proteindatenbank (`ftp.ncbi.nih.gov/blast/db/nr.tar.gz`)

- Darstellung der Proteine und deren Annotation auf einer Webseite

Dieses Werkzeug ist relativ schnell und ermöglicht so einen guten ersten Einblick in die vorhandene Proteinausstattung. Es kann jedoch eine manuelle Annotation nicht ersetzen.

Details

Das Konzept von Quann besteht darin, den Arbeitsaufwand für die Annotation so gering wie möglich zu halten. Der Benutzer kann den kompletten Prozess auf einer einzigen Webseite, unter Angabe der gewünschten minimalen ORF-Länge, starten. Dafür ist keine Interaktion mit einem Administrator notwendig. Nachdem der Benutzer seine Sequenzen an das System übermittelt hat, wird der Annotationprozess gestartet. Nach Beendigung desselben wird der Nutzer *via* e-mail über das Ende informiert und kann dann seine Ergebnisse abrufen.

Im Hintergrund laufen die im Folgenden aufgeführten Prozesse. Nachdem die Sequenzen in das System geladen wurden, startet das *ORF finding* mit GLIMMER. Dabei wird der größte verfügbare Contig als Trainingsmenge benutzt. Die Koordinaten der Gene (Start- und Stoppositionen) werden in Dateien gespeichert und im nächsten Schritt ausgewertet. Das Programm übersetzt anhand der Genpositionen die übermittelte DNA-Sequenz in Proteinsequenzen und erstellt n *multiple fasta*-Dateien, wobei n frei einstellbar ist. Diese *fasta*-Dateien werden im nächsten Schritt auf n-BLAST-Prozesse aufgeteilt. Die Anzahl n ist dabei von der Anzahl an freien Computern („Klienten“) für die BLAST-Suche abhängig. Der eigentliche Annotationsvorgang besteht nun darin, die Proteinsequenzen auf den einzelnen Klienten mit der jeweils aktuellen Proteindatenbank des NCBI abzugleichen. Die ersten zehn BLAST-Treffer werden anschließend auf einer Webseite zusammen mit den Protein-Alignments dargestellt. Außerdem erhält der Benutzer einen Überblick über das Ergebnis der Annotation, der folgendes enthält: Anzahl vorhergesagter Proteine, Anzahl von Proteinen mit BLAST-Treffer und die Anzahl von Proteinen ohne BLAST-Treffer. Die Annotation selber wird Contig- oder proteinweise auf eigenen Seiten dargestellt. Und kann als *FASTA*-Datei zur weiteren Verarbeitung exportiert werden.

Zusammenfassung

Quann eignet sich zur einfachen, vollautomatischen Annotation zu einem frühen Zeitpunkt von Genomprojekten. Die durchschnittliche Contiglänge sollte dabei über 20 kbp liegen, da sonst die Trainingsmenge für den *ORF-finder* zu klein ist. Die Annotation erfolgt nach dem Start ohne Benutzerinteraktion. Zur endgültigen Annotation ist das Programm jedoch nicht geeignet, da keine manuelle Änderung der automatischen Annotation implementiert ist.

3.4.2. Cogger

Die Einordnung von Proteinen in COGs (*cluster of orthologous groups*) [90] erleichtert viele bioinformatische Aufgaben und ist von nicht unerheblichem Nutzen bei der Annotation eines Genoms. Das Konzept der COGs beruht auf der Annahme, dass sich eine Gruppe orthologer Gene (*COG*) aus einem Gen eines gemeinsamen Vorfahren durch Ereignisse der Artenbildung und -verdopplung entwickelt hat [90]. Die erste COG-Datenbank wurde dadurch erstellt, dass alle damals verfügbaren Proteinsequenzen (17.967) aus Genomen miteinander durch BLAST verglichen wurden. Aus diesen Ergebnissen wurde für jedes Protein der beste Treffer extrahiert und zu Minimal-COGs zusammengefasst. Ein Minimal-COG besteht dabei aus drei Proteinen, die untereinander die besten Treffer (Orthologe) repräsentieren. Bei dem Vergleich aller dieser Proteine mit BLAST wurde kein Schwellenwert gesetzt. Man geht davon aus, dass die Wahrscheinlichkeit, dass ein Gen eines der verglichenen Genome seinen besten Treffer in zwei anderen Genomen hat, höchst gering ist, es sei denn, es handelt sich wirklich um orthologe Gene.

Die COG Datenbanken werden vom NCBI (`www.ncbi.nlm.nih.gov/COG/`) gepflegt. Zur Einordnung neuer Proteine in die Datenbanken ist der Benutzer allerdings auf die Arbeit des NCBI angewiesen und somit nicht in der Lage, eigene Genomdaten einzuarbeiten (zu „coggen“). Es existiert zwar eine Webseite, mit deren Hilfe man einzelne Proteine coggen kann, sie basiert jedoch auf einer Datengrundlage von 1999. Eine neuere Webseite mit aktualisierten Datenbanken ist seit Monaten „under construction“.

Cogger wurde im Rahmen der vorliegenden Arbeit entwickelt und ist ein Programm, das es dem Benutzer ermöglicht, eigene Genomdaten (Proteinsequenzen) zu coggen, d. h. den einzelnen COGs zuzuordnen. Als Datengrundlage dient dabei die jeweilige COG-Datenbank `myva` (`ftp.ncbi.nih.gov/pub/COG/COG/`). Die Einord-

nung von Proteinsequenzen in die COGs mit Cogger erfolgt in mehreren Arbeitsschritten:

1. BLAST der Eingabesequenzen gegen die COG-Datenbank *myva*:
   `blastall -i input -p blastp -d myva -o output`

2. Verarbeitung der BLAST-Ausgabe: `cogger.pl output > output.cogs`

3. Zusammenfassung der Ergebnisse: `cogger-summary.pl output.cogs`

4. cogger-summary.pl erstellt außerdem eine erweiterte COG-Datenbank (`whog`), die in anderen Programmen weiterverarbeitert werden kann.

Das Programm cogger.pl verarbeitet die Ausgabe des BLAST-Abgleichs gegen die COG-Datenbank. Die BLAST-Ausgabe enthält dabei die Teile der Query-Sequenz, die ein Ergebnis bei dem BLAST-Vergleich mit der COG-Datenbank ergeben. Diese Proteine werden in einer weiteren Datenbank („whog“) in die einzelnen COG-Gruppen eingeteilt. Folgende Schritte werden durchgeführt, um eine Query-Sequenz in einen COG einzuordnen:

1. Die COG-Nummer des ersten Treffers (Treffer 0) in der COG-Datenbank und der dazugehörige Organismus werden gespeichert.

2. Alle nachfolgenden COG-Nummern und Organismen werden ebenfalls gespeichert.

3. Sind alle Treffer zu einer Query-Sequenz bearbeitet, wird untersucht, ob es Treffer zu mindestens drei verschiedenen Organismen mit immer demselben COG gibt.

4. Bei 3. werden nur die ersten n Treffer berücksichtigt, die im selben COG liegen wir Treffer 0.

5. Trifft 3. zu, wird die Query-Sequenz in den COG eingeordnet. Wenn nicht, wird kein COG zugewiesen.

Cogger-summary liefert eine Liste aller Proteine mit den dazugehörigen Gruppen. Die Ausgabe von cogger-summary.pl besteht aus mehreren Teilen. Dabei werden zuerst alle Proteine mit ihren COGs und den entsprechenden Gruppen ausgegeben. Zuletzt wird eine Übersicht über die COG-Gruppen mit der Anzahl an enthaltenen

Proteinen erstellt. Mit einem weiteren Programm (`makewhog_tng.pl`) kann man die COG-Ergebnisse eines neuen Genoms in eine vorhandene COG-Übersichtsdatei (`whog`) einfügen. Diese kann dann in anderen Programmen, z. B. SIGI, weiterverwendet werden.

Zusammenfassung

Die systematische Einordnung von Proteinen in Funktionsgruppen erleichtert die Arbeit mit großen Sequenzmengen (z. B. Genomen) erheblich. Das Einordnen in COGs ermöglicht darüber hinaus, Proteine als COGs zu erfassen, was wiederum die bioinformatische Analyse erheblich vereinfacht, da man nicht auf die Annotation der Proteine angewiesen ist.

3.4.3. secureBLAST

BLAST ist wegen seiner Geschwindigkeit und Sensitivität zu einer Standardmethode in vielen bioinformatischen Anwendungen geworden [1]. Die Suche in Sequenzdatenbanken, z. B. der am NCBI, über Webseiten (*Webinterfaces*) gehört zu einer der häufigsten Anwendungen im Bereich der Genomsequenzierung. Die Bedienung von BLAST über Webseiten ist sehr einfach und wird innerhalb der Forschergemeinschaft häufig genutzt. Diese Vorliebe vieler Benutzer wurde im Programmpaket secureBLAST [97] ausgenutzt, das im Rahmen dieser Arbeit entwickelt wurde. Das Programm besteht aus drei Teilen, die jeweils durch eigene Webseiten dargestellt werden:

- Hauptseite: hier wird die Query-Sequenz vom Benutzer eingefügt und die BLAST-Suche gestartet
- Update-Seite: hier kann der Benutzer die hinterlegten Sequenzen erneuern
- Administrations-Seite: hier können Benutzer angelegt/gelöscht, Passwörter gesetzt und Schreib-/Leserechte vergeben werden.

Mit secureBLAST ist es möglich, über *Webinterfaces* auf einfache Weise Sequenzdatenbanken zu erstellen und zu pflegen. Diese Datenbanken können von einzelnen Benutzern, die in einer Benutzerverwaltung abgelegt sind, selbst erstellt und erneuert werden. Sie dienen als Grundlage für BLAST-Suchen in den Sequenzen, die in

den jeweiligen Datenbanken gespeichert sind. Das Programm beinhaltet eine Benutzerverwaltung, die es einem festgelegten Administrator ermöglicht, Benutzer zu definieren und ihnen bestimmte Rechte zuzuweisen. So ist es möglich zu definieren, auf welche Sequenzen der Benutzer lesend oder schreibend zugreifen darf. Die Möglichkeit des Schreibzugriffs auf Sequenzdatenbanken im secureBLAST-System ermöglicht es dem Leiter eines Sequenzierprojektes jederzeit über einfach gestaltete Webseiten, die Sequenz seines Projektes in der Datenbank zu erneuern.

Der Vorteil eines solchen web-basierten BLAST-Systems mit Benutzerverwaltung liegt nicht nur darin, dass man keinen Administrator benötigt, der jede einzelne Sequenzdatenbank nach Absprache mit dem Projektleiter erneuert. Es besteht außerdem die Möglichkeit, Forschern außerhalb bestimmter Institutionen über Webbrowser durch Benutzernamen und Passwörter kontrollierten Zugriff auf einzelne Datensätze zu gewähren. Mitarbeiter, die auf der ganzen Welt verteilt sein können, sind so in der Lage, den Fortschritt eines Sequenzierprojektes zu verfolgen und können außerdem jederzeit im aktuellen Datensatz mit BLAST nach für sie interessanten Sequenzen suchen.

Zusammenfassung

Das hier skizzierte System zur einfachen Verwaltung von Sequenz- bzw. BLAST-Datenbanken ermöglicht in Kombination mit der vorgestellten Benutzerverwaltung einen sicheren und kontrollierten Zugriff auf Sequenzdatenbanken. Die Möglichkeit der Erneuerung der hinterlegten Sequenzen direkt durch den Mitarbeiter eines Sequenzierprojekts macht es möglich, immer den aktuellen Datensatz zur Sequenzsuche anzubieten. Das secureBLAST-System kommt mittlerweile im Genomsequenzierungsprojekt *E. blattae* und in mehreren anderen Sequenzierprojekten des Laboratoriums für Genomanalyse zum Einsatz. Die Benutzer des Systems sind bei diesen Projekten über ganz Deutschland verteilt.

4. Diskussion

Seit der ersten bakteriellen Genomsequenz von 1995 [26] hat die Anzahl vollständig sequenzierter Genome rapide zugenommen, so dass zum heutigen Zeitpunkt bereits 159 mikrobielle Genome öffentlich zugänglich sind (Quelle: Genbank, 19.4.2004). Hinzu kommen 186 unvollständige und eine unbekannte Anzahl nicht publizierter Genome. Unter den vollendeten Genomsequenzen befinden sich 141 Genome aus dem Reich der *Bacteria* und nur 18 aus dem der *Archaea*. Der wachsenden Menge an Sequenzinformationen, nicht nur aus Prokaryoten, wurde mit der Gründung der EBI Datenbanken [23] Rechnung getragen. Sieht man von der Aufgabe, die Sequenzdatenbanken zu pflegen, einmal ab, so ist die bioinformatische Auswertung von Genomdaten in den letzten Jahren eine große wissenschaftliche Herausforderung in der Biologie geworden. So stellt sich, nachdem eine Genomsequenz fertiggestellt ist, die Frage nach der biologischen Information, die in der Aneinanderreihung von A, C, T und G zu finden ist. Diese Arbeit zeigt anhand einiger Beispiele, wie die Sequenzierung eines Organismus und die anschließende Auswertung der Sequenzen aussehen kann. Die Ergebnisse werden im Folgenden diskutiert.

4.1. Genomsequenzierungsprojekt *Escherichia blattae*

Die vollständige Genomsequenz des Enterobakteriums *Escherichia blattae* soll es ermöglichen, den Organismus als Labor- bzw. Produktionsstamm für die biotechnologischen Industrie zu erschließen. *E. blattae* erfüllt dafür alle Voraussetzungen, wie z. B. die niedrigste Sicherheitseinstufung (S1) oder aerobes Wachstum bei 37 °C. So ist eine Anzucht im Großmaßstab ohne besondere Sicherheitsvorkehrungen möglich. Da der Organismus bei mesophilen Temperaturen wächst, müssen Fermenter nicht besonders stark geheizt oder gekühlt werden, was zu Energieeinsparungen führt. Der Hauptvorteil besteht aber darin, dass alle bekannten genetischen Werkzeuge aus *E. coli*, wie z. B. Expressionsvektoren, in *E. blattae* funktionieren. Die Sequen-

zierung ist weitgehend abgeschlossen, und in dieser Arbeit kann bereits eine erste Interpretation der Sequenzdaten gegeben werden.

4.1.1. Sequenzierung

In der Datenbank des Genomsequenzierungsprojektes *E. blattae* befinden sich zur Zeit ca. 52.000 Sequenzläufe, die in 43 Contigs mit insgesamt 4.115.744 sequenzierten Basen angeordnet werden konnten (s. Tabelle 3.1, Seite 41). Die Endphase der Genomsequenzierung gestaltet sich problematisch, da mittlerweile fast alle „Brückeninformationen“ ausgenutzt wurden. Das bedeutet, dass in einigen Bereichen des Genoms keine Informationen darüber vorhanden sind, wie die Contigs angeordnet werden müssen. Ein Grund hierfür ist ein nicht unerheblicher Anteil an Plasmiden aus den Genbanken, die offensichtlich misligiert sind. Das bedeutet, dass die Sequenz eines Plasmidinserts so nicht in der Sequenz des Organismus vorkommt. Solche Plasmide entstehen bei der Erstellung der Genbanken durch Ligation von DNA-Fragmenten, die im Genom so nicht anzutreffen sind. Man kann sich also auf die Informationen über die Länge der Plasmidinserts im Editierungsprogramm GAP4 (s. 2.9, Seite 25) nicht verlassen, so dass die Anordnung der Sequenzläufe in bestimmten Bereichen unklar ist. Deswegen muss vermehrt mit PCRs gearbeitet werden, die mit chromosomaler DNA als Matrize die Anordnung bestimmter Plasmide aufklären sollen. Dieses Verfahren kostet erheblich mehr Zeit als einfaches *plasmid walking* und verzögert so den Ablauf des Projekts.

4.1.2. Annotation

Zur kompakten Darstellung der bisher annotierten Proteine aus *E. blattae* wurde eine Aufteilung in die COG-Gruppen vorgenommen, deren Ergebnis in Tabelle 3.2 auf Seite 43 zusammengefasst ist. Diese Zahlen allein sagen allerdings recht wenig aus, so dass ein Vergleich mit einer Auswahl anderer Bakterien aus der Familie der *Enterobacteriaceae* durchgeführt wurde. Die Proteine dieser Organismen wurden ebenfalls in ihre COG-Gruppen eingeordnet. Die entsprechende Abbildung 3.2 auf Seite 44 zeigt, wie groß der Anteil aller Proteine in den einzelnen COG-Gruppen ist.

Die Proteine aus *E. blattae* sind in ungefähr den gleichen Verhältnissen auf die COG-Gruppen verteilt wie in den anderen Enterobakterien. In Abbildung 3.2 fällt lediglich auf, dass von *Yersinia pestis* und *Shigella flexneri* deutlich mehr Proteine in Gruppe L (*Replikation, Rekombination und Reparatur*) fallen (10 und 16 %) als

in den anderen untersuchten Enterobakterien (zwischen 4 und 6 %). Dieses Ergebnis deckt sich mit denen von Jin und Deng [16, 39]. Sie beschreiben in den Veröffentlichungen zu den Genomsequenzen von *Y. pestis* und *Sh. flexneri*, dass beide Organismen einen hohen Anteil an IS-Elementen besitzen. In *Sh. flexneri* sind 314 und in *Y. pestis* 111 IS-Elemente zu finden, was über den Gehalt an IS-Elementen anderer Enterobakterien hinausgeht.

4.1.3. Horizontaler Gentransfer in *Escherichia blattae*

Eine Suche nach Fremdgenen kann grundsätzlich anhand von zwei Methoden durchgeführt werden. Man unterscheidet dabei phylogenetische und parametrische Verfahren (Näheres zum Vergleich der Methoden befindet sich unter 4.2.3 auf Seite 95) [51]. Die erste Methode beruht auf dem Vergleich von Sequenzen und daraus resultierenden phylogenetischen Verwandtschaftsbeziehungen, und Letztere untersucht unabhängig von Sequenzvergleichen die Komposition der zu Grunde liegenden DNA. Dabei wird häufig die *codon usage* als ein Maß zur Charakterisierung der Zusammensetzung verwendet.

Für die Fremdgensuche in *E. blattae* wurde das Programm SIGI benutzt [62], für das ein eher konservativer parametrischer Ansatz gewählt wurde und das auf die Erkennung von Genominseln, d. h. Clustern von Fremdgenen, spezialisiert ist. Im vorliegenden Fall wurden in 3.980 Proteinen aus *E. blattae* 337 (8,5 %) putative Fremdgene gefunden. Der größte Teil dieser Fremdgene verteilt sich auf drei Bereiche, die als Prophagen identifiziert werden konnten (s. Tabelle 3.3, Seite 45). Untersucht man diese Genominseln anhand der Annotation ihrer Proteine genauer, so wird schnell klar, dass es sich hierbei um Bakteriophagen handelt. Die Untersuchung aller Fremdgene auf ihre Zugehörigkeit zu COG-Gruppen und die anschließende Berechnung des Verhältnisses zwischen der Anzahl von Fremdgenen und „normalen“ Genen in diesen Gruppen ist in Abbildung 3.4 auf Seite 47 gezeigt. Wie im Ergebnisteil unter 3.1.7 auf Seite 45 beschrieben ist, kann man hier erkennen, dass Fremdgene überproportional häufig in bestimmte COG-Gruppen fallen. Hier sind besonders die Gruppen L (*Replikation, Rekombination und Reparatur*) und M (*Zellwand-/ Membran-/ Hüllenbiogenese*) zu nennen. Mit einem χ^2-Test konnte gezeigt werden, dass Fremdgene in diesen Gruppen statistisch relevant überproportional vertreten sind. Bei genauerer Betrachtung der einzelnen Proteine bei dieser Einteilung wird dieses Unverhältnis klarer. Gruppe L beinhaltet fast ausschließlich Transposasen, Integrasen und

IS-Elemente. Diese Proteine sind hauptsächlich Bestandteile der Genominseln aus Prophagen. Die Gruppe L enthält neben einigen Glycosyltransferasen noch Galactosyltransferasen und andere Transferasen. Das Vorkommen dieser Transferasen in Genominseln wurde schon mehrfach beschrieben und ist ein bekanntes Phänomen [42, 41, 62]. Es wird angenommen, dass die Glycosyl- und Galactosyltransferasen durch Modifikation der Lipopolysaccharidschicht Gram-negativer Bakterien den Zellen einen Selektionsvorteil verschaffen.

Die weiteren COG-Gruppen K (*Transkription*), T (*Signaltransdukstionsmechanismen*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) sind in der Menge der Fremdgene schwach, aber dennoch überproportional vertreten. Die Gruppe K enthält fast ausschließlich regulatorische Proteine, R und S beinhalten einige Strukturproteine aus Prophagen und hypothetische Gene.

Zusammenfassend kann man sagen, dass bei der Fremdgenerkennung in *E. blattae* die integrierten Prophagen sehr gut erkannt wurden. Hierbei handelt es sich allerdings auch um das einzige mehr oder weniger verifizierbare Ergebnis einer Suche nach Fremdgenen. Das spiegelt sich auch in der Verteilung der Fremdgene auf die COG-Gruppen wider.

4.1.4. Vergleich mit anderen Enterobakterien

Aus der Gruppe der Enterobakterien sind bisher einige Genome vollständig sequenziert worden, so dass das *Escherichia blattae*-Genom mit denen anderer Enterobakterien verglichen werden konnte. Dieser Vergleich wurde anhand der Proteinsequenzen mit der Methode des bi-direktionalen BLASTs (s. 2.11.1, Seite 31) durchgeführt.

Aus den Vergleichen wurde deutlich, dass *E. blattae* zwischen 54 und 66 % seiner Proteine mit denen anderer Enterobakterien teilt (s. Tabelle 3.4, Seite 48), wobei *Buchnera aphidicola* wegen des kleinen Genoms nicht in die Vergleiche einbezogen wurde. Im Alignment der Proteinsequenzen von *E. blattae* und denen der anderen untersuchten Enterobakterien sind zwischen 68 und 74 % identische Residuen zu finden (s. Tabelle 3.4, Seite 48). Diese Ähnlichkeit ist im Vergleich zu der aus anderen Vergleichen (s. 3.2.2, Seite 56 und 3.3.2, Seite 68) relativ hoch. Bedenkt man allerdings, dass man mit *E. blattae* und *E. coli* zwei Arten desselben Genus untersucht, so wie es auch bei *Methanosarcina mazei* und *Methanosarcina acetivorans* der Fall ist (s. 3.2.2), dann erscheinen 74 % Identität im Alignment der Proteinsequenzen eher gering. Denn die beiden *Methanosarcina*-Spezies weisen eine mittlere Identität

von 82 % auf Proteinsequenzniveau auf (s. 3.2.2, Seite 58).

Aus Tabelle 3.4 wird allerdings auch deutlich, dass 2.619 Proteine aus *E. blattae* ein homologes Protein in *Salmonella typhimurium* besitzen. Die durchschnittliche Übereinstimmung der Residuen im Proteinsequenz-Alignment beträgt dabei 73 %. In der Tabelle (s. 3.4) nehmen die Übereinstimmungen der Proteinsequenzen und die Anzahl gemeinsamer Proteine zwischen *E. blattae* und *Salmonella enterica*, *Shigella flexneri*, *Yersinia pestis* und *B. aphidicola* kontinuierlich ab. Aus Sicht des Proteomvergleichs bleibt also festzuhalten, dass *E. blattae* die höchsten Ähnlichkeiten zu *E. coli* und *S. typhimurium* aufweist, obwohl *S. typhimurium* in einer anderen Gattung eingeordnet ist. Phylogenetisch betrachtet ist *E. blattae* wahrscheinlich weder mit *E. coli* noch mit *S. typhimurium* näher verwandt.

Nachdem nun die Gemeinsamkeiten zwischen *E. blattae* und den anderen Enterobakterien besprochen wurden, sollte auch erwähnt werden, dass es eine Menge von 650 Proteinen gibt, die sich nur im Genom von *E. blattae* und nicht in denen der Vergleichsorganismen finden lassen. Ein Einblick in diese Gruppe von *E. blattae*-spezifischen Proteinen wird in Abbildung 3.5 auf Seite 49 gegeben. Aus dieser Abbildung geht hervor, dass bestimmte COG-Gruppen in den *E. blattae*-eigenen Genen überrepräsentiert sind. Bei diesen Gruppen handelt es sich um K (*Transkription*), L (*Replikation, Rekombination und Reparatur*), N (*Zellbeweglichkeit*), Q (*Biosynthese von Sekundärmetaboliten, Transport und Katabolismus*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*). Zusammenfassend lässt sich sagen, dass hier einige *E. blattae*-spezifischen Merkmale zu finden sind. Diese beschränken sich allerdings hauptsächlich auf Phagenproteine und hypothetische Proteine. Aus den wenigen spezifischen Proteinen mit bekannter Funktion kann auf keine spezielle Besonderheit von *E. blattae* geschlossen werden. Bedenkt man, dass 46,6 % der hier besprochenen *E. blattae*-spezifischen Proteine als Fremdgene erkannt wurden, so ist für diese Proteine mit dem Wissen über ihr alleiniges Vorkommen in *E. blattae* womöglich ein zweiter Hinweis auf ihre fremde Herkunft gefunden worden.

4.2. Untersuchung des *Methanosarcina mazei* Gö1-Genoms

Der Organismus *Methanosarcina mazei* ist biochemisch gut untersucht und wird nun auch bioinformatisch näher betrachtet. Die vorgestellten Untersuchungen zum

Replikationsursprung der Fremdgene und der Vergleich mit *Ms. acetivorans* sollen einen Einblick in diese Untersuchungen geben.

4.2.1. Der Replikationsursprung

Der Replikationsvorgang prokaryotischer Genome ist komplex und erfordert die Interaktion einer Reihe von Proteinen, um die zeitlich koordinierte Vervielfältigung eines Chromosoms zu gewährleisten. Hier werden in einem ersten Schritt der Replikation die beiden Stränge der DNA entwunden und voneinander getrennt, so dass die Replikationsenzyme an den beiden Einzelsträngen mit der Replikation der DNA beginnen können. Die Sequenz dieses Bereichs ist allerdings kaum konserviert.

Bei der Annotation von *(Eury)archaeota* fällt auf, dass die Proteine für die Transkription, Translation und DNA-Replikation mehr Ähnlichkeit zu eukaryotischen als zu bakteriellen Proteinen aufweisen [94, 33]. Der archaeelle Replikationsapparat enthält allerdings nur einen Bruchteil der Proteine, die bei Eukaryoten anzutreffen sind. Das trifft jedoch nicht auf die Gruppe der *Crenarchaeota* zu. In den Genomen dieser Gruppe lassen sich keine Histon- und RepA-ähnlichen Proteine finden, wie es in den *Euryachaeota* der Fall ist. Einen Überblick über die Initiation der Replikation gibt Abbildung 4.1:

1. Bindung der *Initiator Elemente* am *Origin of Replication*
2. Entwinden der doppelsträngigen DNA
3. Binden von zwei Helikase Molekülen
4. Weiteres Entwinden der DNA durch Anlagerung der *sliding clamps*
5. Start der Replikation durch die DNA-Polymerase

Die archaeelle DNA-Replikation unterscheidet sich von der bakteriellen dadurch, dass die beteiligten Proteine nur äußerst geringe Sequenzähnlichkeit zur bakteriellen Replikationsmaschinerie aufweisen [33]. Stattdessen sind die archaeellen Replikationsproteine näher mit denen der Eukaryoten verwandt. Aus diesem Grund wurde anfangs angenommen, die archaeelle DNA-Replikation sei eine Minimalversion der eukaryotischen [22, 73]. Die nahe Verwandtschaft zwischen archaeellen und eukaryotischen Replikationsproteinen legt die Vermutung nahe, dass auch in Archaeen multiple Startpunkte der Replikation vorhanden seien. Dieses konnte in der Tat

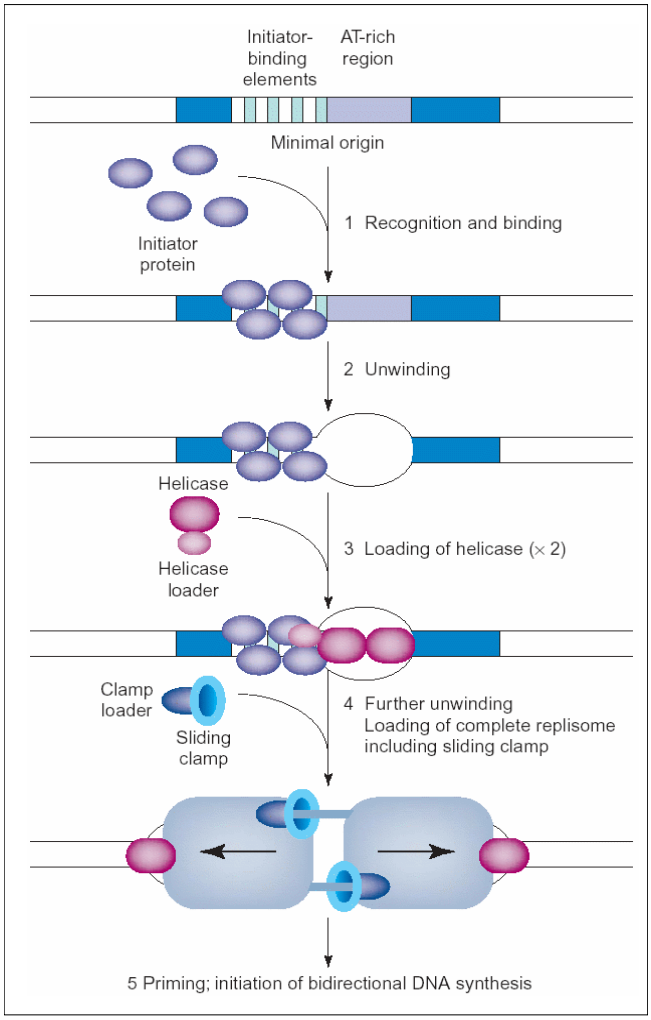

Abbildung 4.1.: Schematische Darstellung der Replikationsinitiation. [5]

für *Halobacterium sp.* NRC1 und kürzlich auch für *Sulfolobus solfataricus* gezeigt werden [100, 80].

Die *in silico*-Bestimmung des *Origin of Replication* ist bei den *Bacteria* kein allzugroßes Problem, da Lobry [52] mit dem GC-Skew und später Grigoriev [35] mit den kumulativen Skews Verfahren gefunden haben, die im Normalfall direkt zu der Region führen, in der die DNA-Replikation beginnt. Bei den *Archaea* sind die bisherigen Methoden allerdings häufig nicht anzuwenden, da sie nicht wie Bakterien über eine Präferenz von G vor C auf dem codierenden Strang verfügen [66]. Unter 3.2.1 wird die Kombination mehrerer Skew-Methoden beschrieben, die letztendlich zu einem Bereich im Genom von *Ms. mazei* führt, der mit hoher Wahrscheinlichkeit als der Replikationsursprung anzusehen ist. Hierfür sprechen mehrere Anzeichen,

z. B. der niedrige GC-Gehalt oder die unmittelbare Nähe des Cdc6-Proteins [43]. SALZBERG [81] beschreibt außerdem die Nähe bestimmter Replikationsfaktoren, wie DNA-Polymerasen oder Helikasen, zum *Origin of Replication*. Diese räumliche Nähe konnte auch in *Ms. mazei* gezeigt werden (s. Abbildung 3.10 und Tabelle 3.5). Hinweise auf die genaue Position des Origin wurden durch die unterschiedlichen Skews (s. Abbildungen 3.6, 3.8 und 3.9) gegeben. Die Ergebnisse zum Replikationsursprung von *Ms. mazei* stimmen überein mit später publizierten Ergebnissen von ZHANG aus [99]. Abschließend bleibt zu sagen, dass die unterschiedlichen Skew-Methoden in dem untersuchten Archaeon gute Hinweise auf die Umgebung des Replikationsursprungs gegeben haben. Es war jedoch eine Kombination aus mehreren Verfahren notwendig, um ausreichende Evidenz zu erhalten. Die genaue Position des Replikationsstarts konnte letztendlich aber nur durch Betrachtung der Proteine in der Nachbarschaft des *Origin of Replication* festgelegt werden. Die Suche nach Repeats (Daten nicht gezeigt) in der vermuteten *Origin*-Sequenz bestätigte letztendlich die Position des *Ori*. Die Lokalisation bestimmter Proteine (s. Tabelle 3.5, Seite 55), die an der Replikation beteiligt sind, in der Nähe des *Origin of Replication* scheint zumindest teilweise konserviert zu sein [81].

4.2.2. Vergleich mit *Ms. acetivorans*

Aus der Familie der *Methanosarcinaceae* sind bisher zwei Genome vollständig sequenziert worden: *Ms. mazei* Gö1 [18] und *Ms. acetivorans* C2A [31]. Ihre Genome sind mit 4.096 und 5.751 Mbp erstaunlich groß, vergleicht man die Genomgröße mit der anderer methanogener Archaeen (s. Tabelle 4.1), deren Genome nur ca. 1,7 Mbp umfassen.

Tabelle 4.1.: Genomgrößen vollständig sequenzierter methanogener Archaeen. M.j. [11], M.k. [87], M.t. [88], Ms.m. [18], Ms.ac. [31]

Organismus	Genomgröße (Mbp)
Methanococcus jannaschii	1,665
Methanopyrus kandleri	1,695
Methanobacterium thermoautotrophicum	1,751
Methanosarcina mazei	4,096
Methanosarcina acetivorans	5,751

Betrachtet man die Lebensräume und die metabolischen Fähigkeiten dieser Methanogenen, so ist anzumerken, dass alle, außer der beiden *Methanosarcina*-Spezies

thermophil sind und in einer relativ engen ökologischen Nische leben. Sie sind außerdem auf die Methanproduktion durch $4H_2 + CO_2 \rightarrow CH_4 + 2H_2O$ beschränkt. Die Spezies der Gattung *Methanosarcina* sind hingegen mesophil und können aus den weiteren Substraten Acetat, Methanol und Methylaminen Methan produzieren [56]. Sicherlich ist für dieses zusätzliche metabolische Vermögen auch ein gewisses Maß an zusätzlicher genetischer Ausstattung notwendig. Es ist aber fraglich, ob dafür ca. 2,3 Mbp (*Ms. mazei* - andere Methanogene: 4,096 Mbp - 1,70 Mbp) notwendig sind. Bei geschätzen 1000 bp pro Gen entspräche das immerhin ungefähr 2.300 zusätzlichen Proteinen.

In der vorliegenden Arbeit soll untersucht werden, worin der Unterschied zwischen den beiden *Methanosarcinaceae* liegt, der sich durch das viel größere *Ms. acetivorans*-Genom andeutet. In Tabelle 3.6 auf Seite 56 sind die offensichtlichen Unterschiede den beiden Genomsequenzen aufgelistet. Der auffälligste Unterschied sind die schon erwähnten 1,655 Mbp, die das *Ms. acetivorans* Genom größer ist, als das von *Ms. mazei*. Damit einher gehen 1.169 zusätzliche Proteine.

Welche Funktionen üben die *Ms. mazei*-spezifischen Proteine aus? Diese Frage soll mit Hilfe der COG-Gruppen Einteilungen geklärt werden (s. Abbildungen 3.14 und 3.15). Wie schon erwähnt fallen hier die Gruppen L (*Replikation, Rekombination und Reparatur*), I (*Lipidtransport und -metabolismus*), Q (*Biosynthese von Sekundärmetaboliten, Transport und Katabolismus*), M (*Zellwand-/ Membran-/ Hüllenbiogenese*) und O (*Posttranslationelle Modifikation, Proteinumsatz, Chaperone*) durch besonders häufiges Auftreten in den *Ms. mazei* eigenen Proteinen auf. Untersucht man die Annotation der Proteine dieser Gruppen, so zeigt die Gruppe Q 4 *Ms. mazei*-spezifische Proteine, eine Isochorismutase, eine Sulfatase und eine putative Methyltransferase sowie 9 Transferasen und Epimerasen sind in der Gruppe M zu finden. Zusammenfassend statten diese Proteine *Ms. mazei* im Vergleich zu *Ms. acetivorans* nicht mit besonderen Fähigkeiten aus.

Der in Abbildung 3.11 auf Seite 57 gezeigte Genomplot der beiden Genome gibt einen ersten Eindruck über die Ähnlichkeit der beiden Genome. Die Graphik stellt ein globales Genom- Alignment der beiden Organismen basierend auf einer bidirektionalen BLAST-Analyse dar. In der Abbildung ist relativ große Synthenie erkennbar, was an einer über große Bereiche sichtbaren Diagonalen deutlich wird. Bei ca. 2 Mbp und 5 Mbp auf der X-Achse sind zwei Inversionen zu erkennen. In der Region um den *Origin of Replication*, der in der Abbildung am Anfang und am Ende der Diagonalen aus BLAST-Treffern angesiedelt ist, weisen die beiden Genome eine

konservierte Genanordnung auf. In der Mitte der Diagonalen des Genomplots (ca. 2..4,5 Mbp auf der X-Achse) findet sich der Replikationsterminus wieder und außerdem der Bereich der größten Unterschiede zwischen den beiden Genomen. Neben einzelnen homologen Proteinen ist in diesem Bereich keinerlei Synthenie sichtbar.

In der Region des Replikationsterminus finden sich 550 Gene, die spezifisch für *Ms. acetivorans* sind, d. h. im nah verwandten *Ms. mazei* nicht vorkommen (s. 3.2.2). Bei diesen 550 Genen handelt es sich um 52 % der insgesamt 1.051 für diesen Organismus spezifischen. Nimmt man für ein Protein eine durchschnittliche Länge von etwa 1000 bp an, so sind in diesem Bereich um den Terminus 500.000 bp *Ms. acetivorans*-spezifische „Proteine“ codiert, deren Funktion unbekannt ist. Man kann also festhalten, dass über die Hälfte der *Ms. acetivorans* eigenen Proteine im Bereich des Terminus angesiedelt ist.

Bei genauerer Betrachtung der jeweils nur in einem der beiden Vergleichsorganismen vorkommenden Proteine (Tabelle 3.7, Seite 59, „kein Treffer“) muss beachtet werden, dass die Zahlen jeweils um 61 bzw. 10 Proteine nach unten korrigiert werden müssten, was zu 364 *Ms. mazei*- und 1.041 *Ms. acetivorans*-spezifischen Proteinen führen würde. Diese Korrektur kann man vornehmen, da die in der letzten Spalte der Tabelle aufgeführten Proteine ein ähnliches Protein im jeweils anderen Organismus haben. Der Grund hierfür sind unterschiedliche Verfahren bzw. Schwellen für das Finden von ORFs. Im *Ms. acetivorans*-Genomprojekt wurde zum *ORF finding* das Programm GLIMMER benutzt, das für seine *overprediction* (d. h. es werden im Schnitt zu viele Gene vorhergesagt) bekannt ist [31, 91]. Im *Ms. mazei* Projekt kam hingegen CRITICA [2] zum Einsatz. Das Benutzen unterschiedlicher ORF-Vorhersage Programme kann bereits diesen Unterschied in der Anzahl erklären.

Abschließend bleibt festzuhalten, dass in *Ms. acetivorans* 912 hypothetische Proteinen zu finden sind, die kein homologes Protein in *Ms. mazei* haben. Daraus ergeben sich bei einer geschätzen durchschnittlichen Genlänge von 1.000 bp schon 912.000 bp. Das enspricht 61 % der 1,5 Mbp, die das *Ms. acetivorans* Genom größer ist als das von *Ms. mazei*. Weiterhin sind ca. 36 % der zusätzlichen Sequenz von *Ms. acetivorans* in hypothetischen Proteinen im Bereich des Terminus kodiert. Die überwiegende Menge an zusätzlicher genetischer Information würde man wohl als „Abfall“ bezeichnen. In der Region um den Terminus von *Ms. acetivorans* kann man einen großen Teil der zusätzlichen 1,5 Mbp Genomgröße wiederfinden. Der hohe Anteil an hypothetischen und lediglich vorhergesagten (*predicted*) Proteinen könnte

zwei Ursachen haben. Entweder befinden sich in dieser Region neuartige Proteine, die in keinem bekannten Organismus aufzufinden sind, was das Fehlen an Datenbankeinträgen erklären würde. Oder diese Region ist so etwas wie die „Spielwiese“ des *Ms. acetivorans*-Genoms, in der sich viele Transposasen und nicht funktionale Proteine, d. h. Artefakte der ORF-Vorhersage, tummeln. Auch beim Vergleich von *Pyrococcus abyssii* mit *Pyrococcus horikoshii* wurde beobachtet, dass die Region um den Terminus ein *hot spot* für Umlagerungen der Genome sind [68, 53]. Diese Umlagerungen konnten auch schon in der Domäne der *Bacteria* beobachtet werden [6].

Ungeklärt bleibt die Frage, warum sich *Ms. acetivorans* das zusätzliche genetische Material, das in 1,5 Mbp codiert ist, leisten kann. Sieht man von wenigen „echten“ und „sinnvollen“ Proteinen wie der ATP-Synthase einmal ab, dann verbleibt viel DNA, die bei dem Vergleich der Genome als überflüssig erscheint. Ein großes Genom benötigt auch mehr Energie zur DNA-Replikation und zur Reparatur. Eine mögliche Erklärung hierfür wäre der Unterschied im Lebensraum der beiden *Methanosarcina*-Spezies. *Ms. mazei* wurde aus einer Kläranlage bei 25 °C isoliert, wogegen *Ms. acetivorans* aus marinen Sedimenten (14 °C) stammt (K.R. Sower, G. Gottschalk, persönliche Mitteilung). Für *Ms. acetivorans* wurde sogar von einer minimale Wachstumstemperatur von 9 °C berichtet (G. Gottschalk, persönliche Mitteilung). Daraus ergibt sich eine um ca. 10 °C geringere Habitatstemperatur für *Ms. acetivorans*. Berücksichtigt man nun, dass die Fehlerraten in DNA mit steigender Temperatur zunehmen, so steigt z. B. die Desaminierung von Cytosin (C→A) um das 5,7 fache bei einer Temperaturerhöhung um 10 °C [30]. Nimmt man an, dass diese Zahlen auch *in vivo* stimmen, dann wäre das eine Erklärung dafür, warum *Ms. acetivorans*, das bei niedrigeren Temperaturen lebt als *Ms. mazei*, ein relativ großes Genom hat, ohne dabei bekannte physiologische Fähigkeiten aufzuweisen. Die niedrigen Fehlerraten bei der „Pflege“ seiner DNA geben *Ms. acetivorans* die Möglichkeit, bei langen Generationszeiten eine größeres Genoms zu unterhalten.

4.2.3. Horizontaler Gentransfer in *Ms. mazei*

Horizontaler Gentransfer (HGT) wird als einer der wichtigsten Faktoren in der prokaryotischen Evolution angesehen [44, 21, 20, 19]. Insbesondere in den *Archaea* wurde über HGT berichtet [44, 87], wobei hier anzumerken ist, dass in hyperthermophilen Organismen dieser Domäne weniger horizontal transferierte Gene gefunden

werden als in den mesophilen Organismen [87]. Zu Letzteren zählt *Ms. mazei* Gö1, in dessen Genom eine nicht unerhebliche Anzahl vermutlich horizontal aufgenommener Gene entdeckt wurde [18].

Die Suche nach Genen, die über horizontalen Gentransfer erworben wurden, wurde mit zwei unterschiedliche Methoden für die Suche Erkennung von Fremdgenen durchgeführt (s. 3.2.3, Seite 64). Es handelt sich zum einen um eine Abwandlung der s. g. *first hit*-Methode und zum anderen um das Programm SIGI [62]. Die beiden unterschiedlichen Methoden können als phylogenetisch (*first hit*) und parametrisch (SIGI) bezeichnet werden [51]. Die erste Methode wird deshalb als phylogenetisch bezeichnet, weil sie sich auf den direkten Sequenzvergleich bezieht oder auf den Vergleich phylogenetischer Bäume, die mit Hilfe von Sequenz-Alignments hergestellt wurden. Dagegen beruht die parametrische Methode nicht auf Sequenzvergleichen, sondern auf der Suche nach Genen, die in ihrer Umgebung atypisch erscheinen. Untersucht werden dabei z. B. die Nukleotidzusammensetzung [49, 50], Di-Nukleotidfrequenzen [40] oder eine Codon-Usage-Verzerrung [67, 61, 65]. Diese Verfahren beruhen auf der *Genome Hypothesis*, die davon ausgeht, dass jedes Genom seine für sich typische *codon usage* aufweist [34]. Das in dieser Arbeit verwendete Programm SIGI nutzt ein *score* basiertes System aus Vergleichen mittlerer Codon-Usage-Werte.

Tabelle 3.17 gibt einen Überblick über die gefundene Anzahl von Fremdgenen in verschiedenen archaeellen Genomen. Die Vorhersage der *first hit*-Methode zeigt für *Ms. mazei* eine große Anzahl von Fremdgenen. Ein Grund hierfür könnte sein, dass die anderen Archaeen mit Ausnahme von *Halobacterium* alle thermophil sind und somit in einer relativ engen ökologischen Nische leben. Auf diese Weise haben sie nicht die Möglichkeit, fremde DNA aufzunehmen, weil in ihrem Habitat keine Fremd-DNA aus anderen Organismen vorhanden ist. *Ms. mazei* hingegen ist in seinem mesophilen Habitat von einer Vielzahl von Bakterien umgeben und hat somit ausreichend Fremd-DNA in seiner Umgebung. Ordnet man die als fremd eingestuften Proteine den COG-Gruppen zu (s. Anhang A.4), so wird offensichtlich, dass die meisten der 544 *bakteriellen* Proteine (13,7 %) in die Gruppe R (*Generelle Funktionsvorhersage*) fallen. Danach folgen die Gruppen L (10 %, *Replikation, Rekombination und Reparatur*), P (5,9 %, *Transport und Metabolismus anorganischer Ionen*), C (5,7 %, *Energieproduktion und -umwandlung*), V (5,5 %, *Verteidigungsmechanismen*) und M (4,5 %, *Zellwand-/ Membran-/ Hüllenbiogenese*). Diese Befunde decken sich mit denen von Koonin [57], der berichtet, dass bakterielle Gene in Archaeen hauptsächlich für metabolische Enzyme und Proteine kodieren, die in die

Zellwandsynthese und Zellteilung involviert sind.

Unter Verwendung von SIGI wurden deutlich weniger Proteine als Fremdgene vorhergesagt. Die 295 von SIGI vorhergesagten Fremdgene (8,7 %) ergeben eine ähnliche Verteilung wie die *first hit*-Methode, wobei die meisten Proteine in den COG-Gruppen L (12 %, *Replikation, Rekombination und Reparatur*), R (7,8 %, *Generelle Funktionsvorhersage*) und M (7,8 %, *Zellwand-/ Membran-/ Hüllenbiogenese*) zu finden sind.

Vergleicht man die Vorhersagen von SIGI und der *first hit*-Methode (s. Abbildung 3.18), so besteht die Schnittmenge der beiden Vorhersagen aus 15 Proteinen, von denen die meisten als *hypothetical proteins* annotiert sind. Es wird häufig beobachtet, dass phylogenetische und parametrische Verfahren unterschiedliche Ergebnisse liefern. Für *Thermotoga maritima* hat ein parametrisches 6,4 % und ein phylogenetisches Verfahren 25 % atypische Gene vorhergesagt [51]. Für so unterschiedliche Zahlen kann es mehrere Ursachen geben. Im Fall der *first hit*-Methode ist die Zusammensetzung der verwendeten Sequenzdatenbanken entscheidend. Im vorliegenden Fall waren z. B. nur 9 % der Genome, mit denen die Proteine aus *Ms. mazei* verglichen wurden, archaeell. Einige als bakteriell eingestuften Proteine könnten eventuell viel ähnlicher zu einem Archaeon sein, das aber noch nicht sequenziert wurde. Es gibt durchaus kritische Stimmen zur Verwendung von BLAST bei dieser phylogenetischen Art der Suche nach Fremdgenen [45]. Für die Ergebnisse bei *Ms. mazei* wurde allerdings eine abgewandelte Variante der beschriebenen Methode [45] verwendet, d. h. nicht nur der erste Treffer wurde betrachtet. Durch die Konstruktion phylogenetischer Bäume [18] für einen Teil der bakteriellen Proteinen aus *Ms. mazei* konnte außerdem gezeigt werden, dass diese mit denen aus Bakterien clustern. Dieselben Proteinen aus anderen Archaeen bilden hingegen einen eigenen Cluster, in dem keine Bakterien zu finden sind.

Der Vorteil der parametrischen Methoden, die z. B. in SIGI [62] benutzt werden, liegt darin, dass man nicht auf den Sequenzvergleich angewiesen ist. Diese Unabhängigkeit von Sequenzdatenbanken sorgt dafür, dass man nicht in dem Dilemma gefangen ist, das besagt, dass der nächste Verwandte womöglich noch gar nicht sequenziert wurde oder gänzlich unbekannt ist. Parametrische Verfahren beruhen darauf, dass unterschiedliche Genome unterschiedlichen Mutationsraten unterliegen. Daher ist es sehr schwierig, horizontal transferierte Gene aus solchen Organismen vorherzusagen, die demselben Mutationsdruck ausgesetzt sind wie das zu untersuchende Genom.

Die beiden *Methanosarcina*-Spezies besitzen innerhalb der *Archaea* die größten

Genome. *Ms. acetivorans* trägt wahrscheinlich einen großen Pool an „eigener“ DNA deren Funktion noch ungeklärt ist. Die Ergebnisse der Fremdgenerkennung decken sich mit denen von MAKAROVA [59]. Hier wurde berichtet, dass die *Methanosarcinaceae* womöglich einen großen Pool an horizontal übertragenen Genen besitzen, was durch die Auswertung der phylogenetischen Muster von COG-Analysen gefunden wurde.

4.3. Untersuchung des *Thermus thermophilus* HB27-Genoms

Mit der von *Thermus thermophilus* HB27 ist die erste vollständige Genomsequenz aus der Gattung *Thermus* verfügbar [37]. Die Auswertung des *T. thermophilus*-Genoms in dieser Arbeit enthält neben der Lokalisation des Replikationsursprungs, den Vergleich mit *Deinococcus radiodurans* R1 und die Suche nach möglichen Fremdgenen.

4.3.1. Lokalisation des Replikationsursprungs

Es existieren diverse Methoden zur *in silico*-Bestimmung des Replikationsursprungs in Prokaryoten (s. 3.3.1). Ihre Anwendung und nicht ganz einfache Auswertung bei den *Archaea* wurde am Beispiel von *Ms. mazei* beschrieben (s. 3.2.1). Die Suche nach dem *Origin of replication* in *T. thermophilus* HB27 (s. 3.3.1) gestaltete sich wenig problematisch, da die Anwendung des kumulativen GC-Skews (s. 3.2.1 und Abbildung 3.19) schnell zum Erfolg führte. Der kumulative GC-Skew [35] nutzt ebenso wie der nicht-kumulative GC-Skew [52] das häufigere Auftreten von Guanin-Basen im *leading strand* aus. Der Unterschied zwischen den beiden Skew-Varianten besteht darin, dass der kumulative Skew durch das Kumulieren der Werte nicht so empfindlich auf lokale Minima und Maxima reagiert. Der nicht kumulative GC-Skew stellt eine Funktion dar, die um eine Achse oszilliert und am *Ori* das Vorzeichen wechselt. Der kumulative GC-Skew hingegen ergibt eine Funktion, die am Replikationsursprung ihr Minimum und am Terminus ihr Maximum aufweist (s. Abbildung 3.19). Diese Darstellung macht die Interpretation aber vor allem dann deutlich einfacher, wenn die Skews, wie bei den meisten *Archaea*, nicht den klassischen Verlauf nehmen [35].

Bei den *Bacteria* kommt erleichternd hinzu, dass der *Origin of Replication* im Normalfall neben dem Gen für das Protein DnaA zu finden ist. Außerdem sind in den *Bacteria* im Bereich des *Origin of Replication* stark konservierte DnaA-Bindemotive und ein auffallend niedriger GC-Gehalt zu finden [63]. Der Abstand zwischen Minimum (*Ori*) und Maximum (Terminus) im GC-Skew entspricht ca. 50 % der Genomsequenz, was für einen bi-direktionale Replikation mit einem Ursprung spricht [35]. Mit Hilfe des kumulativen GC-Skews, dessen Graph bei 1,52 Mbp ein Minimum zeigt, konnte der *T. thermophilus* Replikationsursprung zwischen den Genen TTC1608 (DnaA) und TTC1609 (DnaN) ausfindig gemacht werden (s. Abbildung 3.20). In diesem intergenen Bereich ist der GC-Gehalt von 50 % auffällig und die DnaA-Bindemotive aus *T. thermophilus* HB8 [85] sind ebenfalls eindeutig nachzuweisen (s. Abbildung 3.21).

Betrachtet man die Region um den Ursprung in *Thermus thermophilus* HB27 genauer (s. Abbildung 3.20), so sind dort neben DnaA und DnaN einige Proteine (s. Tabelle 3.8) anzutreffen, die ebenfalls eine Rolle bei der Replikation spielen, z. B. ATPasen, Polymerasen oder Helikasen. Außerdem fällt die große Anzahl von Proteinen auf, bei denen ein Zusammenhang mit der natürlichen Kompetenz dieses Organismus nachgewiesen wurde [28] oder anzunehmen ist [29]. Zu diesen Proteinen gehören die in Tabelle 3.8 aufgelisteten Kompetenzproteine ComEA, ComEC, PilT und PilF. Die Kompetenzgene könnten in der Nähe des Origins liegen, weil sie in einer bisher unbekannten Weise mit dem Zellzyklus reguliert werden.

4.3.2. Vergleich mit *Deinococcus radiodurans* R1

Das Bakterium *Deinococcus radiodurans* R1 [96], das extrem unempfindlich gegenüber radioaktiver Strahlung ist, wird als der nächste Verwandte von *T. thermophilus* angesehen [98, 78]. Die Genomsequenzen beider Organismen sind vollständig bekannt. Diese Punkte waren ausschlaggebend für den Vergleich der beiden Organismen in dieser Arbeit.

Tabelle 3.9 zeigt, dass *D. radiodurans* R1 ca. 1 Mbp DNA und 964 Proteine mehr besitzt als *Thermus thermophilus* HB27. Die Genomplots (s. Abbildungen 3.22 und 3.23) der *T. thermophilus*-Sequenz gegen die beiden *Deinococcus radiodurans* Chromosomen zeigen, dass keine Folgen konservierter Genanordnungen (Synthenie) zwischen den beiden Organismen existieren. Dieses würde sich durch Ansammlungen von Kreuzen (bi-direktionale BLAST-Treffer) auf einer Diagonalen in der Graphik

bemerkbar machen. Je höher die Synthenie dabei ist, desto weniger BLAST-Treffer gibt es außerhalb der Diagonalen.

Wie im Ergebnisteil unter 3.3.2 auf Seite 68 beschrieben, sind die gemeinsamen Proteine, die durch bi-direktionale BLAST-Analysen zwischen *T. thermophilus* und *D. radiodurans* gefunden wurden, nicht gleichmäßig auf alle Replikons der beiden Organismen verteilt. Die gemeinsamen Proteine konzentrieren sich hauptsächlich auf das *T. thermophilus*-Chromosom und das *D. radiodurans*-Chromosom 1. Dadurch lässt sich festhalten, dass die beiden Genome offenbar nur entfernt verwandt sind und sich *D. radiodurans* nicht etwa durch die Aufnahme eines oder mehrerer großer DNA-Fragmente aus einem gemeinsamen Vorgänger von *T. thermophilus* und *Deinococcus radiodurans* entwickelt hat. Die eher entfernte Verwandtschaft von *T. thermophilus* und *D. radiodurans* wird auch darin deutlich, dass die Genom-Alignments auf DNA-Ebene mit MUMmer noch viel weniger gemeinsame Bereiche aufzeigen als die Genom-Alignments auf Proteinebene (Daten nicht gezeigt).

Proteomvergleich

Genomplots beruhen, wie oben schon beschrieben, auf bi-direktionalen BLAST-Analysen. Diese Analysen kann man auch tabellarisch darstellen, was für den Vergleich von *T. thermophilus* und *D. radiodurans* in Tabelle 3.10 (Seite 72) zu sehen ist.

Aus den Zahlen der Genomplots wird deutlich, dass der gemeinsame Kern von *D. radiodurans* R1 und *T. thermophilus* HB27 aus 1.196 Proteinen besteht. Diese Zahl ist, wie auch schon beim Vergleich von *Ms. mazei* mit *Ms. acetivorans*, nur ein Richtwert, denn auch hier konnte ein Teil der Proteine, die zunächst keinen Treffer im Vergleichsgenom hatten, durch eine Suche mit `tblastn` gefunden werden. Die Ursache dieses Unterschieds liegt höchstwahrscheinlich in den unterschiedlichen *ORF-findern* und Maßstäben, die die Annotatoren für das Vorhandensein von Proteinen angelegt haben. Die 1.196 gemeinsamen Proteine verteilen sich auf alle Replikons von *Deinococcus radiodurans* (s. 3.3.2 Seite 72), sind aber zum größten Teil auf Chromosom 1 zu finden. Angenommen *T. thermophilus* und *D. radiodurans* hätten sich aus einem gemeinsamen Vorfahren entwickelt, wäre wohl ein größeres Maß an Synthenie zu beobachten. Eine mögliche Erklärung wäre in diesem Zusammenhang, dass das Chromosom 2 und die Plasmide in *Deinococcus radiodurans* erst dann aufgenommen wurden, als sich aus einer gemeinsamen Linie bereits *T. thermophilus* und

D. radiodurans entwickelt hatten. Die aufgenommene Fremd-DNA in *D. radiodurans* hätte nach dieser Hypothese keine Ähnlichkeit zu *T. thermophilus* gehabt. Die vereinzelten Treffer, die man heute erkennt, könnten durch Austausch der DNA innerhalb von *D. radiodurans* entstanden sein. Diese Hypothese ist heute sicherlich nicht mehr beweisbar, aber es könnte eine mögliche Erklärung für die Unverteilung der Treffer zwischen *T. thermophilus* und *D. radiodurans* sein.

Viel interessanter sind die Zahlen, die uns sagen, wie viele Proteine für den jeweiligen Organismus spezifisch sind, denn hier sind die Unterschiede in den beiden Bakterien verankert. Interessant ist in diesem Zusammenhang auch, dass 50 % respektive 58 % der „eigenen“ Gene von *T. thermophilus* und *D. radiodurans* als hypothetisch annotiert wurden. Hier sind also keine Ähnlichkeitsinformationen in den Sequenzdatenbanken zu finden, so dass es sich bei diesen Proteinen entweder um Artefakte des *ORF-findings* handelt oder um wirklich neue Proteine. Letztere Annahme bestätigt Makarova [58], die durch Genkontext-Analysen zeigt, dass auch aus *hypothetischen* Proteinen Informationen gewonnen werden können. So wurde im genannten Beispiel wahrscheinlich ein für hypothermophile Organismen spezifisches DNA-Reparatursystem entdeckt. Etwas Ähnliches ist für *T. thermophilus* und *D. radiodurans* nicht auszuschließen, denn in den hypothetischen Proteinen, die in beiden Organismen zu finden sind, könnte durchaus ein Teil der Information kodiert sein, die *T. thermophilus* und *D. radiodurans* zum Bewohnen solch extremer Lebensräume befähigt.

Welche Funktionen besitzen die für *T. thermophilus* oder *D. radiodurans* spezifischen Proteine? Die Antwort hierauf soll eine Einordnung dieser Proteine in die COG-Gruppen liefern. Die Abbildungen 3.24 und 3.25 zeigen auf Seite 74, in welchen Gruppen die jeweils „eigenen“ Proteine eingeordnet werden konnten. Dabei ist das Verhältnis zwischen den Proteinen in der jeweiligen Gruppe bezogen auf das Gesamtgenom („total“) und den Proteinen in der Gruppe bezogen auf die eigenen („no hit“) Proteine dargestellt. Auffällig ist beim Vergleich der beiden Diagramme, dass beide Organismen in den Gruppen L (*Replikation, Rekombination und Reparatur*), N (*Zellbeweglichkeit*), G (*Kohlenhydratmetabolismus*), P (*Transport und Metabolismus anorganischer Ionen*), R (*Generelle Funktionsvorhersage*) und S (*Funktion unbekannt*) mehr spezifische Proteine besitzen, als die „totale“ Anzahl Proteine. Diese Beobachtung kann auch anhand von Abbildung 3.26 nachvollzogen werden. In diesem Diagramm sieht man, in welche COG-Gruppen die organismenspezifischen Proteine über- oder unterproportional vertreten sind. So weisen die Balken für die

COG-Gruppen K, L, N, G, P, R und S in beiden Organismen einen Wert größer als 1 auf, was für einen überproportionalen Anteil organismenspezifischer Proteine an dieser COG-Gruppe spricht.

Zusätzlich zu den schon genannten Gruppen, die sowohl in *T. thermophilus* als auch in *D. radiodurans* überrepräsentiert sind, werden aus Abbildungen 3.26 und 3.27 auch Unterschiede deutlich. In *T. thermophilus* ist die Gruppe C (*Energieproduktion und Energieumwandlung*) mit 31 Proteinen, überrepräsentiert. Unter den Proteinen der Gruppe C befinden sich z. B. 7 Oxidoreductasen, 2 Oxidase Untereinheiten, 4 Transferasen und 5 hypothetische Proteine (s. Anhang A.8).

D. radiodurans hat im Unterschied zu *T. thermophilus* überproportional viele spezifische Proteine in den Gruppen K (*Transkription*) und V (*Verteidigungsmechanismen*). Die Proteine der Gruppe K (s. Anhang A.10) enthalten auffällig viele Transkriptions-Repressoren. Die Regulation der Transkription ist in *D. radiodurans* offensichtlich stärker ausgeprägt als in *T. thermophilus*. In die Gruppe V (s. Anhang A.11) fallen 2 Proteine eines Restriktionssystems, eine Streptomycin-Kinase, 2 Acetyltransferasen und eine Acyltransferase.

Trotzdem zeigt der Vergleich von *T. thermophilus* und *D. radiodurans* deutlich, dass diese beiden Organismen, so unterschiedlich wie sie auf den ersten Blick erscheinen, doch einige Gemeinsamkeiten haben. Dieses drückt sich z. B. in einer mittleren Proteinidentität von 46 % aus (s. Abbildung 3.28). Dieser Wert ist verglichen mit durchschnittlich 82 % identischer Residuen im Proteinsequenz-Alignment von *Ms. mazei*- und *Ms. acetivorans*- Proteinen relativ gering (s. Abbildung 3.16). Hierbei muss man aber bedenken, dass mit *Ms. mazei* und *Ms. acetivorans* zwei Arten einer Gattung verglichen werden und im Fall von *T. thermophilus* und *D. radiodurans* zwei Organismen einer Klasse (*Deinococci*). Die durchschnittliche Übereinstimmung im Proteinsequenz-Alignment von 46 % zwischen *T. thermophilus* und *D. radiodurans* ist durchaus konsistent in Bezug auf den phylogenetischen Baum. Unterschiede zwischen *T. thermophilus* und *D. radiodurans* werden vor allem deutlich, wenn man die Genom-Alignments betrachtet und daraus den Schluss von nicht vorhandener Synthenie ziehen muss. Auch das Auftreten jeweils organismenspezifischer Proteine in den oben genannten COG-Gruppen zeigt, dass hier substanzielle Unterschiede vorhanden sind.

Die Untersuchung des Verwandtheitsgrads von *T. thermophilus* zu andere thermophilen Bakterien zeigt *Thermosynechococcus elongatus* BP-1 mit 40 % Identität auf der Ebene der Proteinsequenz-Alignments als nächsten Verwandten ne-

ben *D. radiodurans*, obwohl *Thermosynechococcus elongatus* zu den Cyanobakterien zählt. Diese Verwandtschaftsverhältnisse stimmen sehr gut mit der phylogenetischen Einordnung überein, denn *Thermosynechococcus elongatus* gehört zur Gruppe der *Cyanobacteria*, die im phylogenetischen Baum nicht weit von *Deinococcus* entfernt ist. Die mittlere Proteinidentität, die unter 3.3.2 auf Seite 72 beschrieben wird, weist *Thermoanaerobacter tengcongensis* [3] und *Thermotoga maritima* als weiter entfernt verwandt aus, was ebenfalls zu ihrer phylogenetischen Einordnung in die *Clostridia* bzw. *Thermotogae* übereinstimmt.

4.3.3. Horizontaler Gentransfer in *Thermus thermophilus*

In *T. thermophilus* konnten mit dem Programm SIGI 122 Gene identifiziert werden, die möglicherweise durch horizontalen Gentransfer aufgenommen wurden. Das entspricht 5,5 % aller Gene und erscheint relativ gering, bedenkt man die natürliche Kompetenz des Organismus.

Dieses Ergebnis kann mehrere Ursachen haben. Zum einen muss man bedenken, dass *T. thermophilus* in einer relativ engen ökologischen Nische lebt, in der nur thermophile und hyperthermomphile Organismen wachsen können. Die Anzahl möglicher DNA-Donor-Organismen ist damit stark eingeschränkt. Außerdem ist freie DNA bei 85 °C deutlich weniger stabil als bei mesophilen Temperaturen, so dass ein Teil dieser DNA bereits zerstört ist, bevor sie aufgenommen werden kann. Zum anderen kann der Effekt der Amelioration [49] in *T. thermophilus* mehr ausgeprägt sein als in anderen Organismen, so dass die Detektion fremder Gene hier einfach fehlschlägt. Nimmt man an, dass DNA eines fremden Organismus aufgenommen wird, besitzt dieser nach der *Genome Hypothesis* [34] seine eigene *codon usage*. Nachdem diese fremde DNA nun inseriert wurde, wirkt ein anderer Mutationsdruck auf die DNA ein, was die Zusammensetzung der aufgenommenen DNA verändert und Schritt für Schritt der des Empfängers anpasst. Durch diese Amelioration der Fremd-DNA entzieht sie sich der Entdeckung durch Mechanismen der Fremdgenerkennung. Dieser Effekt könnte auch bei der A/V-Typ-ATPase (Gene TTC0905..0913) eingetreten sein. Diese ATPase ist laut Gogarten archaeellen Typs und wurde deshalb durch Hilario als vermutlich horizontal erworben eingestuft [32, 38]. Diese Proteine wurden aber durch die Fremdgensuche mit SIGI nicht entdeckt (s. 3.3.3).

In *T. thermophilus* wurden dagegen 122 andere Proteine gefunden, die als vermutlich horizontal erworben eingestuft werden (s. 3.3.3, Seite 76). Auf dem Chromosom

sind 53 und auf dem Megaplasmid 69 dieser Gene lokalisiert. Unter den Fremdgenen auf dem Chromosom befinden sich 33 und auf dem Plasmid 60 hypothetische Gene. Auffällig ist die Tatsache, dass 99 der in *T. thermophilus* gefundenen Fremdgene kein homologes Protein in *D. radiodurans* besitzen. Von diesen 99 Genen sind insgesamt 83 als hypothetische Proteine annotiert. Der hohe Anteil von hypothetischen Genen im Zusammenhang mit Fremdgenen ist auffallend und könnte eine Bestätigung für die Fremdgenvermutung sein. Möglich wäre auch das Vorhandensein gänzlich unbekannter Proteine, die auch durch die Basenzusammensetzung ihrer Gene auffallen. Diese „besondere“ Zusammensetzung könnte ein Hinweis auf eine Funktion der Proteine sein, die sich von der des Restes abhebt.

Die Einteilung der von SIGI als fremd eingestuften Proteine in die COG-Gruppen zeigt, dass die Gruppen L und C besonders überrepräsentiert sind (s. 3.3.3). Die Gruppe L (*Replikation, Rekombination und Reparatur*) enthält hypothetische Proteine und Transposasen. Als einziges interessantes „fremdes“ Protein ist ein *DNA Integration/ Rekombination/ Invertions Protein* annotiert, das in *D. radiodurans* als Integrase bezeichnet wird. In der Gruppe V (*Verteidigungsmechanismen*) sind 2 hypothetische Proteine und eine Permease zu finden. Auch die anderen überproportional vertretenen Gruppen K, D, T, M und U enthalten hauptsächlich hypothetische Gene. Eine Ausnahme ist die Gruppe M (*Zellwand-/ Membran-/ Hüllenbiogenese*), in der Transferasen, eine Epimerase und ein regulatorisches Protein enthalten sind. Es erstaunt auf den ersten Blick, dass in den genannten COG-Gruppen als hypothetisch annotierte Proteine zu finden sind. Nach der COG Definition sollten in den einzelnen Gruppen nur solche Proteine zu finden sein, die zur Definition der jeweiligen COG-Gruppe passen. Die Proteine der in dieser Arbeit beschriebenen Genome wurden nach der unter 3.4.2 dargestellten Methode in COG-Gruppen unterteilt. Mit dem Programm Cogger konnten allerdings nur ca. 95 % vom NCBI vorgenommenen COG-Einteilungen reproduziert werden. Dieses wurde am Beispiel von *E. coli* K12 und *Bacillus subtilis* überprüft (Daten nicht gezeigt). Die hypothetischen Proteine, die nicht in die Gruppen R oder S eingeordnet sind, wurden möglicherweise zum Teil falsch annotiert. Es besteht auch die Möglichkeit, dass sie nach den vom NCBI verwendeten Kriterien mit manueller Kontrolle nicht in die entsprechenden COG-Gruppen gehören.

Zusammenfassend lässt sich über die Auswertung der Fremdgensuche sagen, dass anhand der putativen Fremdgene, von denen die meisten als hypothetische Proteine annotiert sind, keine Aussage über eine mögliche Funktion dieser Gruppe von

Proteinen gemacht werden kann. Auffallend sind allerdings die vielen hypothetischen Proteine, die beim Vergleich von *T. thermophilus* mit *D. radiodurans* als für *T. thermophilus* spezifisch bezeichnet wurden. Viele dieser Proteine sind zusätzlich als putative Fremdgene erkannt worden.

4.4. Programme

In diesem Abschnitt werden die Programme besprochen, die im Rahmen dieser Arbeit entstanden sind.

4.4.1. Quann

Quann ist ein Programm zur automatischen Annotation von Proteinen (s. 3.4.1). Das Konzept besteht dabei darin, dass möglichst viel automatisch und ohne die Notwendigkeit eines Administrators ablaufen soll. Dieses Vorgehen hat den Vorteil, dass man den Anwender nicht im Umgang mit komplexen Programmen schulen muss, und außerdem erspart man es sich so, einen Administrator zur Schulung oder Überwachung des Annotationsprozesses abzustellen. Quann läuft, nachdem der Benutzer seine Sequenzen eingegeben und eine minimale ORF-Länge gewählt hat, vollautomatisch. Es ist keine weitere Interaktion von Seiten des Benutzers oder eines Administrators notwendig. Nachdem Quann das Programm GLIMMER [14] zur Vorhersage der offenen Leserahmen verwendet hat, beginnt es *via* BLAST die Proteine mit der jeweils aktuellen Proteindatenbank des NCBI abzugleichen. Dabei wird der beste Treffer in der Datenbank als automatische Annotation für das jeweilige Protein vermerkt und die ersten zehn BLAST-Treffer werden als weitere Information für den Anwender gespeichert. Der Nachteil dieses Verfahrens besteht darin, dass man als Anwender nicht die Möglichkeit hat, eine Annotation im Programm selber zu ändern. Der Vorteil ist allerdings ein erheblicher Geschwindigkeitsgewinn.

Annotationen werden anhand der Treffer 2 bis 10 sehr einfach und schnell auf Plausibilität geprüft. Kommt es dabei zu Unstimmigkeiten, können einzelne Proteinsequenzen manuell weiterverarbeitet und genauer untersucht werden. Technisch gesehen liegt einer der Vorteile von Quann darin, dass es sehr gut auf mehrere Rechner skaliert. So ist es einfach festzulegen, wieviele Rechner für die BLAST-Suche verwendet werden sollen. Damit steigt die Verarbeitungsgeschwindigkeit des Programms mit der Anzahl von Computern, die für die BLAST-Analysen verwendet

werden können. Die Entscheidung, GLIMMER als *ORF-finder* zu verwenden, ist aus einer Abwägung zwischen Geschwindigkeit und Anzahl vorhergesagter Gene getroffen worden. GLIMMER neigt generell dazu, zu viele Gene vorherzusagen, ist aber im Vergleich zu anderen *ORF-findern* sensitiver [91]. Hinzu kommt, dass das Programm schneller ist als andere, z. B. Critica. Die Vorhersage zu vieler Gene ist jedoch kein größeres Problem, da man diese einfach ignorieren kann. Quann wird vollständig über Webseiten bedient, so dass eine zentrale Installation möglich ist und die Benutzer über das Internet von jedem beliebigen Ort aus das Programm bedienen können.

Zusammenfassend bleibt festzuhalten, dass Quann eine schnelle, einfache und wenig arbeitsintensive Möglichkeit bietet, Genomsequenzen automatisch zu annotieren. Zur endgültigen Annotation ist es jedoch einfacher, auf ein komplexes Annotationsprogramm wie z. B. ERGO auszuweichen.

4.4.2. Cogger

In dieser Arbeit werden große Mengen an Proteinsequenzen dadurch charakterisiert, dass sie in ihre COG-Gruppen eingeordnet werden. Durch diese Einordnung ist es möglich, einen Überblick über die Funktionen einer großen Menge an Proteinen zu gewinnen. Zur Einordnung der Sequenzen in die zugehörigen COGs wurde im Rahmen dieser Arbeit das Programm Cogger entwickelt und eingesetzt.

Der Vorteil des Programms liegt darin, dass man Proteine, die bisher nicht in den COG-Datenbanken vorhanden sind, selber „coggen“ kann (s. 3.4.2). Somit ist man nicht von der Arbeit des NCBI abhängig. Das ist besonders bei Genomprojekten von Vorteil, da man COGs auch zur Annotation verwenden kann. Selbst geringe bis sehr schwache Sequenzähnlichkeiten zwischen Datenbankeinträgen und den zu annotierenden Proteinen können für eine Annotation genutzt werden, denn das COG-Prinzip ist unabhängig von Schwellen bei Sequenzähnlichkeiten. Hat ein Protein BLAST-Treffer zu mindestens drei anderen Proteinen aus drei unterschiedlichen Spezies und sind diese drei Proteine demselben COG zugeordnet, wird auch die Query-Sequenz diesem COG zugeordnet. Die Theorie dahinter besagt, dass die Wahrscheinlichkeit, dass diese vier Proteine aus vier unterschiedlichen Spezies durch Zufall untereinander die höchsten Ähnlichkeiten aufweisen, sehr gering ist.

Man muss beim COG-System des NCBI grundsätzlich zwei Datenbanken unterscheiden. Zum einen die Datei `myva` und zum anderen die Datei `whog`. Bei ersterer

```
[D] COG2385  Sporulation protein and related proteins
  Dra:  DR0572
  Syn:  sll0252 sll1283 slr0191
  Bsu:  BS_lytB_2 BS_spoIID
```

Abbildung 4.2.: Auszug aus der `whog`-Datei des COG-Systems. Gezeigt ist COG2385 in der Gruppe D (*Zelluläre Kontrolle, Zellteilung und Chromosomenaufteilung*). Die Organismenabkürzungen stehen vor den jeweiligen Proteinnamen. Abkürzungen: Dra: *Deinococcus radiodurans*, Syn: *Synechocystis*, Bsu: *Bacillus subtilis*

handelt es sich um die Sequenzdatei, in der im *multiple fasta*-Format alle Proteinsequenzen der COGs abgelegt sind. Diese Datei wird für das „coggen“ der Proteine mit BLAST genutzt. Die Datei `whog` enthält dagegen detaillierte Informationen über den Inhalt der einzelnen COGs (s. Abbildung 4.2). Hier sind für jeden COG alle Proteine, sortiert nach Organismen, aufgeführt, die in dem jeweiligen COG zu finden sind.

Die `whog`-Datei ist es, die vom Programm makewhog_tng um einzelne Genome erweitert werden kann (s. 3.4.2). Diese Datei kann von anderen Programmen gelesen und verarbeitet werden. Man erweitert mit makewhog_tng allerdings nicht die `myva`-Datei, die die Sequenzdaten des COG-Systems enthält. Diese Sequenzdaten werden vom NCBI in einem automatischen Verfahren mit manueller Nachbearbeitung erstellt.

Die Funktionalität von Cogger wurde anhand bereits gecoggter Genome überprüft. Es wurden die Genome von *Bacillus subtilis* und *E. coli* [47, 7] zu Testzwecken gecoggt und die Zusammenfassung der Ergebnisse, d. h. wie viele Proteine in den einzelnen COG-Gruppen zu finden sind, mit der Angabe auf der COG-Webseite verglichen (`www.ncbi.nlm.nih.gov/COG`). Dabei ist deutlich geworden, dass Cogger verglichen mit den Angaben auf den Webseiten nur ca. 95 % aller Proteine in die richtigen COGs einordnet. Dazu muss man allerdings sagen, dass auch die original NCBI-Angaben über die Anzahl bestimmter Proteine in den einzelnen COG-Gruppen nicht mit den Webseiten übereinstimmten. Genauere stichprobenartige Untersuchungen ergaben, dass Cogger offenbar bei Proteinen, die nur sehr wenige Treffer (ca. fünf) in der COG-Datenbank haben, den Proteinen COGs zuweist, die das NCBI verwirft. Cogger geht in diesem Fall allerdings streng nach den Vorgaben vor. Der Grund für die Nichteinordnung dieser Proteine auf der NCBI-Webseite konnte bisher nicht geklärt werden.

Abschließend bleibt festzuhalten, dass das COG-Prinzip eine einfache und schnelle Kategorisierung großer Mengen von Proteinsequenzen erlaubt. Es ist außerdem eine zusätzliche Hilfe bei der Annotation. Das hier vorgestellte Programm Cogger eignet sich für die Aufgabe, Proteinsequenzen zu coggen, d. h. sie in die einzelnen COGs einzuordnen. Die Probleme des Programms bei Proteinen mit sehr wenigen BLAST-Treffern sollen jedoch noch behoben werden.

4.5. Ausblick

Die in dieser Arbeit vorgestellten Untersuchungen an unterschiedlichen mikrobiellen Genomen machen deutlich, welche Möglichkeiten die *in silico*-Analyse von Genomdaten bietet.

Genomprojekt *Escherichia blattae*

Im Rahmen des Genomprojekts *Escherichia blattae* wurden erste Einblicke in die fast vollständige Genomsequenz gegeben. Außerdem wurde die verwandtschaftliche Stellung von *E. blattae* innerhalb der Enterobakterien durch Vergleiche der Proteome einiger Vertreter dieser Gruppe verdeutlicht. Die Vervollständigung der Genomsequenz kann zu weiterführenden Untersuchungen des Genoms führen. Als Beispiele seien hier nur die Untersuchungen zum Replikationsursprung, die vollständige Erfassung und Annotation des Proteoms und die *in silico*-Untersuchung der metabolischen Fähigkeiten des Organismus aufgeführt. Aber auch die Forschung im Labor könnte von der vollständigen Sequenz profitieren, indem sie die Grundlage für die gezielte Konstruktion spezieller Produktionsstämme von *E. blattae* schaffen würde.

Das *Methanosarcina mazei* Gö1-Genom

Mit der Sequenz des Archaeons *Methanosarcina mazei* Gö1 steht ein interessanter Vertreter der „dritten Domäne des Lebens“ für die bioinformatische Auswertung zur Verfügung. Die bioinformatische Erschließung dieser Domäne hat im Vergleich zu den *Bacteria* erst begonnen und birgt ein großes Potential. Hierbei seien vor allem komparative Untersuchungen erwähnt, die zu neuen Erkenntnissen über viele der extremen Vertreter (z. B. hyperthermophile) dieser Domäne führen können.

Das *Thermus thermophilus* HB27-Genom

Thermus thermophilus HB27 wird wegen seines aeroben Wachstums und seiner leichten Handhabung auch als „*E. coli* der Thermophilen“ bezeichnet. Sowohl Forscher an der Laborbank als auch vor der Computerbildschirm profitieren von der Verfügbarkeit der *T. thermophilus*-Genomsequenz. Es ist jetzt möglich, die besonderen Aspekte dieses Organismus, zu denen zweifelsohne die Thermophilie gehört, zu untersuchen. In dieser Arbeit wurden erste genomweite Analysen vorgestellt, die die Verwandtschaft von *T. thermophilus* HB27 und *D. radiodurans* R1 beschreiben. Gleichzeitig zeigen sie mit der bi-direktionalen BLAST Methodem auf, wie zukünftig der Vergleich zwischen thermophilen und nicht thermophilen nah verwandten Organismen aussehen kann, um den Aspekt der Thermophilie zu untersuchen.

Programme

Die *Cluster of Orthologous Groups* haben seit ihrer ersten Veröffentlichung 1997 an Bedeutung gewonnen. Die Anwendung reicht dabei von ihrer Verwendung bei der Annotation bis hin zur Suche nach konservierten Genanordnungen in ganzen Domänen von Organismen. Eine Einschränkung, mit der man bisher leben musste, war die, dass man keine eigenen Proteine in die jeweils aktuelle COG-Datenbank einorden konnte und somit auf die Arbeit des NCBI angewiesen war. Das im Rahmen dieser Arbeit enwickelte Programm Cogger macht eben diese, vom NCBI unabhängige, Einordnung von Proteinen möglich. In Zukunft soll die Genauigkeit des Programms weiter verbessert werden.

SecureBLAST ist ebenfalls ein Produkt dieser Arbeit und wird bereits ausgiebig zum kontrollierten Zugriff auf Sequenzdatenbanken genutzt. Externe Forschungsgruppen haben so die Möglichkeit, den Fortgang einen Sequenzierprojektes zu kontrollieren und können in noch unvollständigen Genomdaten nach interessanten Sequenzen suchen. Es ist eine weitere Vereinfachung der Installation und Administration des Programms beabsichtigt, um den personellen Aufwand in Zukunft noch weiter zu verringern.

5. Zusammenfassung

5.1. Genomprojekt *Escherichia blattae*

1. Es wurden 4.115.744 bp des *Escherichia blattae*-Genoms sequenziert. Das entspricht bei einer geschätzten Genomgröße von 4,3 Mbp ca. 95 % des Gesamtgenoms. Zur Zeit existieren noch 43 Lücken. Der durchschnittliche GC-Gehalt des Genoms beträgt 57 %.

2. In der bisher bekannten Sequenz wurden 3.980 offene Leserahmen (ORFs) identifiziert. 3.534 Proteinen konnte eine Funktion zugewiesen werden, wobei 486 Proteine als hypothetisch annotiert wurden.

3. Im Genom von *E. blattae* wurden bis zum jetzigen Zeitpunkt drei Prophagen gefunden. Die Gene dieser Phagen konnten zusammen mit weiteren 337 als Fremdgene identifiziert werden.

4. In *E. blattae* befinden sich 650 Proteine, die keine Ähnlichkeit zu denen anderer sequenzierter Enterobakterien zeigten. 417 Proteine aus *E. blattae* hatten homologe Proteine in allen anderen untersuchten Enterobakterien.

5.2. Das *Methanosarcina mazei* Gö1-Genom

1. Der Replikationsursprung des *Methanosarcina mazei* Gö1-Genoms lag zwischen den Genen MM1314 und MM1315 (1.564.656 bis 1.566.242 bp).

2. Das Genom zeigt große Ähnlichkeit zu der Genomsequenz von *Ms. acetivorans* C2A, dessen Genom allerdings ca. 1,6 Mbp größer ist. Die beiden Organismen besitzen 2.449 gemeinsame Proteine. Der mittlere Anteil identischer Residuen im Proteinsequenz-Alignment beträgt 82 %.

3. Das zusätzliche genetische Potential von *Ms. acetivorans* ist in 1.051 Proteinen kodiert und befindet sich zu 52 % in der Region um den Replikationsterminus.

Die Annotation dieser zusätzlichen Proteine in *Ms. acetivorans* liefert keine Informationen über ihre globale Funktion.

4. Im Genom von *Ms. mazei* konnten mit unterschiedlichen Methoden 293 bzw. 544 horizontal transferierte Gene nachgewiesen werden.

5.3. Das *Thermus thermophilus* HB27-Genom

1. Der Replikationsursprung von *Thermus thermophilus* HB27 lag zwischen den Genen TTC1608 und TTC1609 (1.524.340 .. 1.524.642 bp), wobei der kumulative GC-Skew einen klassischen Verlauf zeigte. Die Region enthielt konservierte DnaA-Bindemotive.

2. *T. thermophilus* HB27 und *Deinococcus radiodurans* R1 enthielten 1.196 gemeinsame Proteine, wobei im Mittel 46 % der Aminosäureresiduen übereinstimmten. Die meisten gemeinsamen Proteine waren dabei zwischen dem *T. thermophilus*-Chromosom und dem *D. radiodurans*-Chromosom 1 zu finden. 1.508 Proteine aus *D. radiodurans* zeigten keinen Treffer in *T. thermophilus*. Umgekehrt sind es 780 Proteine aus *T. thermophilus*, die sich nicht in *D. radiodurans* wiederfinden ließen.

3. In *T. thermophilus* konnten 122 Gene identifiziert werden, die möglicherweise durch horizontalen Gentransfer erworben wurden. 99 dieser Proteine gehören zu den Proteinen, die kein homologes Protein in *D. radiodurans* besitzen.

5.4. Programmentwicklungen

1. Das Programm Quann wurde zur vollautomatischen Annotation entwickelt. Der Administrationsaufwand ist minimal und der Benutzer kann schnell einen Überblick über die Proteine bekommen, die in „seinen“ Sequenzen kodiert sind.

2. Das Programm secureBLAST ermöglicht kontrollierten Zugriff auf BLAST-Datenbanken über das Internet. Eine Zugriffskontrolle auf Basis von Benutzernamen und Passwörter wurde implementiert. Der Administrationsaufwand für secureBLAST ist minimal. Der Benutzer kann eigenständig Datenbanken erstellen und erneuern. Eine Benutzerverwaltung erlaubt das einfache Anlegen und Löschen von Benutzern.

3. Das Programm Cogger dient der unabhängigen automatischen, Einteilung von Proteinsequenzen in die entsprechenden COGs.

Literaturverzeichnis

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[2] J.H. Badger and G.J. Olsen. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol*, 16(4):512–524, Apr 1999.

[3] Q. Bao, Y. Tian, W. Li, Z. Xu, Z. Xuan, S. Hu, W. Dong, J. Yang, Y. Chen, Y. Xue, Y. Xu, X. Lai, L. Huang, X. Dong, Y. Ma, L. Ling, H. Tan, R. Chen, J. Wang, J. Yu, and H. Yang. A complete sequence of the *T. tengcongensis* genome. *Genome Res*, 12(5):689–700, May 2002.

[4] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res*, 28(1):263–266, Jan 2000.

[5] R. Bernander and K. Skarstad. Mapping of a chromosome replication origin in an archaeon. *Trends Microbiol*, 8(12):535–537, Dec 2000.

[6] H. Bierne and B. Michel. When replication forks stop. *Mol Microbiol*, 13(1):17–23, Jul 1994.

[7] F.R. Blattner, G.3rd. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–1474, Sep 1997.

[8] J.K. Bonfield, K.F. Smith, and R. Staden. A new DNA sequence assembly program. *Nucleic Acids Res*, 23(24):4992–4999, Dec 1995.

[9] T.D. Brock and H. Freeze. *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. *J Bacteriol*, 98(1):289–297, Apr 1969.

[10] H. Bruggemann, S. Baumer, W.F. Fricke, A. Wiezer, H. Liesegang, I. Decker, C. Herzberg, R. Martinez-Arias, R. Merkl, A. Henne, and G. Gottschalk. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci U S A*, 100(3):1316–1321, Feb 2003.

[11] C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen, and J. C. Venter. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273(5278):1058–1073, Aug 1996.

[12] N.R.H. Burgess, S.N. McDermott, and J. Whiting. Aerobic bacteria ourring in the hind-gut of the cockroach, Blatta orientalis. *J. Hyg., Camb.*, 71:1–7, 1973.

[13] N. Chester and D.R. Marshak. Dimethyl sulfoxide-mediated primer Tm reduction: a method for analyzing the role of renaturation temperature in the polymerase chain reaction. *Anal Biochem*, 209(2):284–290, Mar 1993.

[14] A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27(23):4636–4641, Dec 1999.

[15] A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg. Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369–2376, Jun 1999.

[16] W. Deng, V. Burland, Guy. 3rd. Plunkett, A. Boutin, G.F. Mayhew, P. Liss, N.T. Perna, D.J. Rose, B. Mau, S. Zhou, D.C. Schwartz, J.D. Fetherston, L.E. Lindler, R.R. Brubaker, G.V. Plano, S.C. Straley, K.A. McDonough, Matthew. L. Nilles, J.S. Matson, F.R. Blattner, and Robert. D. Perry. Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*, 184(16):4601–4611, Aug 2002.

[17] W. Deng, S.-R. Liou, G. 3rd. Plunkett, G.F. Mayhew, D.J. Rose, V. Burland, V. Kodoyianni, D.C. Schwartz, and F.R. Blattner. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol*, 185(7):2330–2337, Apr 2003.

[18] U. Deppenmeier, A. Johann, T. Hartsch, R. Merkl, R.A. Schmitz, R. Martinez-Arias, A. Henne, A. Wiezer, S. Baumer, C. Jacobi, H. Bruggemann, T. Lienard, A. Christmann, M. Bomeke, S. Steckel, A. Bhattacharyya, A. Lykidis, R. Overbeek, H.-P. Klenk, R.P. Gunsalus, H.-J. Fritz, and G. Gottschalk. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol*, 4(4):453–461, Jul 2002.

[19] U. Dobrindt, B. Hochhut, U. Hentschel, and J. Hacker. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*, 2(5):414–424, May 2004.

[20] W.F. Doolittle. Lateral genomics. *Trends Cell Biol*, 9(12):M5–8, Dec 1999.

[21] W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2129, Jun 1999.

[22] D.R. Edgell and W.F. Doolittle. Archaea and the origin(s) of DNA replication proteins. *Cell*, 89(7):995–998, Jun 1997.

[23] D.B. Emmert, P.J. Stoehr, G. Stoesser, and G.N. Cameron. The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res*, 22(17):3445–3449, Sep 1994.

[24] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

[25] S. Fiedler and R. Wirth. Transformation of bacteria with plasmid DNA by electroporation. *Anal Biochem*, 170(1):38–44, Apr 1988.

[26] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, and J.M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, Jul 1995.

[27] J.M. Freeman, T.N. Plasterer, T.F. Smith, and S.C. Mohr. Patterns of Genome Organization in Bacteria. *Science*, 279(5358):1827, March 1998.

[28] A. Friedrich, T. Hartsch, and B. Averhoff. Natural transformation in mesophilic and thermophilic bacteria: identification and characterization of novel,

closely related competence genes in *Acinetobacter sp.* strain BD413 and *Thermus thermophilus* HB27. *Appl Environ Microbiol*, 67(7):3140–3148, Jul 2001.

[29] A. Friedrich, J. Rumszauer, A. Henne, and B. Averhoff. Pilin-like proteins in the extremely thermophilic bacterium *Thermus thermophilus* HB27: implication in competence for natural transformation and links to type IV pilus biogenesis. *Appl Environ Microbiol*, 69(7):3695–3700, Jul 2003.

[30] K J Fryxell and E Zuckerkandl. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*, 17(9):1371–1383, Sep 2000.

[31] J.E. Galagan, C. Nusbaum, A. Roy, M.G. Endrizzi, P. Macdonald, W. FitzHugh, S. Calvo, R. Engels, S. Smirnov, D. Atnoor, A. Brown, N. Allen, J. Naylor, N. Stange-Thomann, K. DeArellano, R. Johnson, L. Linton, P. McEwan, K. McKernan, J. Talamas, A. Tirrell, W. Ye, A. Zimmer, R.D. Barber, I. Cann, D.E. Graham, D.A. Grahame, A.M. Guss, R. Hedderich, C. Ingram-Smith, H.C. Kuettner, J.A. Krzycki, J.A. Leigh, W. Li, J. Liu, B. Mukhopadhyay, J.N. Reeve, K. Smith, T.A. Springer, L.A. Umayam, O. White, R.H. White, E. Conway de. Macario., J.G. Ferry, K.F. Jarrell, H. Jing, A.J.L. Macario, I. Paulsen, M. Pritchett, K.R. Sowers, R.V. Swanson, S.H. Zinder, E. Lander, W.W. Metcalf, and B. Birren. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res*, 12(4):532–542, Apr 2002.

[32] J.P. Gogarten, T. Starke, H. Kibak, J. Fishman, and L. Taiz. Evolution and isoforms of V-ATPase subunits. *J Exp Biol*, 172:137–147, Nov 1992.

[33] B. Grabowski and Z. Kelman. Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu Rev Microbiol*, 57:487–516, 2003.

[34] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*, 8(1):r49–r62, Jan 1980.

[35] A. Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res*, 26(10):2286–2290, May 1998.

[36] D. Hanahan. Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol*, 166(4):557–580, Jun 1983.

[37] A. Henne, H. Brüggemann, C. Raasch, A. Wiezer, T. Hartsch, H. Liesegang, A. Johann, T. Lienard, O. Gohl, R. Martinez-Arias, C. Jacobi, V. Starkuviene, S. Schlenczeck, S. Dencker, R. Huber, H.P. Klenk, W. Kramer, R. Merkl, G. Gottschalk, and H.J. Fritz. The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat Biotechnol*, 22(5):289–297, May 2004.

[38] E. Hilario and J.P. Gogarten. Horizontal transfer of ATPase genes–the tree of life becomes a net of life. *Biosystems*, 31(2-3):111–119, 1993.

[39] Q. Jin, Z. Yuan, J. Xu, Y. Wang, Y. Shen, W. Lu, J. Wang, H. Liu, J. Yang, F. Yang, X. Zhang, J. Zhang, G. Yang, H. Wu, D. Qu, J. Dong, L. Sun, Y. Xue, A. Zhao, Y. Gao, J. Zhu, B. Kan, K. Ding, S. Chen, H. Cheng, Z. Yao, B. He, R. Chen, D. Ma, B. Qiang, Y. Wen, Y. Hou, and J. Yu. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Res*, 30(20):4432–4441, Oct 2002.

[40] S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, 1(5):598–610, Oct 1998.

[41] S. Karlin. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol*, 9(7):335–343, Jul 2001.

[42] S. Karlin and J. Mrazek. Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. *Proc Natl Acad Sci U S A*, 98(9):5240–5245, Apr 2001.

[43] L.M. Kelman and Z. Kelman. Archaea: an archetype for replication initiation studies. *Mol Microbiol*, 48(3):605–615, May 2003.

[44] E.V. Koonin, K.S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001.

[45] L.B. Koski and G.B. Golding. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, Jun 2001. Letter.

[46] Y. Koyama, T. Hoshino, N. Tomizuka, and K. Furukawa. Genetic transformation of the extreme thermophile *Thermus thermophilus* and of other Thermus spp. *J Bacteriol*, 166(1):338–340, Apr 1986.

[47] F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, and A. Danchin. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, 390(6657):249–256, Nov 1997.

[48] E.S. Lander and M.S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, Apr 1988.

[49] J.G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, 44(4):383–397, Apr 1997.

[50] J.G. Lawrence and H. Ochman. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A*, 95(16):9413–9417, Aug 1998.

[51] J.G. Lawrence and H. Ochman. Reconciling the many faces of lateral gene transfer. *Trends Microbiol*, 10(1):1–4, Jan 2002. News.

[52] J.R. Lobry. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 13(5):660–665, May 1996.

[53] J. Louarn, F. Cornet, V. Francois, J. Patte, and J.M. Louarn. Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization. *J Bacteriol*, 176(24):7524–7531, Dec 1994.

[54] T.M. Lowe and S.R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–964, Mar 1997.

[55] M.T. Madigan, J.M. Martinko, and J. Parker. *Brock Biology of Microorganisms*. Prentice Hall, 2000.

[56] R.H. Mah and D.R. Boone. *Bergey´s Manual of Systematic Bacteriology*, volume 3. Williams and Wilkins, Baltimore, 1989.

[57] K.S. Makarova, L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf, and E.V. Koonin. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res*, 9(7):608–628, Jul 1999.

[58] K.S. Makarova, L. Aravind, N.V. Grishin, I.B. Rogozin, and E.V. Koonin. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*, 30(2):482–496, Jan 2002.

[59] K.S. Makarova and E.V. Koonin. Comparative genomics of Archaea: how much have we learned in six years, and what's next. *Genome Biol*, 4(8):115, 2003.

[60] M. McClelland, K.E. Sanderson, J. Spieth, S.W. Clifton, P. Latreille, L. Courtney, S. Porwollik, J. Ali, M. Dante, F. Du, S. Hou, D. Layman, S. Leonard, C. Nguyen, K. Scott, A. Holmes, N. Grewal, E. Mulvaney, E. Ryan, H. Sun, L. Florea, W. Miller, T. Stoneking, M. Nhan, R. Waterston, and R.K. Wilson. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, 413(6858):852–856, Oct 2001.

[61] C. Medigue, T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol*, 222(4):851–856, Dec 1991.

[62] R. Merkl. SIGI: Score-based identification of genomic islands. *BMC Bioinformatics*, 5(22), Mar 2004.

[63] W. Messer, F. Blaesing, D. Jakimowicz, M. Krause, J. Majka, J. Nardmann, S. Schaper, H. Seitz, C. Speck, C. Weigel, G. Wegrzyn, M. Welzeck, and J. Zakrzewska-Czerwinska. Bacterial replication initiator DnaA. Rules for DnaA binding and roles of DnaA in origin unwinding and helicase loading. *Biochimie*, 83(1):5–12, Jan 2001.

[64] F. Meyer, A. Goesmann, A.C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, and A. Puhler. GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, 31(8):2187–2195, Apr 2003.

[65] I. Moszer, E.P. Rocha, and A. Danchin. Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr Opin Microbiol*, 2(5):524–528, Oct 1999.

[66] J. Mrazek and S. Karlin. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci U S A*, 95(7):3720–3725, Mar 1998.

[67] J. Mrazek and S. Karlin. Detecting alien genes in bacterial genomes. *Ann N Y Acad Sci*, 870:314–329, May 1999.

[68] H. Myllykallio, P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe, and P. Forterre. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, 288(5474):2212–2215, Jun 2000. Comment.

[69] Y. Nakamura, T. Kaneko, S. Sato, M. Ikeuchi, H. Katoh, S. Sasamoto, A. Watanabe, M. Iriguchi, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada, and S. Tabata. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res*, 9(4):123–130, Aug 2002.

[70] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[71] K.E. Nelson, R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen, and C.M. Fraser. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, 399(6734):323–329, May 1999.

[72] L. Olendzenski, L. Liu, O. Zhaxybayeva, R. Murphey, D.G. Shin, and J.P. Gogarten. Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. *J Mol Evol*, 51(6):587–599, Dec 2000.

[73] G.J. Olsen and C.R. Woese. Archaeal genomics: an overview. *Cell*, 89(7):991–994, Jun 1997.

[74] R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E.Jr. Selkov, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman,

V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides. The ERGO genome analysis and discovery system. *Nucleic Acids Res*, 31(1):164–171, Jan 2003.

[75] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, Apr 1988.

[76] S. Porwollik and M. McClelland. Lateral gene transfer in *Salmonella*. *Microbes Infect*, 5(11):977–989, Sep 2003.

[77] Williams R. and Sharp R. *Thermus species. Biotechnology handbooks.* Plenum Press, New York., 1995.

[78] F.A. Rainey, M.F. Nobre, P. Schumann, E. Stackebrandt, and M.S. da Costa. Phylogenetic diversity of the deinococci as determined by 16S ribosomal DNA sequence comparison. *Int J Syst Bacteriol*, 47(2):510–514, Apr 1997.

[79] R.J. Robins. Challenges in the human genome projekt. *IEEE Engineering in Medicine and Biology*, 11(1):25–34, Mar 1992.

[80] N.P. Robinson, I. Dionne, M. Lundgren, V.L. Marsh, R. Bernander, and Stephen. D. Bell. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell*, 116(1):25–38, Jan 2004.

[81] S.L. Salzberg, A.J. Salzberg, A.R. Kerlavage, and J.F. Tomb. Skewed oligomers and origins of replication. *Gene*, 217(1-2):57–67, Sep 1998.

[82] J. Sambrook, E.F. Fritsch, and T. Maniatis. *Molecular cloning: a laboratory manual.* Cold Spring Harbour Laboratory, 1989.

[83] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991.

[84] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*, 24:104–108, 1992. Biography.

[85] S. Schaper, J. Nardmann, G. Luder, R. Lurz, C. Speck, and W. Messer. Identification of the chromosomal replication origin from *Thermus thermophilus* and its interaction with the replication initiator DnaA. *J Mol Biol*, 299(3):655–665, Jun 2000.

[86] S. Shuman. Recombination mediated by vaccinia virus DNA topoisomerase I in *Escherichia coli* is sequence specific. *Proc Natl Acad Sci U S A*, 88(22):10104–10108, Nov 1991.

[87] A.I. Slesarev, K.V. Mezhevaya, K.S. Makarova, N.N. Polushin, O.V. Shcherbinina, V.V. Shakhova, G.I. Belova, L. Aravind, D.A. Natale, I.B. Rogozin, R.L. Tatusov, Y.I. Wolf, K.O. Stetter, A.G. Malykh, E.V. Koonin, and S.A. Kozyavkin. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc Natl Acad Sci U S A*, 99(7):4644–4649, Apr 2002.

[88] D. R. Smith, L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, and J. N. Reeve. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol*, 179(22):7135–7155, Nov 1997.

[89] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[90] R.L. Tatusov, E.V. Koonin, and D.J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.

[91] M. Tech and R. Merkl. YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol*, 3(4):441–451, 2003.

[92] H. Tettelin, D. Radune, S. Kasif, H. Khouri, and S.L. Salzberg. Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. *Genomics*, 62(3):500–507, Dec 1999.

[93] R.C.H.J. van Ham., J. Kamerbeek, C. Palacios, C. Rausell, F. Abascal, U. Bastolla, J.M. Fernandez, L. Jimenez, M. Postigo, F.J. Silva, J. Tamames, E. Viguera, A. Latorre, A. Valencia, F. Moran, and A. Moya. Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci U S A*, 100(2):581–586, Jan 2003.

[94] A. Vas and J. Leatherwood. Where does DNA replication start in archaea. *Genome Biol*, 1(3):REVIEWS1020, 2000.

[95] M.S. Waterman. Efficient sequence alignment algorithms. *J Theor Biol*, 108(3):333–337, Jun 1984.

[96] O. White, J.A. Eisen, J.F. Heidelberg, E.K. Hickey, J.D. Peterson, R.J. Dodson, D.H. Haft, M.L. Gwinn, W.C. Nelson, D.L. Richardson, K.S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J.J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K.S. Makarova, L. Aravind, M.J. Daly, and C.M. Fraser. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, 286(5444):1571–1577, Nov 1999.

[97] A. Wiezer and R. Merkl. secureBLAST. *In Silico Biol*, 3(4):405–409, 2003.

[98] R.A. Williams, K.E. Smith, S.G. Welch, J. Micallef, and R.J. Sharp. DNA relatedness of Thermus strains, description of *Thermus brockianus* sp. nov., and proposal to reestablish Thermus thermophilus (Oshima and Imahori. *Int J Syst Bacteriol*, 45(3):495–499, Jul 1995.

[99] R. Zhang and C.-T. Zhang. Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. *Biochem Biophys Res Commun*, 297(2):396–400, Sep 2002.

[100] R. Zhang and C.-T. Zhang. Multiple replication origins of the archaeon *Halobacterium species* NRC-1. *Biochem Biophys Res Commun*, 302(4):728–734, Mar 2003.

A. Anhang

A.1. bi-direktionales BLAST

Bei zwei blast Ergebnissen a.blast und b.blast könnte der Aufruf und die Auswertung wie folgt aussehen: `.pr_both.pl a.blast b.blast | grep BLAST2`. Um die Positionen der einzelnen Treffer auf den jeweiligen Chromosomen auszugeben, muß die *identyfier line* im FASTA-file zum Beispiel wie folgt aussehen:
`>NP-293728.1_(DR0002)_1904-3268 chromosomal replication protein`
(Genname_Abkürzung_Start_Stop Annotation). Diese Notation ist wichtig, weil das Skript (`pr_both.pl`) sonst die Positionen nicht erkennt, was allerding für Plots der Genomvergleiche entscheidend ist. Wenn der Benutzer Histogramme erstellen möchten, die die Qualität der Treffer aus dem Vergleich abgeben, so kann dies über `.pr_both.pl a.blast b.blast | grep HIST` geschehen. Die Werte der Ausgabe sind dabei die identität bzw. die Ähnlichkeit der Proteinsequenzen des Alignment in Prozent.

```
#!/usr/bin/perl
# Nur hits anzeigen, die in beiden Richtungen
# auf das selbe Protein zeigen

use lib '/usr/lib/perl5/site_perl/5.6.0/Bio/';
use Bio::SearchIO;
use strict;

my $debug = 1;

my ($start, $stop, $tt_start, $tt_stop, $second_start, \
 $second_stop, $orf, $tt, $second, $ba, $ab, $cblast, \
 $searchio, $query_name, $bresult, $entry, $hsp, $cons,\
 $ident, $temp, $query_desc, $bla, $pos, $cont, $desc, \
 $hit_name, $tt_desc, $second_desc, $hit_desc);

my @blasts = (@ARGV);
```

```
if (scalar(@ARGV) < 2) {
    die "Usage: pr_both.pl 1st-blastresult 2nd-blastresult\n";
}

foreach $cblast (@blasts)
{
    if (!-e $cblast) {
        die "$cblast was not found !\n";
    }
}

foreach $cblast (@blasts)
{
    $searchio = new Bio::SearchIO ('-format' => 'blast', '-file' \
     => $cblast);
    while ( my $result = $searchio->next_result() )
    {
        $query_name = $result->query_accession;
        ($bla, $bla, $start, $stop, $query_desc) = \
        split(/_/, $result->query_name);
        $second->{$query_name}->{"start"} = $start;
        $second->{$query_name}->{"stop"}  = $stop;
        $second->{$query_name}->{"desc"}  = $query_desc;

        if (defined(my $hit = $result->next_hit)) {
            $hit_name = $hit->name;
            $hit_desc = $hit->description;
             $temp = $hit->name." ".$hit->description;
             $bla = rindex($temp, "_")+1;
             $hit_desc = substr($temp, $bla);

            if ($debug) {
                print "BLAST:".$query_name." ".$query_desc.\
                "-> ".$hit_name."\n";
            }

            $hsp   = $hit->next_hsp();
            $cons  = sprintf "%3.2f", ($hsp->frac_conserved*100);
            $ident = sprintf "%3.2f", ($hsp->frac_identical*100);

            $bresult->{$query_name}->{"hit"}   = $hit_name;
            $bresult->{$query_name}->{"ident"} = $ident;
            $bresult->{$query_name}->{"cons"}  = $cons;
        } else {
            if ($debug){
```

```
                print "BLAST:".$query_name." ".$query_desc.\
                "-> NO HIT\n";
            }
            $bresult->{$query_name}->{"hit"} = "NO HIT";
        }
    }
}


foreach $entry (keys %$bresult)
{
    $ab = $bresult->{$entry}->{"hit"};
    $ba = $bresult->{$ab}->{"hit"};

    $a = substr($ab, 0, 2);
    $b = substr($ba, 0, 2);

    $cons     = $bresult->{$entry}->{"cons"};
    $ident    = $bresult->{$entry}->{"ident"};

    if ($debug) {
            print "-----------------------------------\
            --------\n";
            print "ENTRY:$entry\tvalue: ".$bresult->\
            {$entry}->{"hit"}."\n";
            print "ab: $ab\tba: $ba\n";
        }

    if ($entry eq $ba) {
            $tt_start = $tt->{$ab}->{"start"};
            $tt_stop  = $tt->{$ab}->{"stop"};
            $tt_desc  = $tt->{$ab}->{"desc"};

            $second_start = $second->{$ba}->{"start"};
            $second_stop  = $second->{$ba}->{"stop"};
            $second_desc  = $second->{$ba}->{"desc"};

            print "HIST\t$entry\t$ident\t$cons\n";
            print "BLAST2:$entry\t->\t".$bresult->{$entry}\
            ->{"hit"}."\t".$tt_desc."\n";
          }
}
```

A.2. *Ms. mazei*-spezifische Proteine

Genname, COG, (COG-Gruppe), Annotation

MM00003	NOCOG	hypothetical protein
MM00010	NOCOG	hypothetical protein
MM00012	COG5507 (S)	RNA signal recognition particle 4.5S RNA
MM00013	NOCOG	conserved protein
MM00021	NOCOG	conserved protein
MM00023	COG1724 (N)	conserved protein
MM00025	NOCOG	hypothetical protein
MM00026	NOCOG	conserved protein
MM00029	NOCOG	conserved protein
MM00032	COG1724 (N)	conserved protein
MM00034	NOCOG	hypothetical protein
MM00036	COG0419 (L)	hypothetical protein
MM00045	COG0666 (R)	hypothetical protein
MM00051	NOCOG	hypothetical protein
MM00053	NOCOG	hypothetical protein
MM00065	NOCOG	hypothetical protein
MM00076	NOCOG	hypothetical protein
MM00083	NOCOG	hypothetical protein
MM00094	NOCOG	hypothetical protein
MM00095	NOCOG	conserved protein
MM00104	NOCOG	hypothetical protein
MM00105	NOCOG	hypothetical protein
MM00149	COG3967 (M)	hypothetical protein
MM00150	COG0702 (MG)	putative nucleoside-diphosphate-sugar epimerase
MM00151	COG3378 (R)	conserved protein
MM00159	NOCOG	hypothetical protein
MM00165	NOCOG	hypothetical protein
MM00186	COG4274 (S)	conserved protein
MM00189	COG4274 (S)	conserved protein
MM00192	COG0258 (L)	hypothetical protein
MM00193	COG2189 (L)	type III restriction-modification system methylation subunit
MM00194	NOCOG	conserved protein
MM00195	NOCOG	conserved protein
MM00196	COG0433 (R)	conserved protein
MM00197	NOCOG	conserved protein
MM00200	NOCOG	conserved protein
MM00201	NOCOG	conserved protein
MM00204	COG0454 (KR)	hypothetical protein
MM00227	NOCOG	conserved protein
MM00229	NOCOG	hypothetical protein
MM00250	COG3436 (L)	Transposase
MM00257	NOCOG	hypothetical protein
MM00270	NOCOG	conserved protein
MM00273	NOCOG	hypothetical protein
MM00274	COG1502 (I)	putative endonuclease
MM00275	NOCOG	hypothetical protein
MM00276	NOCOG	Superfamily II DNA and RNA helicase
MM00278	COG0270 (L)	DNA-cytosine methyltransferase
MM00298	COG1732 (M)	Glycine betaine-binding protein
MM00324	NOCOG	hypothetical protein
MM00343	NOCOG	hypothetical protein
MM00347	NOCOG	hypothetical protein
MM00349	NOCOG	hypothetical protein
MM00352	NOCOG	hypothetical protein
MM00367	COG3945 (S)	hypothetical protein
MM00396	COG3436 (L)	Transposase
MM00404	NOCOG	hypothetical protein
MM00410	NOCOG	hypothetical protein
MM00452	NOCOG	hypothetical protein
MM00470	NOCOG	hypothetical protein
MM00479	NOCOG	conserved protein
MM00480	NOCOG	hypothetical protein
MM00494	NOCOG	hypothetical protein
MM00501	COG3865 (S)	3-demethylubiquinone-9 3-methyltransferase
MM00502	COG0251 (J)	Translation initiation inhibitor
MM00519	NOCOG	hypothetical protein
MM00520	NOCOG	hypothetical protein
MM00521	NOCOG	conserved protein
MM00523	NOCOG	hypothetical protein
MM00525	NOCOG	hypothetical protein
MM00526	NOCOG	hypothetical protein
MM00532	NOCOG	hypothetical protein
MM00538	NOCOG	hypothetical protein
MM00539	NOCOG	hypothetical protein
MM00542	NOCOG	hypothetical protein
MM00543	NOCOG	hypothetical protein
MM00544	COG0557 (K)	Exoribonuclease II
MM00547	NOCOG	conserved protein
MM00548	NOCOG	hypothetical protein
MM00549	COG3654 (R)	conserved protein
MM00554	NOCOG	conserved protein
MM00556	NOCOG	hypothetical protein
MM00557	COG1468 (L)	hypothetical protein

```
MM00558 COG1343 (L)     hypothetical protein
MM00562 NOCOG           hypothetical protein
MM00563 NOCOG           hypothetical protein
MM00564 NOCOG           hypothetical protein
MM00565 COG0640 (K)     transcriptional regulator
MM00579 COG0675 (L)     Transposase
MM00604 NOCOG           conserved protein
MM00622 COG3415 (L)     Transposase
MM00623 COG3335 (L)     Transposase
MM00640 NOCOG           hypothetical protein
MM00655 NOCOG           hypothetical protein
MM00683 NOCOG           hypothetical protein
MM00693 COG1418 (R)     conserved protein
MM00710 NOCOG           hypothetical protein
MM00731 NOCOG           hypothetical protein
MM00735 COG1742 (S)     conserved protein
MM00739 COG3436 (L)     Transposase
MM00746 NOCOG           conserved protein
MM00758 NOCOG           conserved protein
MM00762 NOCOG           conserved protein
MM00768 NOCOG           glycosyltransferase
MM00769 COG4186 (R)     phosphohydrolase
MM00773 COG4572 (R)     putative cation transport regulator
MM00791 NOCOG           hypothetical protein
MM00795 NOCOG           hypothetical protein
MM00796 NOCOG           hypothetical protein
MM00816 COG3436 (L)     Transposase
MM00851 NOCOG           conserved protein
MM00875 NOCOG           hypothetical protein
MM00879 NOCOG           conserved protein
MM00885 NOCOG           hypothetical protein
MM00894 NOCOG           CbiM protein
MM00908 NOCOG           conserved protein
MM00919 COG5549 (O)     conserved protein
MM00924 NOCOG           conserved protein
MM00925 NOCOG           conserved protein
MM00930 COG0347 (E)     Nitrogen regulatory protein P-II
MM00931 NOCOG           hypothetical protein
MM00932 NOCOG           hypothetical protein
MM00942 NOCOG           hypothetical protein
MM00954 NOCOG           hypothetical protein
MM00965 NOCOG           hypothetical protein
MM00971 NOCOG           hypothetical protein
MM00978 NOCOG           hypothetical protein
MM00996 COG2073 (H)     cobalamin biosynthesis protein G
MM00997 COG2073 (H)     cobalamin biosynthesis protein G
MM01001 NOCOG           hypothetical protein
MM01016 COG0454 (KR)    putative ribosomal-protein-alanine acetyltransferase
MM01021 NOCOG           hypothetical protein
MM01049 COG1055 (P)     Arsenite permease
MM01065 NOCOG           hypothetical protein
MM01066 NOCOG           hypothetical protein
MM01091 COG0526 (OC)    thiol-disulfide isomerase
MM01092 COG0526 (OC)    thiol-disulfide isomerase
MM01099 NOCOG           hypothetical protein
MM01102 NOCOG           hypothetical protein
MM01130 NOCOG           hypothetical protein
MM01150 NOCOG           conserved protein
MM01152 COG0399 (M)     Aspartate aminotransferase
MM01158 NOCOG           conserved protein
MM01170 COG0381 (M)     UDP-N-acetylglucosamine 2-epimerase
MM01173 NOCOG           conserved protein
MM01174 NOCOG           hypothetical protein
MM01181 NOCOG           conserved protein
MM01187 NOCOG           hypothetical protein
MM01188 NOCOG           conserved protein
MM01230 NOCOG           hypothetical protein
MM01250 NOCOG           Cation transporter
MM01251 NOCOG           Cation transporter
MM01252 NOCOG           conserved protein
MM01253 NOCOG           Cation transporter
MM01260 COG3436 (L)     Transposase
MM01268 NOCOG           hypothetical protein
MM01282 NOCOG           conserved protein
MM01337 NOCOG           hypothetical protein
MM01368 COG0695 (O)     Glutaredoxin-like protein
MM01372 NOCOG           hypothetical protein
MM01389 COG1266 (R)     conserved protein
MM01392 NOCOG           hypothetical protein
MM01400 NOCOG           hypothetical protein
MM01422 COG1917 (S)     hypothetical protein
MM01432 NOCOG           hypothetical protein
MM01434 NOCOG           Monomethylamine permease
MM01446 NOCOG           hypothetical protein
MM01459 NOCOG           hypothetical protein
MM01460 NOCOG           conserved protein
MM01466 COG3677 (L)     hypothetical protein
MM01468 NOCOG           hypothetical protein
MM01481 NOCOG           Iron-dependent repressor
MM01482 NOCOG           hypothetical protein
MM01483 NOCOG           Ferrous iron transport protein B
MM01532 NOCOG           hypothetical protein
```

MM01624	COG3436 (L)	Transposase
MM01645	NOCOG	hypothetical protein
MM01646	NOCOG	conserved protein
MM01653	NOCOG	conserved protein
MM01657	COG0639 (T)	putative serine/threonine protein phosphatase
MM01659	NOCOG	hypothetical protein
MM01660	COG2239 (P)	MG2+ transporter MGTE
MM01663	NOCOG	conserved protein
MM01666	NOCOG	ATPase involved in DNA repair
MM01672	COG2226 (H)	methyltransferase
MM01701	COG0702 (MG)	putative nucleoside-diphosphate-sugar epimerase
MM01718	NOCOG	conserved protein
MM01733	NOCOG	metalloproteinase
MM01736	COG3436 (L)	Transposase
MM01765	NOCOG	hypothetical protein
MM01768	NOCOG	hypothetical protein
MM01779	NOCOG	hypothetical protein
MM01782	COG0786 (E)	Sodium/glutamate symport carrier protein
MM01784	NOCOG	hypothetical protein
MM01785	NOCOG	hypothetical protein
MM01787	COG3436 (L)	Transposase
MM01822	NOCOG	hypothetical protein
MM01823	COG2426 (S)	putative small multi-drug export protein
MM01832	COG3436 (L)	Transposase
MM01841	NOCOG	hypothetical protein
MM01848	NOCOG	hypothetical protein
MM01850	NOCOG	hypothetical protein
MM01866	NOCOG	hypothetical protein
MM01887	NOCOG	hypothetical protein
MM01898	COG1226 (P)	Potassium channel protein
MM01899	NOCOG	hypothetical protein
MM01900	NOCOG	hypothetical protein
MM01905	NOCOG	hypothetical membrane spanning protein
MM01907	COG1335 (Q)	Isochorismatase
MM01921	NOCOG	hypothetical protein
MM01922	COG1522 (K)	hypothetical protein
MM01926	NOCOG	hypothetical protein
MM01938	NOCOG	hypothetical protein
MM01943	NOCOG	conserved protein
MM01951	NOCOG	hypothetical protein
MM01959	COG4095 (S)	hypothetical protein
MM01972	NOCOG	hypothetical protein
MM01982	COG2015 (Q)	Alkyl sulfatase
MM01984	NOCOG	hypothetical cytosolic protein
MM01985	NOCOG	hypothetical cytosolic protein
MM01989	NOCOG	hypothetical protein
MM01991	NOCOG	hypothetical protein
MM01992	NOCOG	hypothetical protein
MM02025	NOCOG	hypothetical protein
MM02026	NOCOG	hypothetical protein
MM02027	NOCOG	hypothetical protein
MM02029	NOCOG	conserved protein
MM02096	COG0472 (M)	UDP-N-Acetylglucosamine-1-phosphate transferase
MM02097	NOCOG	hypothetical protein
MM02100	NOCOG	conserved protein
MM02103	COG0500 (QR)	putative methyltransferase
MM02108	NOCOG	conserved protein
MM02114	NOCOG	hypothetical protein
MM02116	NOCOG	hypothetical protein
MM02118	NOCOG	hypothetical protein
MM02119	NOCOG	hypothetical protein
MM02120	NOCOG	hypothetical protein
MM02121	NOCOG	hypothetical protein
MM02165	COG3415 (L)	Transposase
MM02166	NOCOG	conserved protein
MM02183	NOCOG	hypothetical protein
MM02191	COG1403 (V)	conserved protein
MM02200	NOCOG	hypothetical protein
MM02201	NOCOG	hypothetical protein
MM02202	NOCOG	hypothetical protein
MM02203	COG3943 (R)	hypothetical protein
MM02207	NOCOG	conserved protein
MM02211	NOCOG	hypothetical protein
MM02232	NOCOG	Transposase
MM02255	NOCOG	conserved protein
MM02285	NOCOG	conserved protein
MM02290	COG0860 (M)	Cell surface protein
MM02293	COG0732 (V)	type I restriction-modification system specificity subunit
MM02323	NOCOG	Ech Hydrogenase, Subunit
MM02338	NOCOG	DNA-directed RNA polymerases I, II, and III, 7.3 kDa polypeptide
MM02342	NOCOG	hypothetical protein
MM02377	COG1714 (S)	hypothetical protein
MM02382	NOCOG	hypothetical protein
MM02390	COG3328 (L)	Transposase
MM02401	NOCOG	conserved protein
MM02407	NOCOG	conserved protein
MM02413	NOCOG	conserved protein
MM02414	NOCOG	hypothetical protein
MM02415	NOCOG	hypothetical protein
MM02421	COG0326 (O)	Chaperone protein
MM02430	COG3436 (L)	Transposase
MM02436	COG3436 (L)	Transposase

```
MM02442 NOCOG          conserved protein
MM02449 NOCOG          conserved protein
MM02450 COG3328 (L)    Transposase
MM02451 NOCOG          conserved protein
MM02452 COG0785 (O)    conserved protein
MM02461 NOCOG          conserved protein
MM02476 COG1670 (J)    Ribosomal-protein-alanine acetyltransferase
MM02477 NOCOG          hypothetical protein
MM02521 NOCOG          hypothetical protein
MM02530 COG3328 (L)    Transposase
MM02543 NOCOG          hypothetical protein
MM02544 NOCOG          hypothetical protein
MM02548 NOCOG          hypothetical protein
MM02555 NOCOG          hypothetical protein
MM02557 COG0753 (P)    Catalase
MM02558 NOCOG          hypothetical protein
MM02559 COG1479 (S)    hypothetical protein
MM02560 NOCOG          hypothetical protein
MM02580 NOCOG          hypothetical protein
MM02588 COG3436 (L)    Transposase
MM02590 NOCOG          Transposase
MM02623 COG0689 (J)    Ribonuclease
MM02626 COG1996 (K)    DNA-directed RNA polymerase subunit P
MM02643 COG1335 (Q)    hypothetical protein
MM02651 NOCOG          hypothetical protein
MM02655 COG1835 (I)    hypothetical protein
MM02672 NOCOG          conserved protein
MM02677 NOCOG          conserved protein
MM02687 NOCOG          hypothetical protein
MM02688 COG1032 (C)    hypothetical protein
MM02690 COG3436 (L)    Transposase
MM02701 NOCOG          hypothetical protein
MM02724 NOCOG          conserved protein
MM02726 NOCOG          hypothetical protein
MM02735 NOCOG          conserved protein
MM02747 NOCOG          hypothetical protein
MM02755 NOCOG          hypothetical protein
MM02756 NOCOG          hypothetical protein
MM02757 NOCOG          hypothetical protein
MM02758 NOCOG          hypothetical protein
MM02759 NOCOG          hypothetical protein
MM02761 NOCOG          hypothetical protein
MM02762 NOCOG          hypothetical protein
MM02763 COG1961 (L)    putative DNA recombinase
MM02764 NOCOG          hypothetical protein
MM02765 NOCOG          hypothetical protein
MM02766 NOCOG          hypothetical protein
MM02767 NOCOG          hypothetical protein
MM02775 COG2339 (S)    conserved protein
MM02776 NOCOG          hypothetical protein
MM02781 NOCOG          conserved protein
MM02824 COG2141 (C)    glucose-6-phosphate dehydrogenase [coenzyme F420 dependent]
MM02838 NOCOG          conserved protein
MM02843 NOCOG          hypothetical protein
MM02863 NOCOG          conserved protein
MM02873 COG2133 (G)    L-sorbosone dehydrogenase
MM02874 NOCOG          conserved protein
MM02890 NOCOG          hypothetical protein
MM02898 NOCOG          conserved protein
MM02899 COG2985 (R)    conserved protein
MM02908 NOCOG          hypothetical protein
MM02920 COG1670 (J)    Ribosomal-protein-alanine acetyltransferase
MM02922 NOCOG          hypothetical protein
MM02926 NOCOG          hypothetical protein
MM02928 NOCOG          hypothetical protein
MM02929 NOCOG          hypothetical protein
MM02957 COG3328 (L)    Transposase
MM02966 COG2248 (R)    conserved protein
MM02969 NOCOG          conserved protein
MM02971 NOCOG          conserved protein
MM02975 NOCOG          hypothetical protein
MM02976 COG0610 (V)    type I restriction-modification system restriction subunit
MM02977 NOCOG          hypothetical protein
MM02978 COG0732 (V)    type I restriction-modification system specificity subunit
MM02980 NOCOG          hypothetical protein
MM02984 NOCOG          putative ferredoxin
MM02986 NOCOG          hypothetical protein
MM02987 COG4548 (P)    hypothetical protein
MM02988 COG0714 (R)    conserved protein
MM02995 NOCOG          hypothetical protein
MM02997 NOCOG          hypothetical protein
MM03002 NOCOG          hypothetical protein
MM03003 NOCOG          conserved protein
MM03025 NOCOG          hypothetical protein
MM03029 COG0346 (E)    Lactoylglutathione lyase
MM03031 NOCOG          hypothetical protein
MM03033 NOCOG          conserved protein
MM03035 NOCOG          hypothetical protein
MM03036 COG0454 (KR)   hypothetical protein
MM03039 NOCOG          hypothetical protein
MM03040 NOCOG          hypothetical protein
MM03041 NOCOG          hypothetical protein
```

MM03046	NOCOG	conserved protein
MM03048	COG2273 (G)	hypothetical protein
MM03049	COG0438 (M)	glycosyltransferase
MM03051	NOCOG	hypothetical protein
MM03055	NOCOG	conserved protein
MM03056	COG2333 (R)	Competence-like protein
MM03057	NOCOG	hypothetical protein
MM03060	NOCOG	hypothetical protein
MM03061	NOCOG	hypothetical protein
MM03062	NOCOG	hypothetical protein
MM03063	COG1404 (O)	calcium dependent protease
MM03064	NOCOG	hypothetical protein
MM03065	NOCOG	hypothetical protein
MM03079	NOCOG	hypothetical protein
MM03083	NOCOG	conserved protein
MM03089	NOCOG	hypothetical protein
MM03124	NOCOG	hypothetical protein
MM03126	NOCOG	hypothetical protein
MM03128	COG0671 (I)	hypothetical protein
MM03158	COG0326 (O)	hypothetical protein
MM03159	NOCOG	hypothetical protein
MM03160	COG3436 (L)	Transposase
MM03162	NOCOG	hypothetical protein
MM03163	NOCOG	conserved protein
MM03165	NOCOG	hypothetical protein
MM03166	NOCOG	hypothetical protein
MM03184	NOCOG	conserved protein
MM03192	NOCOG	hypothetical membrane spanning protein
MM03200	NOCOG	hypothetical protein
MM03209	COG0679 (R)	Transporter
MM03236	COG3415 (L)	Transposase
MM03239	NOCOG	hypothetical protein
MM03242	NOCOG	hypothetical protein
MM03245	COG3436 (L)	Transposase
MM03248	NOCOG	hypothetical protein
MM03250	NOCOG	hypothetical protein
MM03251	NOCOG	hypothetical protein
MM03252	NOCOG	hypothetical protein
MM03253	COG0417 (L)	hypothetical protein
MM03254	NOCOG	hypothetical protein
MM03260	COG2323 (S)	hypothetical protein
MM03266	NOCOG	hypothetical protein
MM03273	NOCOG	conserved protein
MM03275	COG1253 (R)	conserved protein
MM03285	COG1408 (R)	conserved protein
MM03287	NOCOG	hypothetical protein
MM03292	COG1553 (P)	hypothetical protein
MM03302	NOCOG	hypothetical protein
MM03306	NOCOG	conserved protein
MM03308	COG1073 (R)	conserved protein
MM03315	NOCOG	hypothetical protein
MM03316	NOCOG	hypothetical protein
MM03317	COG4119 (LR)	hypothetical protein
MM03319	COG0477 (GEPR)	Transport protein
MM03341	NOCOG	conserved protein
MM03347	NOCOG	conserved protein
MM03348	NOCOG	hypothetical protein
MM03349	NOCOG	hypothetical protein
MM03353	NOCOG	conserved protein
MM03354	NOCOG	hypothetical protein
MM03355	COG1336 (L)	conserved protein
MM03356	NOCOG	hypothetical protein
MM03357	NOCOG	hypothetical protein
MM03358	NOCOG	hypothetical protein
MM03367	NOCOG	hypothetical protein
MM03370	NOCOG	hypothetical protein

A.3. *Ms. acetivorans*-spezifische Proteine

Genname, COG, (COG-Gruppe), Annotation

MA0002	NOCOG	predicted protein
MA0017	NOCOG	hypothetical protein
MA0027	NOCOG	predicted protein
MA0028	NOCOG	predicted protein
MA0038	COG5496 (R)	predicted protein
MA0041	NOCOG	predicted protein
MA0044	NOCOG	predicted protein

MA0049 COG2026 (JD)	conserved hypothetical protein
MA0052 NOCOG	predicted protein
MA0066 NOCOG	predicted protein
MA0070 NOCOG	predicted protein
MA0073 NOCOG	predicted protein
MA0078 COG1672 (R)	predicted protein
MA0082 NOCOG	predicted protein
MA0094 NOCOG	predicted protein
MA0095 NOCOG	predicted protein
MA0098 NOCOG	predicted protein
MA0099 COG1708 (R)	conserved hypothetical protein
MA0111 NOCOG	predicted protein
MA0161 COG1528 (P)	ferritin
MA0167 COG2010 (C)	conserved hypothetical protein
MA0177 NOCOG	predicted protein
MA0178 COG1309 (K)	bacterial regulatory protein, tetR family
MA0186 NOCOG	predicted protein
MA0193 NOCOG	predicted protein
MA0196 NOCOG	predicted protein
MA0200 COG1993 (S)	conserved hypothetical protein
MA0226 NOCOG	predicted protein
MA0237 NOCOG	predicted protein
MA0248 NOCOG	predicted protein
MA0268 NOCOG	predicted protein
MA0280 COG0725 (P)	conserved hypothetical protein
MA0288 NOCOG	predicted protein
MA0291 NOCOG	predicted protein
MA0312 NOCOG	predicted protein
MA0317 COG0561 (R)	conserved hypothetical protein
MA0318 COG0561 (R)	conserved hypothetical protein
MA0319 NOCOG	predicted protein
MA0320 COG0456 (R)	acetyltransferase (GNAT) family
MA0347 NOCOG	predicted protein
MA0349 NOCOG	predicted protein
MA0351 COG2801 (L)	integrase
MA0355 NOCOG	predicted protein
MA0361 COG0716 (C)	conserved hypothetical protein
MA0362 COG0596 (R)	conserved hypothetical protein
MA0363 COG0716 (C)	conserved hypothetical protein
MA0371 NOCOG	predicted protein
MA0376 COG2026 (JD)	conserved hypothetical protein
MA0390 NOCOG	predicted protein
MA0391 NOCOG	predicted protein
MA0392 NOCOG	predicted protein
MA0399 COG2801 (L)	retroviral pol related endonuclease
MA0401 NOCOG	predicted protein
MA0402 COG1917 (S)	conserved hypothetical protein
MA0407 COG0716 (C)	predicted protein
MA0409 COG0599 (S)	4-carboxymuconolactone decarboxylase
MA0411 NOCOG	predicted protein
MA0412 NOCOG	predicted protein
MA0413 COG4925 (S)	predicted protein
MA0414 COG1733 (K)	conserved hypothetical protein
MA0416 COG1917 (S)	conserved hypothetical protein
MA0419 COG1073 (R)	membrane protein
MA0421 COG4902 (S)	conserved hypothetical protein
MA0423 NOCOG	predicted protein
MA0425 COG3391 (S)	hypothetical protein (with a domain homologous to a RING finger protein)
MA0438 NOCOG	predicted protein
MA0448 NOCOG	predicted protein
MA0450 NOCOG	predicted protein
MA0457 NOCOG	predicted protein
MA0458 NOCOG	predicted protein
MA0468 NOCOG	predicted protein
MA0470 NOCOG	predicted protein
MA0471 NOCOG	predicted protein
MA0473 NOCOG	predicted protein
MA0474 NOCOG	predicted protein
MA0475 NOCOG	predicted protein
MA0478 NOCOG	predicted protein
MA0483 NOCOG	predicted protein
MA0486 COG3384 (S)	conserved hypothetical protein
MA0489 NOCOG	predicted protein
MA0492 NOCOG	predicted protein
MA0494 NOCOG	predicted protein
MA0495 NOCOG	predicted protein
MA0496 NOCOG	predicted protein
MA0498 NOCOG	predicted protein
MA0501 NOCOG	predicted protein
MA0502 NOCOG	predicted protein
MA0503 NOCOG	predicted protein
MA0505 NOCOG	predicted protein
MA0506 COG0068 (O)	predicted protein
MA0515 NOCOG	predicted protein
MA0524 NOCOG	conserved hypothetical protein
MA0553 NOCOG	predicted protein
MA0565 NOCOG	predicted protein
MA0566 NOCOG	predicted protein
MA0568 NOCOG	predicted protein
MA0570 NOCOG	predicted protein
MA0571 NOCOG	predicted protein
MA0579 COG2801 (L)	transposase

MA0580	NOCOG	predicted protein
MA0589	NOCOG	predicted protein
MA0607	NOCOG	predicted protein
MA0614	COG2084 (I)	3-hydroxyisobutyrate dehydrogenase
MA0615	COG0477 (GEPR)	transmembrane transport protein (ditE homolog)
MA0625	NOCOG	predicted protein
MA0643	COG2260 (J)	conserved hypothetical protein
MA0657	COG1477 (H)	membrane protein
MA0658	NOCOG	predicted protein
MA0659	COG4656 (C)	nitrogen fixation protein
MA0660	COG4658 (C)	nitrogen fixation protein
MA0661	COG4659 (C)	nitrogen fixation protein
MA0662	COG4660 (C)	nitrogen fixation protein
MA0663	COG4657 (C)	nitrogen fixation protein
MA0665	COG4818 (S)	conserved hypothetical protein
MA0684	COG1592 (C)	conserved hypothetical protein
MA0686	NOCOG	predicted protein
MA0702	NOCOG	predicted protein
MA0707	COG4122 (R)	caffeoyl-CoA O-methyltransferase
MA0710	NOCOG	predicted protein
MA0716	NOCOG	predicted protein
MA0731	NOCOG	predicted protein
MA0736	COG0596 (R)	predicted protein
MA0737	NOCOG	hypothetical protein
MA0738	NOCOG	predicted protein
MA0740	NOCOG	predicted protein
MA0741	NOCOG	predicted protein
MA0742	NOCOG	predicted protein
MA0745	NOCOG	predicted protein
MA0753	NOCOG	predicted protein
MA0754	NOCOG	predicted protein
MA0755	COG3049 (M)	choloylglycine hydrolase
MA0756	NOCOG	predicted protein
MA0761	COG1574 (R)	exoenzymes regulatory protein precursor
MA0763	NOCOG	predicted protein
MA0768	NOCOG	predicted protein
MA0773	COG0604 (CR)	NADPH: quinone reductase
MA0774	NOCOG	predicted protein
MA0778	NOCOG	predicted protein
MA0785	NOCOG	proteophosphoglycan
MA0786	NOCOG	predicted protein
MA0792	COG1834 (E)	arginine deiminase
MA0793	NOCOG	predicted protein
MA0801	NOCOG	predicted protein
MA0804	COG0500 (QR)	conserved hypothetical protein
MA0806	NOCOG	conserved hypothetical protein
MA0825	NOCOG	predicted protein
MA0828	NOCOG	predicted protein
MA0852	COG1277 (R)	bacitracin ABC transporter, permease protein
MA0854	COG0454 (KR)	N-acetyltransferase
MA0861	NOCOG	predicted protein
MA0865	NOCOG	hypothetical protein (multi-domain)
MA0886	NOCOG	predicted protein
MA0901	COG0733 (R)	sodium- and chloride-dependent transporter
MA0907	NOCOG	alkaline serine proteinase I precursor
MA0918	COG0454 (KR)	acetyltransferase (GNAT) family
MA0971	COG0735 (P)	ferric uptake regulator
MA0972	COG0376 (P)	catalase
MA0973	NOCOG	predicted protein
MA0979	NOCOG	conserved hypothetical protein
MA0980	NOCOG	predicted protein
MA0982	COG1487 (R)	PIN protein
MA0983	NOCOG	predicted protein
MA0987	COG1994 (R)	conserved hypothetical protein
MA0991	COG3631 (R)	predicted protein
MA0992	NOCOG	predicted protein
MA0995	COG3910 (R)	conserved hypothetical protein
MA0998	NOCOG	predicted protein
MA1001	COG0114 (C)	fumarate hydratase
MA1002	COG0281 (C)	malate dehydrogenase (oxaloacetate decarboxylating) (NADP+)
MA1004	NOCOG	predicted protein
MA1005	COG0735 (P)	ferric uptake regulator
MA1006	COG1271 (C)	cytochrome d ubiquinol oxidase, subunit I
MA1007	COG1294 (C)	cytochrome d ubiquinol oxidase, subunit II
MA1034	COG2801 (L)	transposase
MA1038	NOCOG	predicted protein
MA1039	NOCOG	predicted protein
MA1044	COG3293 (L)	transposase
MA1100	NOCOG	conserved hypothetical protein
MA1101	COG0789 (K)	transcriptional regulator
MA1107	NOCOG	predicted protein
MA1115	NOCOG	predicted protein
MA1116	COG2801 (L)	transposase
MA1123	NOCOG	predicted protein
MA1124	NOCOG	predicted protein
MA1125	COG2801 (L)	transposase
MA1133	COG4832 (S)	conserved hypothetical protein
MA1135	NOCOG	predicted protein
MA1154	NOCOG	predicted protein
MA1158	COG1670 (J)	acetyltransferase (GNAT) family protein
MA1159	NOCOG	predicted protein
MA1161	COG0477 (GEPR)	macrolide-efflux protein

MA1169	NOCOG	predicted protein
MA1178	COG0500 (QR)	conserved hypothetical protein
MA1181	NOCOG	predicted protein
MA1188	NOCOG	predicted protein
MA1194	COG4908 (R)	predicted protein
MA1202	NOCOG	predicted protein
MA1207	NOCOG	conserved hypothetical protein
MA1209	NOCOG	nitrogenase, subunit delta
MA1217	NOCOG	nitrogenase, subunit delta
MA1221	COG0535 (R)	conserved hypothetical protein
MA1224	NOCOG	conserved hypothetical protein
MA1228	COG0378 (OK)	urease
MA1229	COG1840 (P)	predicted protein
MA1240	NOCOG	predicted protein
MA1242	NOCOG	predicted protein
MA1253	NOCOG	predicted protein
MA1254	NOCOG	predicted protein
MA1271	COG1840 (P)	predicted protein
MA1272	COG0378 (OK)	conserved hypothetical protein
MA1283	NOCOG	predicted protein
MA1284	NOCOG	predicted protein
MA1285	NOCOG	predicted protein
MA1287	NOCOG	predicted protein
MA1294	NOCOG	predicted protein
MA1300	NOCOG	predicted protein
MA1301	NOCOG	predicted protein
MA1303	NOCOG	predicted protein
MA1304	NOCOG	predicted protein
MA1323	NOCOG	predicted protein
MA1327	NOCOG	conserved hypothetical protein
MA1329	NOCOG	predicted protein
MA1330	COG3350 (S)	predicted protein
MA1331	COG0730 (R)	conserved hypothetical protein
MA1332	NOCOG	predicted protein
MA1333	NOCOG	predicted protein
MA1335	NOCOG	predicted protein
MA1345	COG1680 (V)	fmtA-like protein
MA1349	NOCOG	predicted protein
MA1356	NOCOG	acidic integral membrane protein
MA1359	NOCOG	predicted protein
MA1365	NOCOG	predicted protein
MA1375	NOCOG	predicted protein
MA1379	COG2801 (L)	integrase core domain
MA1398	NOCOG	predicted protein
MA1399	NOCOG	predicted protein
MA1400	COG4883 (S)	conserved hypothetical protein
MA1402	NOCOG	hypothetical protein (multi-domain)
MA1404	NOCOG	predicted protein
MA1405	COG0398 (S)	predicted protein
MA1406	NOCOG	predicted protein
MA1407	NOCOG	predicted protein
MA1410	NOCOG	predicted protein
MA1415	NOCOG	predicted protein
MA1434	COG2801 (L)	transposase
MA1443	NOCOG	predicted protein
MA1445	NOCOG	predicted protein
MA1451	NOCOG	predicted protein
MA1453	NOCOG	predicted protein
MA1459	NOCOG	hypothetical protein (inconsistent evidence)
MA1465	NOCOG	predicted protein
MA1467	NOCOG	predicted protein
MA1494	NOCOG	predicted protein
MA1511	NOCOG	predicted protein
MA1518	COG0733 (R)	sodium- and chloride-dependent noradrenaline transporter
MA1538	COG1277 (R)	conserved hypothetical protein
MA1539	COG1277 (R)	conserved hypothetical protein
MA1542	NOCOG	predicted protein
MA1544	COG1277 (R)	conserved hypothetical protein
MA1545	COG1277 (R)	conserved hypothetical protein
MA1546	COG1277 (R)	conserved hypothetical protein
MA1547	COG1277 (R)	conserved hypothetical protein
MA1548	COG0823 (U)	predicted protein
MA1556	COG4983 (S)	hypothetical protein (multi-domain)
MA1557	NOCOG	predicted protein
MA1558	NOCOG	predicted protein
MA1559	NOCOG	predicted protein
MA1560	NOCOG	predicted protein
MA1561	NOCOG	predicted protein
MA1562	COG2114 (T)	predicted protein
MA1563	NOCOG	predicted protein
MA1564	NOCOG	predicted protein
MA1565	NOCOG	predicted protein
MA1570	NOCOG	predicted protein
MA1575	NOCOG	predicted protein
MA1577	NOCOG	predicted protein
MA1580	NOCOG	predicted protein
MA1585	NOCOG	predicted protein
MA1592	COG0477 (GEPR)	predicted protein
MA1594	NOCOG	predicted protein
MA1595	NOCOG	predicted protein
MA1602	COG0494 (LR)	hypothetical protein (inconsistent evidence)
MA1604	NOCOG	predicted protein

MA1609	NOCOG	hypothetical protein
MA1612	NOCOG	predicted protein
MA1619	NOCOG	predicted protein
MA1621	NOCOG	predicted protein
MA1625	NOCOG	predicted protein
MA1626	NOCOG	predicted protein
MA1637	COG3293 (L)	transposase
MA1641	COG0366 (G)	alpha-amylase family protein
MA1643	NOCOG	predicted protein
MA1644	NOCOG	predicted protein
MA1647	COG1708 (R)	predicted protein
MA1648	COG0494 (LR)	mutT/NUDIX hydrolases family protein
MA1662	NOCOG	predicted protein
MA1663	NOCOG	predicted protein
MA1670	COG0500 (QR)	conserved hypothetical protein
MA1673	NOCOG	predicted protein
MA1681	NOCOG	predicted protein
MA1688	NOCOG	predicted protein
MA1693	NOCOG	predicted protein
MA1695	NOCOG	predicted protein
MA1698	NOCOG	predicted protein
MA1700	NOCOG	predicted protein
MA1705	NOCOG	predicted protein
MA1707	NOCOG	predicted protein
MA1720	NOCOG	predicted protein
MA1732	NOCOG	predicted protein
MA1734	NOCOG	predicted protein
MA1737	NOCOG	predicted protein
MA1757	COG3385 (L)	conserved hypothetical protein
MA1765	NOCOG	conserved hypothetical protein
MA1780	NOCOG	predicted protein
MA1785	NOCOG	predicted protein
MA1787	NOCOG	predicted protein
MA1795	NOCOG	predicted protein
MA1798	NOCOG	predicted protein
MA1803	NOCOG	predicted protein
MA1807	NOCOG	predicted protein
MA1808	COG5323 (S)	predicted protein
MA1809	NOCOG	predicted protein
MA1814	NOCOG	predicted protein
MA1823	NOCOG	predicted protein
MA1829	NOCOG	predicted protein
MA1830	NOCOG	predicted protein
MA1832	NOCOG	predicted protein
MA1841	NOCOG	predicted protein
MA1843	NOCOG	predicted protein
MA1845	NOCOG	predicted protein
MA1846	NOCOG	predicted protein
MA1849	COG5266 (P)	predicted protein
MA1851	NOCOG	predicted protein
MA1859	NOCOG	predicted protein
MA1864	COG0210 (L)	helicase
MA1865	COG0210 (L)	predicted protein
MA1871	NOCOG	predicted protein
MA1875	NOCOG	predicted protein
MA1876	NOCOG	predicted protein
MA1877	NOCOG	predicted protein
MA1881	NOCOG	predicted protein
MA1884	NOCOG	predicted protein
MA1886	NOCOG	predicted protein
MA1887	NOCOG	predicted protein
MA1888	NOCOG	predicted protein
MA1889	NOCOG	predicted protein
MA1890	COG2801 (L)	transposase
MA1891	NOCOG	predicted protein
MA1893	NOCOG	predicted protein
MA1894	NOCOG	predicted protein
MA1896	NOCOG	predicted protein
MA1897	NOCOG	predicted protein
MA1898	NOCOG	predicted protein
MA1899	NOCOG	predicted protein
MA1900	NOCOG	predicted protein
MA1905	COG1966 (T)	carbon starvation protein A
MA1908	NOCOG	predicted protein
MA1909	COG1898 (M)	hypothetical protein (inconsistent evidence)
MA1917	NOCOG	predicted protein
MA1926	COG2389 (R)	conserved hypothetical protein
MA1929	COG1353 (R)	conserved hypothetical protein
MA1930	COG2801 (L)	transposase
MA1931	COG1353 (R)	conserved hypothetical protein
MA1932	NOCOG	predicted protein
MA1933	COG1337 (L)	conserved hypothetical protein
MA1934	COG1567 (L)	conserved hypothetical protein
MA1935	COG1332 (L)	conserved hypothetical protein
MA1952	NOCOG	predicted protein
MA1960	NOCOG	predicted protein
MA1965	NOCOG	predicted protein
MA1969	COG2027 (M)	serine-type D-Ala-D-Ala carboxypeptidase
MA1983	NOCOG	predicted protein
MA1984	NOCOG	predicted protein
MA1987	NOCOG	predicted protein
MA1990	NOCOG	predicted protein

MA1992	NOCOG	predicted protein
MA1993	NOCOG	predicted protein
MA1999	COG0659 (P)	sulfate transporter
MA2000	COG1523 (G)	glycogen debranching enzyme
MA2001	COG3590 (O)	endothelin converting enzyme homolog PepO
MA2002	COG0596 (R)	conserved hypothetical protein
MA2006	NOCOG	predicted protein
MA2018	NOCOG	predicted protein
MA2020	COG2801 (L)	retroviral pol related endonuclease
MA2022	NOCOG	predicted protein
MA2023	NOCOG	predicted protein
MA2025	NOCOG	predicted protein
MA2026	COG0582 (L)	predicted protein
MA2029	NOCOG	predicted protein
MA2035	NOCOG	predicted protein
MA2036	NOCOG	predicted protein
MA2038	NOCOG	predicted protein
MA2039	NOCOG	predicted protein
MA2040	NOCOG	conserved hypothetical protein
MA2041	COG5529 (R)	predicted protein
MA2042	NOCOG	predicted protein
MA2043	COG3209 (M)	hypothetical protein (multi-domain)
MA2044	NOCOG	predicted protein
MA2045	COG3209 (M)	hypothetical protein (multi-domain)
MA2048	NOCOG	predicted protein
MA2049	NOCOG	predicted protein
MA2053	COG4929 (S)	predicted protein
MA2054	NOCOG	predicted protein
MA2062	NOCOG	predicted protein
MA2065	NOCOG	predicted protein
MA2066	COG0464 (O)	hypothetical protein (multi-domain)
MA2067	NOCOG	predicted protein
MA2069	NOCOG	predicted protein
MA2070	COG4104 (S)	predicted protein
MA2073	NOCOG	conserved hypothetical protein
MA2083	NOCOG	predicted protein
MA2089	NOCOG	predicted protein
MA2092	NOCOG	predicted protein
MA2093	COG1695 (K)	predicted protein
MA2094	NOCOG	predicted protein
MA2101	NOCOG	predicted protein
MA2102	COG3344 (L)	hypothetical protein (multi-domain)
MA2104	NOCOG	predicted protein
MA2107	NOCOG	predicted protein
MA2112	NOCOG	hypothetical protein
MA2115	NOCOG	predicted protein
MA2117	NOCOG	predicted protein
MA2118	COG4268 (V)	conserved hypothetical protein
MA2119	COG1401 (V)	hypothetical protein (multi-domain)
MA2123	NOCOG	predicted protein
MA2124	NOCOG	predicted protein
MA2125	NOCOG	predicted protein
MA2126	NOCOG	predicted protein
MA2127	NOCOG	predicted protein
MA2128	NOCOG	predicted protein
MA2134	COG1598 (S)	predicted protein
MA2135	COG1724 (N)	conserved hypothetical protein
MA2138	NOCOG	predicted protein
MA2139	COG3593 (L)	hypothetical protein (multi-domain)
MA2141	NOCOG	predicted protein
MA2154	NOCOG	predicted protein
MA2157	NOCOG	predicted protein
MA2164	COG0500 (QR)	predicted protein
MA2165	NOCOG	predicted protein
MA2171	COG0707 (M)	capsular polysaccharide biosynthesis protein
MA2172	COG5017 (S)	beta-1,4-galactosyltransferase
MA2176	COG0125 (F)	predicted protein
MA2177	NOCOG	predicted protein
MA2179	NOCOG	predicted protein
MA2185	NOCOG	predicted protein
MA2187	COG2801 (L)	transposase
MA2189	COG2801 (L)	transposase
MA2192	NOCOG	predicted protein
MA2195	NOCOG	predicted protein
MA2196	NOCOG	predicted protein
MA2198	COG3271 (R)	predicted protein
MA2203	COG5337 (M)	cellulosomal protein
MA2204	NOCOG	predicted protein
MA2205	NOCOG	predicted protein
MA2213	NOCOG	predicted protein
MA2216	NOCOG	predicted protein
MA2222	NOCOG	predicted protein
MA2223	NOCOG	predicted protein
MA2225	NOCOG	predicted protein
MA2226	NOCOG	predicted protein
MA2227	NOCOG	predicted protein
MA2228	NOCOG	predicted protein
MA2229	NOCOG	predicted protein
MA2230	NOCOG	predicted protein
MA2231	COG0582 (L)	site-specific recombinase
MA2232	NOCOG	predicted protein
MA2233	NOCOG	conserved hypothetical protein

MA2234	NOCOG	predicted protein
MA2236	NOCOG	predicted protein
MA2237	NOCOG	predicted protein
MA2240	NOCOG	predicted protein
MA2242	NOCOG	predicted protein
MA2248	NOCOG	predicted protein
MA2251	COG0392 (S)	conserved hypothetical protein
MA2260	NOCOG	predicted protein
MA2262	COG4831 (S)	conserved hypothetical protein
MA2263	COG0477 (GEPR)	hypothetical protein (inconsistent evidence)
MA2265	COG1573 (L)	predicted protein
MA2269	NOCOG	predicted protein
MA2271	NOCOG	conserved hypothetical protein
MA2274	NOCOG	predicted protein
MA2283	COG1335 (Q)	conserved hypothetical protein
MA2288	NOCOG	predicted protein
MA2290	COG5470 (S)	predicted protein
MA2293	NOCOG	predicted protein
MA2296	NOCOG	predicted protein
MA2301	NOCOG	hypothetical protein with leucine-rich tandem repeats
MA2302	NOCOG	predicted protein
MA2303	NOCOG	predicted protein
MA2305	NOCOG	predicted protein
MA2306	NOCOG	predicted protein
MA2307	COG1555 (L)	predicted protein
MA2308	NOCOG	predicted protein
MA2314	NOCOG	predicted protein
MA2315	NOCOG	predicted protein
MA2318	NOCOG	predicted protein
MA2323	NOCOG	conserved hypothetical protein
MA2325	COG5361 (S)	predicted protein
MA2327	NOCOG	predicted protein
MA2328	NOCOG	conserved hypothetical protein
MA2329	NOCOG	conserved hypothetical protein
MA2330	NOCOG	predicted protein
MA2334	NOCOG	predicted protein
MA2336	NOCOG	predicted protein
MA2338	COG0456 (R)	acetyltransferase (GNAT) family
MA2339	NOCOG	conserved hypothetical protein
MA2353	NOCOG	conserved hypothetical protein
MA2355	NOCOG	predicted protein
MA2356	NOCOG	predicted protein
MA2359	NOCOG	hypothetical protein (multi-domain)
MA2360	NOCOG	predicted protein
MA2361	NOCOG	predicted protein
MA2362	COG4717 (S)	conserved hypothetical protein
MA2371	NOCOG	conserved hypothetical protein
MA2378	NOCOG	predicted protein
MA2392	NOCOG	predicted protein
MA2393	NOCOG	predicted protein
MA2394	NOCOG	predicted protein
MA2395	NOCOG	predicted protein
MA2396	NOCOG	predicted protein
MA2398	COG2151 (R)	conserved hypothetical protein
MA2402	NOCOG	predicted protein
MA2405	NOCOG	predicted protein
MA2406	COG3464 (L)	transposase
MA2411	NOCOG	predicted protein
MA2416	NOCOG	predicted protein
MA2417	NOCOG	conserved hypothetical protein
MA2420	NOCOG	predicted protein
MA2421	NOCOG	bacterial regulatory protein, arsR family
MA2422	COG2032 (P)	superoxide dismutase
MA2433	COG0224 (C)	H(+)-transporting ATP synthase, subunit gamma
MA2435	COG0711 (C)	H(+)-transporting ATP synthase, subunit B
MA2436	COG0636 (C)	H(+)-transporting ATP synthase, subunit C
MA2437	COG0356 (C)	H(+)-transporting ATP synthase, subunit A
MA2438	NOCOG	conserved hypothetical protein
MA2439	NOCOG	H(+)-transporting ATP synthase, gene 1
MA2440	COG0355 (C)	H(+)-transporting ATP synthase, subunit epsilon
MA2448	COG1226 (P)	potassium channel protein
MA2453	NOCOG	predicted protein
MA2455	COG3853 (P)	klaA protein (telA protein)
MA2456	NOCOG	predicted protein
MA2459	NOCOG	predicted protein
MA2460	NOCOG	predicted protein
MA2461	NOCOG	predicted protein
MA2463	COG0458 (EF)	predicted protein
MA2464	COG2120 (S)	conserved hypothetical protein
MA2466	NOCOG	predicted protein
MA2469	COG2128 (S)	hypothetical protein (multi-domain)
MA2483	NOCOG	predicted protein
MA2487	NOCOG	predicted protein
MA2492	NOCOG	predicted protein
MA2493	COG1309 (K)	bacterial regulatory protein, tetR family
MA2504	NOCOG	predicted protein
MA2505	NOCOG	predicted protein
MA2522	NOCOG	predicted protein
MA2523	NOCOG	predicted protein
MA2524	NOCOG	predicted protein
MA2526	NOCOG	beta/gama crystallin family protein
MA2528	NOCOG	cellulosomal protein

MA2529	COG3227 (E)	extracellular protease
MA2530	NOCOG	predicted protein
MA2534	NOCOG	conserved hypothetical protein
MA2538	NOCOG	predicted protein
MA2541	NOCOG	conserved hypothetical protein
MA2542	NOCOG	predicted protein
MA2552	NOCOG	predicted protein
MA2577	NOCOG	predicted protein
MA2581	NOCOG	predicted protein
MA2583	NOCOG	predicted protein
MA2585	NOCOG	predicted protein
MA2586	NOCOG	predicted protein
MA2587	NOCOG	predicted protein
MA2588	NOCOG	predicted protein
MA2589	NOCOG	predicted protein
MA2590	NOCOG	predicted protein
MA2591	NOCOG	predicted protein
MA2592	NOCOG	predicted protein
MA2593	NOCOG	predicted protein
MA2594	NOCOG	predicted protein
MA2595	NOCOG	predicted protein
MA2596	NOCOG	predicted protein
MA2597	NOCOG	predicted protein
MA2598	COG1989 (NOU)	predicted protein
MA2599	COG0582 (L)	phage integrase
MA2600	NOCOG	predicted protein
MA2601	NOCOG	predicted protein
MA2602	NOCOG	conserved hypothetical protein
MA2603	NOCOG	predicted protein
MA2604	NOCOG	predicted protein
MA2605	NOCOG	predicted protein
MA2606	NOCOG	predicted protein
MA2607	NOCOG	predicted protein
MA2608	NOCOG	predicted protein
MA2610	NOCOG	predicted protein
MA2614	NOCOG	predicted protein
MA2617	COG1895 (S)	predicted protein
MA2618	COG1708 (R)	conserved hypothetical protein
MA2619	NOCOG	predicted protein
MA2620	NOCOG	conserved hypothetical protein
MA2621	NOCOG	hypthetical protein (inconsistent evidence)
MA2629	COG0657 (I)	triacylglycerol lipase
MA2632	COG0385 (R)	transmembrane transport protein
MA2634	NOCOG	predicted protein
MA2636	NOCOG	conserved hypothetical protein
MA2637	NOCOG	predicted protein
MA2638	NOCOG	predicted protein
MA2639	NOCOG	predicted protein
MA2640	NOCOG	predicted protein
MA2643	NOCOG	predicted protein
MA2652	COG1266 (R)	predicted protein
MA2654	NOCOG	predicted protein
MA2668	NOCOG	predicted protein
MA2679	COG3385 (L)	transposase
MA2682	NOCOG	predicted protein
MA2693	NOCOG	predicted protein
MA2697	NOCOG	predicted protein
MA2700	NOCOG	predicted protein
MA2701	COG3039 (L)	transposase
MA2702	NOCOG	hypothetical protein (multi-domain)
MA2703	NOCOG	hypothetical protein (multi-domain)
MA2709	NOCOG	predicted protein
MA2710	NOCOG	predicted protein
MA2711	NOCOG	conserved hypothetical protein
MA2712	NOCOG	predicted protein
MA2713	NOCOG	predicted protein
MA2720	COG0031 (E)	cysteine synthase
MA2727	NOCOG	conserved hypothetical protein
MA2731	NOCOG	predicted protein
MA2735	NOCOG	predicted protein
MA2743	COG0596 (R)	predicted protein
MA2744	NOCOG	predicted protein
MA2747	NOCOG	predicted protein
MA2749	NOCOG	predicted protein
MA2753	NOCOG	predicted protein
MA2755	COG1600 (C)	predicted protein
MA2758	NOCOG	predicted protein
MA2760	COG3315 (Q)	hypothetical protein (multi-domain)
MA2765	COG3315 (Q)	hypothetical protein (multi-domain)
MA2767	NOCOG	predicted protein
MA2768	NOCOG	predicted protein
MA2778	COG1120 (PH)	predicted protein
MA2782	COG1309 (K)	hypothetical protein (multi-domain)
MA2787	NOCOG	predicted protein
MA2788	NOCOG	predicted protein
MA2789	NOCOG	predicted protein
MA2790	NOCOG	predicted protein
MA2791	NOCOG	predicted protein
MA2792	COG0582 (L)	predicted protein
MA2798	NOCOG	predicted protein
MA2801	NOCOG	predicted protein
MA2802	NOCOG	predicted protein

MA2803	NOCOG	predicted protein
MA2804	NOCOG	predicted protein
MA2805	NOCOG	predicted protein
MA2806	NOCOG	predicted protein
MA2808	NOCOG	hypothetical protein (multi-domain)
MA2809	NOCOG	predicted protein
MA2810	NOCOG	predicted protein
MA2811	COG0582 (L)	site-specific recombinase
MA2812	NOCOG	predicted protein
MA2819	NOCOG	conserved hypothetical protein
MA2824	NOCOG	predicted protein
MA2830	NOCOG	predicted protein
MA2832	NOCOG	predicted protein
MA2836	COG2801 (L)	transposase
MA2852	NOCOG	predicted protein
MA2857	COG3385 (L)	conserved hypothetical protein
MA2858	NOCOG	predicted protein
MA2863	NOCOG	predicted protein
MA2864	NOCOG	predicted protein
MA2865	NOCOG	predicted protein
MA2869	NOCOG	albumin-binding protein
MA2876	NOCOG	predicted protein
MA2883	NOCOG	conserved hypothetical protein
MA2887	NOCOG	predicted protein
MA2898	NOCOG	predicted protein
MA2899	COG2334 (R)	conserved hypothetical protein
MA2900	NOCOG	predicted protein
MA2905	COG0500 (QR)	predicted protein
MA2906	COG0500 (QR)	conserved hypothetical protein
MA2908	COG1858 (P)	cytochrome-c peroxidase
MA2913	NOCOG	predicted protein
MA2925	COG1858 (P)	methylamine metabolism protein
MA2928	NOCOG	predicted protein
MA2929	NOCOG	predicted protein
MA2930	NOCOG	predicted protein
MA2931	COG2801 (L)	transposase
MA2932	NOCOG	predicted protein
MA2936	COG0535 (R)	conserved hypothetical protein, nifB-nifX homolog
MA2940	COG2801 (L)	transposase
MA2941	COG1680 (V)	predicted protein
MA2943	NOCOG	predicted protein
MA2949	NOCOG	predicted protein
MA2953	COG1574 (R)	conserved hypothetical protein
MA2954	COG3464 (L)	transposase
MA2955	NOCOG	predicted protein
MA2964	NOCOG	predicted protein
MA2974	NOCOG	predicted protein
MA2977	NOCOG	predicted protein
MA2980	NOCOG	predicted protein
MA2981	NOCOG	predicted protein
MA2983	COG2801 (L)	transposase
MA2998	NOCOG	predicted protein
MA3007	NOCOG	predicted protein
MA3008	COG0406 (G)	phosphoglycerate mutase family protein
MA3011	NOCOG	predicted protein
MA3027	NOCOG	predicted protein
MA3028	NOCOG	predicted protein
MA3032	COG0296 (G)	alpha-amylase
MA3036	NOCOG	predicted protein
MA3037	NOCOG	prion protein precursor
MA3038	NOCOG	predicted protein
MA3039	COG2041 (R)	sulfite oxidase
MA3041	NOCOG	predicted protein
MA3053	NOCOG	conserved hypothetical protein
MA3074	NOCOG	predicted protein
MA3086	NOCOG	predicted protein
MA3089	NOCOG	predicted protein
MA3090	NOCOG	predicted protein
MA3099	COG0467 (T)	predicted protein
MA3100	NOCOG	predicted protein
MA3106	NOCOG	predicted protein
MA3109	NOCOG	predicted protein
MA3110	NOCOG	predicted protein
MA3111	NOCOG	predicted protein
MA3112	NOCOG	predicted protein
MA3114	COG0454 (KR)	acetyltransferase (GNAT) family
MA3115	NOCOG	predicted protein
MA3116	NOCOG	predicted protein
MA3131	NOCOG	predicted protein
MA3132	NOCOG	predicted protein
MA3141	NOCOG	predicted protein
MA3144	NOCOG	predicted protein
MA3156	NOCOG	predicted protein
MA3157	NOCOG	predicted protein
MA3158	COG2801 (L)	transposase
MA3164	NOCOG	predicted protein
MA3175	NOCOG	predicted protein
MA3179	COG3158 (P)	potassium uptake protein kup
MA3180	NOCOG	predicted protein
MA3182	COG1657 (I)	hypothetical protein
MA3184	NOCOG	predicted protein
MA3186	NOCOG	predicted protein

MA3188	NOCOG	predicted protein
MA3196	NOCOG	predicted protein
MA3202	NOCOG	predicted protein
MA3207	COG0122 (L)	DNA-3-methyladenine glycosylase II
MA3208	COG0778 (C)	predicted protein
MA3210	NOCOG	predicted protein
MA3213	COG0675 (L)	predicted protein
MA3214	NOCOG	microbial collagenase
MA3215	NOCOG	predicted protein
MA3216	COG2234 (R)	bacterial leucyl aminopeptidase
MA3219	NOCOG	predicted protein
MA3220	NOCOG	predicted protein
MA3223	NOCOG	predicted protein
MA3224	NOCOG	predicted protein
MA3225	NOCOG	predicted protein
MA3226	NOCOG	predicted protein
MA3227	NOCOG	predicted protein
MA3228	NOCOG	predicted protein
MA3230	NOCOG	predicted protein
MA3231	NOCOG	predicted protein
MA3232	NOCOG	predicted protein
MA3233	NOCOG	predicted protein
MA3234	NOCOG	predicted protein
MA3240	NOCOG	predicted protein
MA3244	NOCOG	predicted protein
MA3249	NOCOG	predicted protein
MA3251	NOCOG	predicted protein
MA3255	NOCOG	predicted protein
MA3266	NOCOG	predicted protein
MA3267	COG0524 (G)	conserved hypothetical protein
MA3272	NOCOG	predicted protein
MA3274	NOCOG	predicted protein
MA3276	NOCOG	predicted protein
MA3295	COG3478 (R)	conserved hypothetical protein
MA3301	NOCOG	predicted protein
MA3303	NOCOG	predicted protein
MA3308	NOCOG	predicted protein
MA3316	COG3422 (S)	conserved hypothetical protein
MA3318	NOCOG	predicted protein
MA3319	NOCOG	predicted protein
MA3329	COG3385 (L)	predicted protein
MA3330	NOCOG	conserved hypothetical protein
MA3331	NOCOG	predicted protein
MA3333	NOCOG	predicted protein
MA3339	NOCOG	predicted protein
MA3340	NOCOG	predicted protein
MA3341	NOCOG	predicted protein
MA3349	NOCOG	hypothetical protein
MA3350	NOCOG	predicted protein
MA3351	NOCOG	predicted protein
MA3356	NOCOG	predicted protein
MA3363	NOCOG	conserved hypothetical protein
MA3364	NOCOG	predicted protein
MA3372	NOCOG	conserved hypothetical protein
MA3377	COG4690 (E)	conserved hypothetical protein
MA3380	NOCOG	conserved hypothetical protein
MA3383	COG0393 (S)	conserved hypothetical protein
MA3384	NOCOG	predicted protein
MA3385	NOCOG	predicted protein
MA3386	NOCOG	predicted protein
MA3401	NOCOG	predicted protein
MA3402	COG1618 (F)	conserved hypothetical protein
MA3403	NOCOG	predicted protein
MA3409	NOCOG	predicted protein
MA3412	NOCOG	predicted protein
MA3413	COG4818 (S)	conserved hypothetical protein
MA3414	NOCOG	predicted protein
MA3417	NOCOG	predicted protein
MA3421	NOCOG	predicted protein
MA3435	NOCOG	predicted protein
MA3443	NOCOG	hypothetical protein (multi-domain)
MA3444	NOCOG	predicted protein
MA3450	NOCOG	predicted protein
MA3460	NOCOG	predicted protein
MA3469	NOCOG	predicted protein
MA3471	NOCOG	hypothetical protein (multi-domain)
MA3472	COG3315 (Q)	hypthetical protein (inconsistent evidence)
MA3474	NOCOG	predicted protein
MA3480	NOCOG	predicted protein
MA3483	NOCOG	predicted protein
MA3492	COG0775 (F)	predicted protein
MA3493	NOCOG	predicted protein
MA3503	NOCOG	predicted protein
MA3508	NOCOG	predicted protein
MA3517	NOCOG	predicted protein
MA3518	NOCOG	predicted protein
MA3521	NOCOG	predicted protein
MA3523	NOCOG	predicted protein
MA3532	COG0384 (R)	phenazine biosynthesis protein
MA3533	NOCOG	predicted protein
MA3536	NOCOG	predicted protein
MA3540	NOCOG	predicted protein

MA3541	COG0454 (KR)	predicted protein
MA3558	NOCOG	predicted protein
MA3569	NOCOG	predicted protein
MA3571	NOCOG	predicted protein
MA3579	NOCOG	predicted protein
MA3580	NOCOG	predicted protein
MA3603	NOCOG	predicted protein
MA3604	COG2866 (E)	carboxypeptidase A
MA3615	NOCOG	predicted protein
MA3616	NOCOG	predicted protein
MA3625	NOCOG	predicted protein
MA3629	NOCOG	predicted protein
MA3649	NOCOG	predicted protein
MA3650	NOCOG	predicted protein
MA3651	NOCOG	predicted protein
MA3652	COG1834 (E)	amidinotransferase
MA3653	NOCOG	predicted protein
MA3654	COG0454 (KR)	acetyltransferase (GNAT) family
MA3656	NOCOG	predicted protein
MA3659	NOCOG	predicted protein
MA3661	NOCOG	predicted protein
MA3663	NOCOG	predicted protein
MA3666	COG1688 (L)	conserved hypothetical protein
MA3667	COG1857 (L)	conserved hypothetical protein
MA3668	NOCOG	conserved hypothetical protein
MA3669	COG1343 (L)	conserved hypothetical protein
MA3671	COG1468 (L)	conserved hypothetical protein
MA3672	COG1583 (L)	conserved hypothetical protein
MA3681	NOCOG	predicted protein
MA3697	NOCOG	predicted protein
MA3701	NOCOG	predicted protein
MA3712	NOCOG	predicted protein
MA3724	NOCOG	predicted protein
MA3725	NOCOG	predicted protein
MA3729	COG0574 (G)	pyruvate,water dikinase
MA4664	NOCOG	conserved hypothetical protein
MA3738	NOCOG	predicted protein
MA3761	NOCOG	predicted protein
MA3763	NOCOG	predicted protein
MA3765	COG0500 (QR)	predicted protein
MA3766	NOCOG	acylneuraminate cytidylyltransferase
MA3767	COG2089 (M)	N-acetylneuraminic acid phosphate synthase
MA3768	NOCOG	predicted protein
MA3773	NOCOG	predicted protein
MA3774	NOCOG	predicted protein
MA3775	NOCOG	predicted protein
MA3785	NOCOG	predicted protein
MA3795	NOCOG	predicted protein
MA3796	NOCOG	predicted protein
MA3797	NOCOG	predicted protein
MA3798	COG3039 (L)	transposase
MA3801	NOCOG	predicted protein
MA3802	NOCOG	predicted protein
MA3803	NOCOG	predicted protein
MA3804	NOCOG	predicted protein
MA3805	COG5280 (S)	hypothetical protein (inconsistent evidence)
MA3806	NOCOG	predicted protein
MA3807	NOCOG	predicted protein
MA3808	NOCOG	predicted protein
MA3809	NOCOG	predicted protein
MA3810	NOCOG	predicted protein
MA3811	NOCOG	predicted protein
MA3812	NOCOG	predicted protein
MA3813	NOCOG	predicted protein
MA3814	NOCOG	predicted protein
MA3815	NOCOG	predicted protein
MA3816	NOCOG	predicted protein
MA3817	COG0100 (J)	predicted protein
MA3818	NOCOG	predicted protein
MA3819	NOCOG	predicted protein
MA3821	NOCOG	predicted protein
MA3822	NOCOG	predicted protein
MA3823	COG5585 (T)	conserved hypothetical protein
MA3824	NOCOG	predicted protein
MA3825	NOCOG	predicted protein
MA3826	NOCOG	terminase
MA3827	NOCOG	predicted protein
MA3828	NOCOG	predicted protein
MA3830	NOCOG	predicted protein
MA3831	NOCOG	predicted protein
MA3832	NOCOG	predicted protein
MA3833	NOCOG	predicted protein
MA3834	NOCOG	predicted protein
MA3835	NOCOG	predicted protein
MA3836	NOCOG	predicted protein
MA3837	NOCOG	predicted protein
MA3838	NOCOG	predicted protein
MA3839	COG0457 (R)	hypothetical protein
MA3840	COG4748 (S)	conserved hypothetical protein
MA3841	COG2801 (L)	transposase
MA3844	NOCOG	conserved hypothetical protein
MA3912	NOCOG	predicted protein

MA3913	NOCOG	predicted protein
MA3923	NOCOG	predicted protein
MA3930	COG0614 (P)	predicted protein
MA3931	COG0614 (P)	predicted protein
MA3935	COG1993 (S)	conserved hypothetical protein
MA3947	COG3439 (S)	conserved hypothetical protein
MA3948	NOCOG	predicted protein
MA3954	COG0454 (KR)	acetyltransferase family
MA3955	NOCOG	predicted protein
MA3970	COG1552 (J)	ribosomal protein L40e
MA3980	NOCOG	predicted protein
MA3982	NOCOG	predicted protein
MA3985	NOCOG	predicted protein
MA3988	COG3464 (L)	transposase
MA4001	NOCOG	predicted protein
MA4003	NOCOG	clostripain
MA4024	NOCOG	hypothetical protein
MA4025	NOCOG	predicted protein
MA4030	NOCOG	predicted protein
MA4037	COG1277 (R)	conserved hypothetical protein
MA4038	NOCOG	predicted protein
MA4049	NOCOG	predicted protein
MA4054	NOCOG	predicted protein
MA4056	NOCOG	predicted protein
MA4057	COG2801 (L)	transposase
MA4073	NOCOG	predicted protein
MA4080	NOCOG	predicted protein
MA4090	NOCOG	predicted protein
MA4096	NOCOG	predicted protein
MA4122	NOCOG	predicted protein
MA4124	COG0582 (L)	site-specific recombinase
MA4127	COG0582 (L)	site-specific recombinase
MA4128	NOCOG	predicted protein
MA4129	NOCOG	predicted protein
MA4130	NOCOG	predicted protein
MA4131	NOCOG	predicted protein
MA4134	NOCOG	predicted protein
MA4135	NOCOG	predicted protein
MA4136	NOCOG	predicted protein
MA4138	NOCOG	predicted protein
MA4139	NOCOG	predicted protein
MA4140	NOCOG	predicted protein
MA4141	NOCOG	predicted protein
MA4142	COG1997 (J)	predicted protein
MA4143	NOCOG	predicted protein
MA4146	NOCOG	predicted protein
MA4148	COG0671 (I)	type 2 phosphatidic acid phosphatase (PAP2) family
MA4168	NOCOG	predicted protein
MA4179	NOCOG	predicted protein
MA4192	NOCOG	predicted protein
MA4210	NOCOG	predicted protein
MA4212	NOCOG	predicted protein
MA4215	NOCOG	predicted protein
MA4229	NOCOG	predicted protein
MA4230	NOCOG	predicted protein
MA4231	NOCOG	predicted protein
MA4270	NOCOG	predicted protein
MA4281	NOCOG	predicted protein
MA4282	NOCOG	predicted protein
MA4283	NOCOG	predicted protein
MA4286	NOCOG	predicted protein
MA4288	NOCOG	predicted protein
MA4290	COG1520 (S)	conserved hypothetical protein
MA4296	NOCOG	predicted protein
MA4298	NOCOG	hypothetical protein
MA4300	NOCOG	predicted protein
MA4302	COG1520 (S)	conserved hypothetical protein
MA4303	NOCOG	predicted protein
MA4307	COG3291 (R)	EF hand family protein (signaling or buffering/transport protein)
MA4308	NOCOG	predicted protein
MA4310	COG1520 (S)	conserved hypothetical protein
MA4311	NOCOG	predicted protein
MA4313	NOCOG	predicted protein
MA4314	COG1520 (S)	conserved hypothetical protein
MA4316	COG1520 (S)	predicted protein
MA4320	COG0842 (V)	conserved hypothetical protein
MA4336	NOCOG	predicted protein
MA4348	NOCOG	predicted protein
MA4358	NOCOG	predicted protein
MA4359	COG1277 (R)	conserved hypothetical protein
MA4361	COG1470 (S)	conserved hypothetical protein
MA4385	NOCOG	predicted protein
MA4387	NOCOG	predicted protein
MA4388	NOCOG	predicted protein
MA4390	NOCOG	predicted protein
MA4411	COG1335 (Q)	isochorismatase
MA4412	NOCOG	predicted protein
MA4439	NOCOG	predicted protein
MA4441	NOCOG	predicted protein
MA4443	NOCOG	predicted protein
MA4447	NOCOG	predicted protein
MA4458	NOCOG	predicted protein

```
MA4463 COG2801 (L)      retroviral pol related endonuclease
MA4477 NOCOG            predicted protein
MA4480 COG0640 (K)      predicted protein
MA4482 NOCOG            predicted protein
MA4483 COG2801 (L)      retroviral pol related endonuclease
MA4495 NOCOG            predicted protein
MA4497 NOCOG            predicted protein
MA4498 NOCOG            predicted protein
MA4499 NOCOG            predicted protein
MA4508 NOCOG            predicted protein
MA4509 NOCOG            predicted protein
MA4510 NOCOG            predicted protein
MA4512 COG3393 (R)      acetyltransferase (GNAT) family protein
MA4521 COG1765 (O)      predicted protein
MA4526 NOCOG            predicted protein
MA4528 NOCOG            predicted protein
MA4538 NOCOG            predicted protein
MA4557 NOCOG            predicted protein
MA4565 NOCOG            predicted protein
MA4566 COG1320 (P)      multiple resistance and pH regulation related protein G (Na+/H+ antiporter)
MA4567 COG2212 (P)      multiple resistance and pH regulation related protein F (Na+/H+ antiporter)
MA4568 COG1863 (P)      multiple resistance and pH regulation related protein E (Na+/H+ antiporter)
MA4570 COG1006 (P)      multiple resistance and pH regulation related protein C (Na+/H+ antiporter)
MA4665 COG2111 (P)      multiple resistance and pH regulation related protein B (Na+/H+ antiporter)
MA4578 NOCOG            predicted protein
MA4583 COG0310 (P)      predicted protein
MA4603 NOCOG            predicted protein
MA4629 NOCOG            predicted protein
MA4639 NOCOG            predicted protein
MA4644 COG0477 (GEPR)   conserved hypothetical protein
MA4646 NOCOG            predicted protein
MA4657 NOCOG            hypothetical protein
```

A.3.1. COG Gruppen

A.4. Bakterielle Gene nach der *first-hit* Methode in *Ms. mazei*

Genname, COG, (COG-Gruppe), Annotation

```
MM00001 COG0747 (E)     Dipeptide ABC transporter, binding protein
MM00002 COG0747 (E)     Dipeptide ABC transporter, binding protein
MM00004 COG0601 (EP)    Dipeptide ABC transporter, permease protein
MM00005 COG1173 (EP)    Dipeptide ABC transporter, permease protein
MM00006 COG0444 (EP)    Dipeptide ABC transporter, ATP-binding protein
MM00007 COG1124 (EP)    Dipeptide ABC transporter, ATP-binding protein
MM00012 COG5507 (S)     RNA signal recognition particle 4.5S RNA
MM00014 COG0020 (I)     Undecaprenyl pyrophosphate synthetase
MM00027 NOCOG           conserved protein
MM00028 COG0210 (L)     Superfamily I DNA and RNA helicase
MM00033 NOCOG           conserved protein
MM00040 COG4175 (E)     Glycine betaine transporter, ATP-binding protein
MM00041 COG4176 (E)     Glycine betaine transporter, permease protein
MM00042 COG2113 (E)     Glycine betaine transporter, substrate-binding protein
MM00045 COG0666 (R)     hypothetical protein
MM00048 COG1012 (C)     Aldehyde dehydrogenase
MM00052 COG3547 (L)     Transposase
MM00067 COG0123 (BQ)    hypothetical protein
MM00068 COG0145 (EQ)    N-methylhydantoinase (ATP-hydrolyzing)
MM00069 COG0385 (R)     Sodium-dependent transporter
MM00075 COG3794 (C)     conserved protein
MM00084 COG0703 (E)     Shikimate kinase
MM00096 COG3547 (L)     Transposase
MM00106 COG3030 (R)     FxsA protein
MM00109 COG1104 (E)     Cysteine desulfhydrase
MM00115 COG0534 (V)     putative transmembrane efflux protein
MM00121 COG1151 (C)     Carbon monoxide dehydrogenase
MM00139 COG1138 (O)     Cytochrome c-type biogenesis protein
MM00148 NOCOG           conserved protein
MM00150 COG0702 (MG)    putative nucleoside-diphosphate-sugar epimerase
MM00153 NOCOG           conserved protein
MM00154 NOCOG           conserved protein
MM00156 COG1002 (V)     type IIS restriction enzyme
```

MM00159	NOCOG	hypothetical protein
MM00163	NOCOG	DNA repair protein
MM00164	NOCOG	ATP-dependent protease La
MM00166	COG5310 (Q)	homospermidine synthase
MM00167	COG0053 (P)	Cobalt-zinc-cadmium resistance protein
MM00168	COG2202 (T)	hypothetical sensory transduction histidine kinase
MM00170	NOCOG	Transposase
MM00176	COG0407 (H)	Methylcobalamin:coenzyme M methyltransferase
MM00180	COG0454 (KR)	Diamine acetyltransferase
MM00181	COG4866 (S)	conserved protein
MM00191	NOCOG	Transposase
MM00196	COG0433 (R)	conserved protein
MM00197	NOCOG	conserved protein
MM00203	COG4804 (S)	hypothetical protein
MM00204	COG0454 (KR)	hypothetical protein
MM00206	COG2865 (K)	conserved protein
MM00209	COG1793 (L)	hypothetical protein
MM00211	COG4870 (O)	Cysteine proteinase
MM00218	NOCOG	conserved protein
MM00228	NOCOG	hypothetical protein
MM00232	COG3324 (R)	hypothetical protein
MM00238	COG2013 (S)	hypothetical protein
MM00249	COG0655 (R)	BRAMP
MM00250	COG3436 (L)	Transposase
MM00269	NOCOG	hypothetical protein
MM00274	COG1502 (I)	putative endonuclease
MM00280	NOCOG	hypothetical protein
MM00281	NOCOG	hypothetical protein
MM00300	COG0449 (M)	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]
MM00309	COG0535 (R)	metallo cofactor biosynthesis protein
MM00310	COG3173 (R)	hypothetical protein
MM00360	COG1251 (C)	Sulfite reductase, assimilatory-type
MM00362	COG1251 (C)	Sulfite reductase, assimilatory-type
MM00375	NOCOG	conserved protein
MM00383	COG5658 (S)	hypothetical protein
MM00396	COG3436 (L)	Transposase
MM00410	NOCOG	hypothetical protein
MM00421	COG0675 (L)	Transposase
MM00429	COG0286 (V)	type I restriction-modification system methylation subunit
MM00430	COG0732 (V)	type I restriction-modification system specificity subunit
MM00431	COG0610 (V)	type I restriction-modification system restriction subnit
MM00438	COG0685 (E)	Methylenetetrahydrofolate reductase
MM00439	NOCOG	zinc-finger protein
MM00441	COG0190 (H)	Methylenetetrahydrofolate dehydrogenase (NADP+)
MM00442	COG0112 (E)	Serine hydroxymethyltransferase
MM00443	COG0299 (F)	Phosphoribosylglycinamide formyltransferase
MM00449	COG2230 (M)	SAM-dependent methyltransferases
MM00451	NOCOG	hypothetical protein
MM00455	COG1305 (E)	hypothetical protein
MM00481	COG0454 (KR)	Acetyltransferase
MM00490	COG1413 (C)	Phycocyanin alpha-subunit phycocyanobilin lyase
MM00495	COG0282 (C)	Acetate kinase
MM00496	COG0280 (C)	Phosphate acetyltransferase
MM00500	COG3832 (S)	hypothetical protein
MM00501	COG3865 (S)	3-demethylubiquinone-9 3-methyltransferase
MM00502	COG0251 (J)	Translation initiation inhibitor
MM00504	NOCOG	Phycocyanin alpha-subunit phycocyanobilin lyase
MM00505	COG0407 (H)	Methylcobalamin:Coenzyme M methyltransferase MtbA
MM00509	NOCOG	Transposase
MM00518	COG2202 (T)	hypothetical sensory transduction histidine kinase
MM00519	NOCOG	hypothetical protein
MM00522	COG3547 (L)	Transposase
MM00525	NOCOG	hypothetical protein
MM00526	NOCOG	hypothetical protein
MM00534	COG1835 (I)	conserved protein
MM00537	COG0662 (G)	hypothetical protein
MM00540	COG1408 (R)	conserved protein
MM00561	COG1203 (R)	conserved protein
MM00563	NOCOG	hypothetical protein
MM00566	COG4912 (L)	hypothetical protein
MM00570	COG1119 (P)	ABC transporter, ATP-binding protein
MM00572	COG1785 (P)	Alkaline phosphatase
MM00612	COG0171 (H)	NH(3)-dependent NAD(+) synthetase
MM00616	NOCOG	conserved protein
MM00619	COG1331 (O)	conserved protein
MM00621	NOCOG	hypothetical protein
MM00633	NOCOG	hypothetical protein
MM00639	COG3527 (Q)	Alpha-acetolactate decarboxylase
MM00653	COG0463 (M)	glycosyltransferase involved in cell wall biogenesis
MM00654	COG0438 (M)	mannosyltransferase
MM00656	COG1134 (GM)	Polysaccharide ABC transporter, ATP-binding protein
MM00657	COG1682 (GM)	Polysaccharide ABC transporter, permease protein
MM00658	COG0451 (MG)	GDP-fucose synthetase
MM00659	COG1089 (M)	GDP-mannose 4,6 dehydratase
MM00664	COG0493 (ER)	glutamate synthase [NADPH]
MM00701	COG3808 (C)	vacuolar-type H+-pyrophosphatase
MM00722	COG2710 (C)	Nitrogenase molybdenum-iron protein alpha chain
MM00723	COG2710 (C)	Nitrogenase molybdenum-iron protein beta chain
MM00724	COG2710 (C)	Nitrogenase iron-molybdenum cofactor biosynthesis protein
MM00726	COG0725 (P)	Molybdate-binding protein
MM00729	COG0834 (ET)	conserved protein
MM00735	COG1742 (S)	conserved protein

MM00736	COG0701 (R)	Transporter
MM00739	COG3436 (L)	Transposase
MM00749	COG1063 (ER)	Glutathione-independent formaldehyde dehydrogenase
MM00750	COG0145 (EQ)	Dihydropyrimidinase
MM00768	NOCOG	glycosyltransferase
MM00773	COG4572 (R)	putative cation transport regulator
MM00788	COG0535 (R)	metallo cofactor biosynthesis protein
MM00801	COG0104 (F)	Adenylosuccinate synthetase
MM00816	COG3436 (L)	Transposase
MM00817	COG0345 (E)	Pyrroline-5-carboxylate reductase
MM00818	COG0263 (E)	glutamate 5-kinase
MM00819	COG0014 (E)	Gamma-glutamyl phosphate reductase
MM00835	COG0474 (P)	Cation-transporting ATPase
MM00844	NOCOG	hypothetical protein
MM00846	NOCOG	conserved protein
MM00864	COG3387 (G)	Glucoamylase
MM00865	COG0172 (J)	Seryl-tRNA synthetase
MM00877	NOCOG	Transposase
MM00898	COG0138 (F)	formyltransferasephosphoribosylaminoimidazolecarboxamide
MM00901	COG0474 (P)	Cation-transporting ATPase
MM00904	COG0696 (G)	Phosphoglycerate mutase
MM00926	NOCOG	Transposase
MM00930	COG0347 (E)	Nitrogen regulatory protein P-II
MM00931	NOCOG	hypothetical protein
MM00932	NOCOG	hypothetical protein
MM00934	COG1509 (E)	lysine 2,3-aminomutase
MM00935	NOCOG	Beta-lysine acetyltransferase
MM00948	COG0642 (T)	Sensory transduction histidine kinases
MM00971	NOCOG	hypothetical protein
MM00974	COG0006 (E)	Xaa-Pro aminopeptidase
MM00992	COG0785 (O)	Cytochrome c-type biogenesis protein
MM01002	COG0119 (E)	2-isopropylmalate synthase
MM01016	COG0454 (KR)	putative ribosomal-protein-alanine acetyltransferase
MM01019	COG1600 (C)	Iron-sulfur cluster-binding protein
MM01020	COG0350 (L)	O6-methylguanine-DNA methyltransferase
MM01035	COG1246 (E)	Amino-acid acetyltransferase
MM01039	COG0619 (P)	Cobalt transport protein
MM01043	NOCOG	Transposase
MM01044	NOCOG	Transposase
MM01057	COG4152 (R)	ABC transporter, ATP-binding protein
MM01059	COG0651 (CP)	Formate hydrogenlyase subunit 3
MM01061	NOCOG	Hydrogenase-4 component E
MM01062	COG0651 (CP)	Hydrogenase-4 component F
MM01064	COG3260 (C)	Formate hydrogenlyase subunit 7
MM01070	COG0407 (H)	Methylcobalamin: Coenzyme M methyltransferase
MM01072	COG0523 (R)	conserved protein
MM01077	COG0145 (EQ)	conserved protein
MM01078	COG0145 (EQ)	conserved protein
MM01090	COG0467 (T)	putative DNA integration/recombination/invertion protein
MM01091	COG0526 (OC)	thiol-disulfide isomerase
MM01115	NOCOG	conserved protein
MM01124	COG2244 (R)	O-antigen translocase
MM01128	NOCOG	conserved protein
MM01139	COG0463 (M)	Dolichyl-phosphate mannose synthase related protein
MM01155	COG0726 (G)	conserved protein
MM01156	COG2244 (R)	O-antigen translocase
MM01171	COG1835 (I)	conserved protein
MM01173	NOCOG	conserved protein
MM01180	COG1063 (ER)	oxidoreductase
MM01182	COG0438 (M)	glycosyltransferase
MM01192	NOCOG	Transposase
MM01217	COG0514 (L)	ATP-dependent DNA helicase
MM01234	COG0262 (H)	conserved protein
MM01235	COG1715 (V)	5-methylcytosine-specific restriction enzyme
MM01260	COG3436 (L)	Transposase
MM01264	NOCOG	hypothetical protein
MM01274	COG0169 (E)	Shikimate 5-dehydrogenase
MM01283	COG2193 (P)	Bacterioferritin
MM01300	NOCOG	Transposase
MM01306	COG2013 (S)	HTH DNA-binding protein
MM01308	COG3666 (L)	Transposase
MM01316	COG0591 (ER)	sodium/proline symporter
MM01331	COG2610 (GE)	Gluconate permease
MM01337	NOCOG	hypothetical protein
MM01343	COG0042 (J)	NifR3-like protein
MM01375	COG0855 (P)	Polyphosphate kinase
MM01376	COG0248 (FP)	Exopolyphosphatase
MM01377	COG5607 (S)	hypothetical protein
MM01395	COG0346 (E)	conserved protein
MM01435	COG0531 (E)	Monomethylamine permease
MM01439	COG0407 (H)	Methylcobalamin:coenzyme M methyltransferase
MM01442	NOCOG	hypothetical protein
MM01443	COG0458 (EF)	hypothetical protein
MM01444	COG0502 (H)	biotin synthetase
MM01445	NOCOG	hypothetical protein
MM01461	COG0474 (P)	Cation-transporting ATPase
MM01467	NOCOG	hypothetical protein
MM01469	COG3315 (Q)	conserved protein
MM01479	COG0553 (KL)	SWF/SNF family helicase
MM01480	COG4279 (S)	conserved protein
MM01582	COG0655 (R)	iron-sulfur flavoprotein
MM01598	NOCOG	conserved protein

```
MM01607 COG0454 (KR)    Acetyltransferase
MM01615 COG1453 (R)     oxidoreductase, ALDO/KETO reductase family
MM01624 COG3436 (L)     Transposase
MM01640 COG0726 (G)     putative polysaccharide deacetylase
MM01641 COG0438 (M)     glycosyltransferase
MM01651 COG3663 (L)     conserved protein
MM01656 COG0665 (E)     oxidoreductase
MM01657 COG0639 (T)     putative serine/threonine protein phosphatase
MM01660 COG2239 (P)     MG2+ transporter MGTE
MM01661 COG4096 (V)     type I restriction-modification system restriction subunit
MM01666 NOCOG           ATPase involved in DNA repair
MM01668 COG2865 (K)     ATP-dependent DNA helicase RecG
MM01669 COG0286 (V)     type I restriction-modification system specificity subunit
MM01675 COG0325 (R)     conserved protein
MM01679 COG3547 (L)     Transposase
MM01683 COG0249 (L)     DNA mismatch repair protein
MM01684 COG3594 (G)     O-acetyl transferase
MM01688 NOCOG           Trimethylamine:corrinoid methyltransferase
MM01689 NOCOG           Trimethylamine:corrinoid methyltransferase MttB
MM01695 COG3212 (S)     conserved protein
MM01700 COG0826 (O)     protease
MM01701 COG0702 (MG)    putative nucleoside-diphosphate-sugar epimerase
MM01713 NOCOG           conserved protein
MM01722 NOCOG           hypothetical protein
MM01730 COG0675 (L)     Transposase
MM01736 COG3436 (L)     Transposase
MM01770 COG0574 (G)     Pyruvate, phosphate dikinase
MM01778 COG0053 (P)     Cobalt-zinc-cadmium resistance protein
MM01779 NOCOG           hypothetical protein
MM01780 COG1926 (R)     hypothetical protein
MM01782 COG0786 (E)     Sodium/glutamate symport carrier protein
MM01785 NOCOG           hypothetical protein
MM01787 COG3436 (L)     Transposase
MM01788 COG2013 (S)     HTH DNA-binding protein
MM01793 NOCOG           hypothetical protein
MM01794 NOCOG           hypothetical protein
MM01797 COG0234 (O)     10 kDa chaperonin
MM01798 COG0459 (O)     60 kDa chaperonin
MM01812 COG1262 (S)     conserved protein
MM01832 COG3436 (L)     Transposase
MM01842 COG1251 (C)     Sulfite reductase, assimilatory-type
MM01847 COG0644 (C)     hypothetical protein
MM01885 COG0019 (E)     Diaminopimelate decarboxylase
MM01886 NOCOG           Nonspecific acid phosphatase precursor
MM01888 NOCOG           Transposase
MM01889 COG3247 (S)     conserved protein
MM01892 COG3119 (P)     Arylsulfatase
MM01893 COG0641 (R)     putative transcriptional regulator
MM01894 COG1796 (L)     DNA polymerase beta
MM01898 COG1226 (P)     Potassium channel protein
MM01902 COG1247 (M)     hypothetical protein
MM01903 COG1846 (K)     transcriptional regulator, MarR family
MM01907 COG1335 (Q)     Isochorismatase
MM01912 NOCOG           hypothetical protein
MM01916 COG1190 (J)     Lysyl-tRNA synthetase
MM01930 COG3547 (L)     Transposase
MM01932 COG0407 (H)     Methylcobalamin:coenzyme M methyltransferase
MM01935 COG0454 (KR)    Acetyltransferase
MM01936 NOCOG           conserved protein
MM01937 COG1739 (S)     hypothetical protein
MM01944 COG1533 (L)     Transporter
MM01948 COG1453 (R)     oxidoreductase, ALDO/KETO reductase family
MM01950 COG0753 (P)     Catalase
MM01953 COG0053 (P)     Cobalt-zinc-cadmium resistance protein
MM01982 COG2015 (Q)     Alkyl sulfatase
MM01985 NOCOG           hypothetical cytosolic protein
MM01987 NOCOG           hypothetical protein
MM02016 COG0316 (S)     HesB protein
MM02018 COG3408 (G)     conserved protein
MM02047 COG5012 (R)     dimethylamine corrinoid protein
MM02048 NOCOG           Trimethylamine:corrinoid methyltransferase
MM02049 NOCOG           Trimethylamine:corrinoid methyltransferase
MM02052 COG5012 (R)     dimethylamine corrinoid protein
MM02076 COG1812 (E)     S-adenosylmethionine synthetase
MM02080 COG0390 (R)     ABC transporter, ATP-binding protein
MM02098 NOCOG           conserved protein
MM02099 COG3594 (G)     hypothetical protein
MM02104 NOCOG           conserved protein
MM02112 COG2348 (V)     hypothetical protein
MM02113 COG0726 (G)     hypothetical protein
MM02122 COG2815 (S)     putative serine/threonine protein kinase
MM02173 NOCOG           conserved protein
MM02178 NOCOG           Two component system histidine kinase
MM02180 COG0518 (F)     GMP synthase [glutamine-hydrolyzing]
MM02181 COG1494 (G)     Fructose-1,6-bisphosphatase
MM02191 COG1403 (V)     conserved protein
MM02192 COG3666 (L)     Transposase
MM02198 COG0286 (V)     type I restriction-modification system specificity subunit
MM02202 NOCOG           hypothetical protein
MM02203 COG3943 (R)     hypothetical protein
MM02204 COG0610 (V)     hypothetical protein
MM02219 COG0500 (QR)    hypothetical protein
```

```
MM02221 COG0463 (M)     glycosyltransferase
MM02224 COG0655 (R)     Trp repressor binding protein
MM02230 NOCOG           Transposase
MM02231 COG3293 (L)     Transposase
MM02234 NOCOG           conserved protein
MM02238 NOCOG           hypothetical protein
MM02253 COG0145 (EQ)    Dihydropyrimidinase
MM02287 COG3619 (S)     hypothetical protein
MM02290 COG0860 (M)     Cell surface protein
MM02291 COG1451 (R)     Zinc metalloprotease
MM02292 COG0610 (V)     type I restriction-modification system restriction subunit
MM02294 COG0286 (V)     type I restriction-modification system specificity subunit
MM02296 NOCOG           conserved protein
MM02302 COG3875 (S)     conserved protein
MM02303 COG3875 (S)     transcriptional regulator
MM02320 COG1009 (CP)    Ech Hydrogenase, Subunit
MM02330 NOCOG           hypothetical protein
MM02333 COG2076 (P)     Multidrug efflux pump
MM02340 COG0028 (EH)    Acetolactate synthase
MM02341 COG1012 (C)     Succinate-semialdehyde dehydrogenase [NADP+]
MM02343 COG3794 (C)     Copper-binding protein
MM02345 COG3666 (L)     Transposase
MM02347 COG0189 (HJ)    Ribosomal protein S6 modification protein
MM02351 COG0784 (T)     transcriptional regulator
MM02377 COG1714 (S)     hypothetical protein
MM02420 COG0188 (L)     DNA gyrase, subunit A
MM02421 COG0326 (O)     Chaperone protein
MM02427 COG0407 (H)     Methylthiol:coenzyme M methyltransferase
MM02430 COG3436 (L)     Transposase
MM02435 COG2202 (T)     hypothetical sensory transduction histidine kinase
MM02436 COG3436 (L)     Transposase
MM02447 COG3335 (L)     Transposase
MM02448 NOCOG           Transposase
MM02456 COG1123 (R)     Dipeptide/oligopeptide transporter, ATP-binding protein
MM02468 COG0535 (R)     metallo cofactor biosynthesis protein
MM02505 COG0443 (O)     Chaperone protein
MM02515 COG2202 (T)     hypothetical sensory transduction histidine kinase
MM02518 NOCOG           Sensory transduction histidine kinases
MM02523 NOCOG           hypothetical protein
MM02524 COG1842 (KT)    Phage shock protein A
MM02529 NOCOG           hypothetical protein
MM02541 COG0474 (P)     Cation-transporting ATPase
MM02542 COG2816 (L)     phosphohydrolase
MM02547 COG1145 (C)     putative ferredoxin
MM02552 COG1357 (S)     hypothetical protein
MM02553 COG0596 (R)     putative hydrolase
MM02554 COG3153 (R)     Acetyltransferase
MM02557 COG0753 (P)     Catalase
MM02561 COG1162 (R)     Conserved hypothetical protein
MM02579 COG3547 (L)     Transposase
MM02588 COG3436 (L)     Transposase
MM02596 COG0145 (EQ)    hypothetical protein
MM02598 COG0145 (EQ)    hypothetical protein
MM02610 COG4096 (V)     Type I restriction enzyme
MM02613 COG1704 (S)     LemA protein
MM02629 NOCOG           hypothetical sensory transduction histidine kinase
MM02632 COG0281 (C)     NAD-dependent malic enzyme
MM02633 COG0500 (QR)    hypothetical protein
MM02638 COG2335 (M)     hypothetical protein
MM02655 COG1835 (I)     hypothetical protein
MM02662 COG1115 (E)     Na(+)-linked D-alanine glycine permease
MM02676 NOCOG           conserved protein
MM02679 NOCOG           conserved protein
MM02682 COG4584 (L)     Transposase
MM02683 COG3344 (L)     reverse transcriptase
MM02685 COG1484 (L)     Transposase
MM02690 COG3436 (L)     Transposase
MM02694 COG4584 (L)     Transposase
MM02695 COG3344 (L)     reverse transcriptase
MM02698 COG3344 (L)     reverse transcriptase
MM02699 COG4584 (L)     Transposase
MM02700 COG1484 (L)     Transposase
MM02704 COG0286 (V)     type I restriction-modification system specificity subunit
MM02705 COG4096 (V)     type I restriction-modification system restriction subunit
MM02706 COG3937 (S)     conserved protein
MM02721 NOCOG           conserved protein
MM02722 COG0590 (FJ)    hypothetical protein
MM02723 COG0574 (G)     Phosphoenolpyruvate synthase
MM02733 COG0610 (V)     hypothetical protein
MM02736 NOCOG           conserved protein
MM02737 NOCOG           conserved protein
MM02738 NOCOG           conserved protein
MM02749 COG2195 (E)     Aminoacyl-histidine dipeptidase
MM02750 COG2862 (S)     hypothetical protein
MM02751 COG4803 (S)     conserved protein
MM02752 COG0628 (R)     hypothetical protein
MM02754 COG5519 (L)     conserved protein
MM02768 NOCOG           conserved protein
MM02769 COG1454 (C)     Iron-containing alcohol dehydrogenase
MM02770 COG3408 (G)     Glycogen debranching enzyme
MM02771 COG2152 (G)     conserved protein
MM02773 COG2202 (T)     Two component system histidine kinase
```

```
MM02778 COG0053 (P)     Cobalt-zinc-cadmium resistance protein
MM02787 NOCOG           conserved protein
MM02790 COG0655 (R)     iron-sulfur flavoprotein
MM02791 COG2003 (L)     DNA repair protein
MM02797 COG1741 (R)     Pirin
MM02800 NOCOG           conserved protein
MM02808 COG1262 (S)     conserved protein
MM02810 NOCOG           conserved protein
MM02811 COG4127 (S)     conserved protein
MM02815 COG1237 (R)     metal dependent hydrolase
MM02824 COG2141 (C)     glucose-6-phosphate dehydrogenase [coenzyme F420 dependent]
MM02826 COG3408 (G)     conserved protein
MM02827 COG3408 (G)     hypothetical protein
MM02837 COG3547 (L)     Transposase
MM02842 COG3467 (R)     5-nitroimidazole antibiotic resistance protein
MM02844 NOCOG           hypothetical protein
MM02846 COG0531 (E)     Putrescine-ornithine antiporter
MM02869 COG2020 (O)     hypothetical protein
MM02871 COG3326 (S)     hypothetical protein
MM02873 COG2133 (G)     L-sorbosone dehydrogenase
MM02876 COG0389 (L)     DNA polymerase IV
MM02879 COG0457 (R)     tetratricopeptide repeat family protein
MM02892 COG0577 (V)     ABC transporter, permease protein
MM02897 COG1741 (R)     Pirin
MM02901 NOCOG           Transposase
MM02902 NOCOG           Transposase
MM02903 COG3666 (L)     Transposase
MM02904 NOCOG           Transposase
MM02905 NOCOG           Transposase
MM02907 COG1741 (R)     Pirin
MM02910 COG1001 (F)     Adenine deaminase
MM02911 COG1932 (HE)    Phosphoserine aminotransferase
MM02912 COG0145 (EQ)    5-oxoprolinase
MM02916 COG1853 (R)     flavoredoxin
MM02917 COG1453 (R)     oxidoreductase, ALDO/KETO reductase family
MM02918 COG1453 (R)     oxidoreductase, ALDO/KETO reductase family
MM02924 COG1295 (S)     Ribonuclease BN
MM02931 COG0491 (R)     hydrolase
MM02932 NOCOG           conserved protein
MM02939 COG1045 (E)     Serine O-acetyltransferase
MM02943 COG0438 (M)     LipopolysaccharideN-acetylglucosaminyltransferase
MM02944 COG2152 (G)     conserved protein
MM02945 COG0438 (M)     Mannosyltransferase
MM02946 NOCOG           hypothetical protein
MM02952 COG2326 (S)     conserved protein
MM02961 COG5012 (R)     dimethylamine corrinoid protein
MM02964 COG0531 (E)     Dimethylamine permease
MM02976 COG0610 (V)     type I restriction-modification system restriction subunit
MM02977 NOCOG           hypothetical protein
MM02978 COG0732 (V)     type I restriction-modification system specificity subunit
MM02979 COG4804 (S)     hypothetical protein
MM02981 COG0286 (V)     type I restriction-modification system specificity subunit
MM02982 COG3467 (R)     antibiotic resistance protein
MM02984 NOCOG           putative ferredoxin
MM02987 COG4548 (P)     hypothetical protein
MM02988 COG0714 (R)     conserved protein
MM02993 COG0588 (G)     Phosphoglycerate mutase
MM02994 NOCOG           conserved protein
MM02998 NOCOG           Transposase
MM03007 COG0784 (T)     putative transcriptional regulator
MM03027 NOCOG           Transposase
MM03028 COG0590 (FJ)    Deoxycytidylate deaminase
MM03038 COG0675 (L)     Transposase
MM03049 COG0438 (M)     glycosyltransferase
MM03052 COG1063 (ER)    oxidoreductase
MM03054 COG2843 (M)     conserved protein
MM03059 NOCOG           conserved protein
MM03063 COG1404 (O)     calcium dependent protease
MM03069 COG0474 (P)     Cation-transporting ATPase
MM03070 COG2326 (S)     hypothetical protein
MM03073 NOCOG           surface layer protein B
MM03086 COG1488 (H)     Nicotinate phosphoribosyltransferase
MM03090 COG1413 (C)     Phycocyanin alpha-subunit phycocyanobilin lyase related protein, CpcE/NblB family
MM03099 COG0642 (T)     Two component system histidine kinase
MM03101 NOCOG           Two component system histidine kinase
MM03105 COG0590 (FJ)    putative cytosine deaminase
MM03106 COG2335 (M)     hypothetical protein
MM03114 NOCOG           conserved protein
MM03115 COG3177 (S)     hypothetical protein
MM03118 COG0466 (O)     ATP-dependent protease La
MM03119 COG0655 (R)     iron-sulfur flavoprotein
MM03121 COG0477 (GEPR)  hypothetical protein
MM03123 COG1280 (E)     hypothetical protein
MM03130 COG0534 (V)     hypothetical protein
MM03131 COG0524 (G)     Fructokinase
MM03132 NOCOG           hypothetical protein
MM03133 NOCOG           hypothetical protein
MM03139 COG1970 (M)     Large-conductance mechanosensitive channel
MM03141 COG4096 (V)     type I restriction-modification system restriction subunit
MM03147 COG0286 (V)     type I restriction-modification system specificity subunit
MM03151 NOCOG           hypothetical protein
MM03152 NOCOG           hypothetical protein
```

MM03156	COG1453 (R)	oxidoreductase
MM03157	COG1680 (V)	putative 6-aminohexanoate-dimer hydrolase
MM03160	COG3436 (L)	Transposase
MM03163	NOCOG	conserved protein
MM03176	NOCOG	hypothetical protein
MM03185	COG0019 (E)	Ornithine decarboxylase
MM03188	COG0174 (E)	Glutamine synthetase
MM03189	COG1226 (P)	Potassium channel protein
MM03191	NOCOG	hypothetical protein
MM03192	NOCOG	hypothetical membrane spanning protein
MM03194	COG0534 (V)	DinF protein
MM03198	COG1073 (R)	hypothetical protein
MM03200	NOCOG	hypothetical protein
MM03201	COG0518 (F)	conserved protein
MM03202	COG2898 (S)	conserved protein
MM03203	COG4870 (O)	hypothetical protein
MM03204	COG0534 (V)	conserved protein
MM03205	COG0642 (T)	Two component system histidine kinase
MM03207	COG0826 (O)	protease
MM03213	COG1151 (C)	Prismane protein
MM03218	COG0346 (E)	conserved protein
MM03222	NOCOG	conserved protein
MM03228	NOCOG	Transposase
MM03229	NOCOG	Transposase
MM03240	COG0641 (R)	putative transcriptional regulator
MM03244	NOCOG	hypothetical protein
MM03245	COG3436 (L)	Transposase
MM03247	COG1357 (S)	conserved protein
MM03259	COG0668 (M)	conserved protein
MM03260	COG2323 (S)	hypothetical protein
MM03262	COG1296 (E)	Branched-chain amino acid transport protein
MM03263	COG1687 (E)	Branched-chain amino acid transport protein
MM03275	COG1253 (R)	conserved protein
MM03276	COG0778 (C)	hypothetical protein
MM03285	COG1408 (R)	conserved protein
MM03294	COG1787 (V)	hypothetical protein
MM03295	COG2202 (T)	hypothetical sensory transduction histidine kinase
MM03308	COG1073 (R)	conserved protein
MM03309	COG0778 (C)	Ferredoxin
MM03310	COG0693 (R)	transcriptional regulator, AraC family
MM03311	COG3153 (R)	Acetyltransferase
MM03312	COG3183 (V)	5-methylcytosine-specific restriction enzyme A
MM03317	COG4119 (LR)	hypothetical protein
MM03325	COG1301 (C)	permease, Na+/H+-dicarboxylate symporter
MM03328	COG0518 (F)	conserved protein
MM03329	COG0546 (R)	Phosphoglycolate phosphatase
MM03343	NOCOG	Transposase
MM03344	NOCOG	Transposase
MM03347	NOCOG	conserved protein
MM03350	NOCOG	hypothetical protein
MM03371	NOCOG	conserved protein

A.5. Fremdgene in *Ms. mazei* SIGI-Vorhersage

Genname, COG, (COG-Gruppe), Annotation

MM00273	NOCOG	hypothetical protein
MM00274	COG1502 (I)	putative endonuclease
MM00275	NOCOG	hypothetical protein
MM00276	NOCOG	Superfamily II DNA and RNA helicase
MM00277	COG1194 (L)	T/G-specific DNA glycosylase
MM00390	COG3391 (S)	putative surface layer protein
MM00391	NOCOG	hypothetical protein
MM00393	NOCOG	conserved protein
MM00394	NOCOG	conserved protein
MM00395	COG3391 (S)	putative surface layer protein
MM00396	COG3436 (L)	Transposase
MM00397	NOCOG	Transposase
MM00398	NOCOG	conserved protein
MM00400	NOCOG	conserved protein
MM00401	NOCOG	conserved protein
MM00402	NOCOG	conserved protein
MM00404	NOCOG	hypothetical protein
MM00405	COG3391 (S)	putative surface layer protein
MM00406	NOCOG	conserved protein
MM00449	COG2230 (M)	SAM-dependent methyltransferases
MM00450	COG2320 (S)	conserved protein
MM00451	NOCOG	hypothetical protein

MM00452	NOCOG	hypothetical protein
MM00453	NOCOG	conserved protein
MM00454	NOCOG	hypothetical protein
MM00539	NOCOG	hypothetical protein
MM00542	NOCOG	hypothetical protein
MM00544	COG0557 (K)	Exoribonuclease II
MM00545	NOCOG	hypothetical protein
MM00546	NOCOG	conserved protein
MM00547	NOCOG	conserved protein
MM00548	NOCOG	hypothetical protein
MM00549	COG3654 (R)	conserved protein
MM00550	NOCOG	hypothetical protein
MM00552	NOCOG	hypothetical protein
MM00553	NOCOG	hypothetical protein
MM00554	NOCOG	conserved protein
MM00555	NOCOG	hypothetical protein
MM00556	NOCOG	hypothetical protein
MM00560	NOCOG	hypothetical protein
MM00561	COG1203 (R)	conserved protein
MM00562	NOCOG	hypothetical protein
MM00563	NOCOG	hypothetical protein
MM00564	NOCOG	hypothetical protein
MM00650	COG0438 (M)	mannosyltransferase
MM00651	COG1216 (R)	glycosyltransferase
MM00652	COG0463 (M)	glycosyltransferase
MM00653	COG0463 (M)	glycosyltransferase involved in cell wall biogenesis
MM00654	COG0438 (M)	mannosyltransferase
MM00655	NOCOG	hypothetical protein
MM00656	COG1134 (GM)	Polysaccharide ABC transporter, ATP-binding protein
MM00658	COG0451 (MG)	GDP-fucose synthetase
MM00874	NOCOG	conserved protein
MM00875	NOCOG	hypothetical protein
MM00876	NOCOG	hypothetical protein
MM00877	NOCOG	Transposase
MM00878	NOCOG	Transposase
MM00879	NOCOG	conserved protein
MM00881	NOCOG	Transposase
MM00882	NOCOG	conserved protein
MM01171	COG1835 (I)	conserved protein
MM01173	NOCOG	conserved protein
MM01174	NOCOG	hypothetical protein
MM01175	COG2244 (R)	oligosaccharide repeat unit transporter
MM01176	NOCOG	Conserved transmembrane protein
MM01177	COG0438 (M)	glycosyltransferase
MM01178	COG0438 (M)	glycosyltransferase
MM01179	COG0438 (M)	glycosyltransferase
MM01427	NOCOG	conserved protein
MM01428	NOCOG	conserved protein
MM01429	NOCOG	Transposase
MM01430	COG3547 (L)	Transposase
MM01431	COG3547 (L)	Transposase
MM01432	NOCOG	hypothetical protein
MM01637	NOCOG	conserved protein
MM01638	COG0463 (M)	glycosyltransferase
MM01639	COG0110 (R)	putative acetyltransferase
MM01640	COG0726 (G)	putative polysaccharide deacetylase
MM01641	COG0438 (M)	glycosyltransferase
MM01642	NOCOG	Conserved transmembrane protein
MM01643	COG2244 (R)	oligosaccharide repeat unit transporter
MM01644	COG0500 (QR)	methyltransferase
MM01645	NOCOG	hypothetical protein
MM01663	NOCOG	conserved protein
MM01664	NOCOG	conserved protein
MM01665	NOCOG	conserved protein
MM01666	NOCOG	ATPase involved in DNA repair
MM01667	COG0732 (V)	type I restriction-modification system specificity subunit
MM01670	NOCOG	conserved protein
MM01675	COG0325 (R)	conserved protein
MM01676	NOCOG	conserved protein
MM01677	COG3291 (R)	conserved protein
MM01679	COG3547 (L)	Transposase
MM02024	NOCOG	conserved protein
MM02025	NOCOG	hypothetical protein
MM02026	NOCOG	hypothetical protein
MM02027	NOCOG	hypothetical protein
MM02028	COG1413 (C)	hypothetical protein
MM02029	NOCOG	conserved protein
MM02099	COG3594 (G)	hypothetical protein
MM02100	NOCOG	conserved protein
MM02101	COG2244 (R)	Heteropolysaccharide repeat unit export protein
MM02102	COG0463 (M)	glycosyltransferase
MM02103	COG0500 (QR)	putative methyltransferase
MM02104	NOCOG	conserved protein
MM02105	COG0438 (M)	putative glycosyltransferase
MM02106	COG0438 (M)	putative glycosyltransferase
MM02107	COG0438 (M)	glycosyltransferase
MM02108	NOCOG	conserved protein
MM02109	COG0463 (M)	glycosyltransferase
MM02111	COG0110 (R)	galactoside-O-acetyltransferase
MM02198	COG0286 (V)	type I restriction-modification system specificity subunit
MM02199	COG0732 (V)	type I restriction-modification system specificity subunit
MM02200	NOCOG	hypothetical protein

MM02201	NOCOG	hypothetical protein
MM02202	NOCOG	hypothetical protein
MM02211	NOCOG	hypothetical protein
MM02212	NOCOG	conserved protein
MM02213	COG1216 (R)	putative glycosyltransferase
MM02214	COG1134 (GM)	Polysaccharide ABC transporter, ATP-binding protein
MM02215	COG1682 (GM)	Polysaccharide ABC transporter, permease protein
MM02217	COG0463 (M)	conserved protein
MM02406	NOCOG	conserved protein
MM02407	NOCOG	conserved protein
MM02408	NOCOG	conserved protein
MM02409	NOCOG	conserved protein
MM02410	NOCOG	conserved protein
MM02411	NOCOG	hypothetical protein
MM02412	NOCOG	conserved protein
MM02413	NOCOG	conserved protein
MM02414	NOCOG	hypothetical protein
MM02415	NOCOG	hypothetical protein
MM02430	COG3436 (L)	Transposase
MM02434	NOCOG	conserved protein
MM02435	COG2202 (T)	hypothetical sensory transduction histidine kinase
MM02436	COG3436 (L)	Transposase
MM02437	NOCOG	Transposase
MM02439	COG1943 (L)	Transposase
MM02441	COG3398 (S)	transcriptional regulator, ArsR family
MM02442	NOCOG	conserved protein
MM02443	COG2834 (M)	conserved protein
MM02444	NOCOG	conserved protein
MM02445	COG0785 (O)	conserved protein
MM02446	COG3398 (S)	transcriptional regulator, ArsR family
MM02447	COG3335 (L)	Transposase
MM02449	NOCOG	conserved protein
MM02450	COG3328 (L)	Transposase
MM02451	NOCOG	conserved protein
MM02452	COG0785 (O)	conserved protein
MM02453	COG2881 (S)	conserved protein
MM02543	NOCOG	hypothetical protein
MM02544	NOCOG	hypothetical protein
MM02545	NOCOG	hypothetical protein
MM02546	COG1309 (K)	Conserved hypothetical protein
MM02547	COG1145 (C)	putative ferredoxin
MM02548	NOCOG	hypothetical protein
MM02549	NOCOG	Conserved hypothetical protein
MM02550	NOCOG	hypothetical protein
MM02551	COG4845 (V)	Acetyltransferase
MM02552	COG1357 (S)	hypothetical protein
MM02554	COG3153 (R)	Acetyltransferase
MM02555	NOCOG	hypothetical protein
MM02556	COG0513 (LKJ)	ATP-dependent RNA helicase
MM02558	NOCOG	hypothetical protein
MM02559	COG1479 (S)	hypothetical protein
MM02560	NOCOG	hypothetical protein
MM02677	NOCOG	conserved protein
MM02678	COG1373 (R)	hypothetical protein
MM02680	NOCOG	conserved protein
MM02681	NOCOG	hypothetical protein
MM02682	COG4584 (L)	Transposase
MM02683	COG3344 (L)	reverse transcriptase
MM02685	COG1484 (L)	Transposase
MM02686	NOCOG	Transposase
MM02687	NOCOG	hypothetical protein
MM02688	COG1032 (C)	hypothetical protein
MM02689	NOCOG	Transposase
MM02690	COG3436 (L)	Transposase
MM02691	COG3576 (R)	hypothetical protein
MM02692	COG1233 (Q)	hypothetical protein
MM02693	COG1244 (R)	hypothetical protein
MM02694	COG4584 (L)	Transposase
MM02695	COG3344 (L)	reverse transcriptase
MM02696	COG1662 (L)	hypothetical protein
MM02697	COG3677 (L)	hypothetical protein
MM02698	COG3344 (L)	reverse transcriptase
MM02701	NOCOG	hypothetical protein
MM02702	COG3593 (L)	hypothetical protein
MM02703	COG0732 (V)	type I restriction-modification system specificity subunit
MM02739	COG0732 (V)	type I restriction-modification system specificity subunit
MM02741	NOCOG	hypothetical protein
MM02742	NOCOG	hypothetical protein
MM02743	COG0494 (LR)	hypothetical protein
MM02753	COG0582 (L)	hypothetical protein
MM02754	COG5519 (L)	conserved protein
MM02755	NOCOG	hypothetical protein
MM02756	NOCOG	hypothetical protein
MM02757	NOCOG	hypothetical protein
MM02758	NOCOG	hypothetical protein
MM02759	NOCOG	hypothetical protein
MM02760	NOCOG	hypothetical protein
MM02761	NOCOG	hypothetical protein
MM02762	NOCOG	hypothetical protein
MM02763	COG1961 (L)	putative DNA recombinase
MM02764	NOCOG	hypothetical protein
MM02765	NOCOG	hypothetical protein

MM02766	NOCOG	hypothetical protein
MM02767	NOCOG	hypothetical protein
MM02768	NOCOG	conserved protein
MM02774	NOCOG	conserved protein
MM02775	COG2339 (S)	conserved protein
MM02776	NOCOG	hypothetical protein
MM02898	NOCOG	conserved protein
MM02899	COG2985 (R)	conserved protein
MM02900	NOCOG	conserved protein
MM02901	NOCOG	Transposase
MM02902	NOCOG	Transposase
MM02903	COG3666 (L)	Transposase
MM02904	NOCOG	Transposase
MM02905	NOCOG	Transposase
MM02906	COG1943 (L)	Transposase
MM02908	NOCOG	hypothetical protein
MM02995	NOCOG	hypothetical protein
MM02996	NOCOG	hypothetical protein
MM02997	NOCOG	hypothetical protein
MM02998	NOCOG	Transposase
MM02999	COG3335 (L)	Transposase
MM03026	COG3391 (S)	hypothetical protein
MM03028	COG0590 (FJ)	Deoxycytidylate deaminase
MM03029	COG0346 (E)	Lactoylglutathione lyase
MM03030	COG0727 (R)	conserved protein
MM03031	NOCOG	hypothetical protein
MM03032	COG3369 (S)	conserved protein
MM03033	NOCOG	conserved protein
MM03034	NOCOG	conserved protein
MM03036	COG0454 (KR)	hypothetical protein
MM03037	COG1943 (L)	Transposase
MM03038	COG0675 (L)	Transposase
MM03039	NOCOG	hypothetical protein
MM03040	NOCOG	hypothetical protein
MM03041	NOCOG	hypothetical protein
MM03047	COG3385 (L)	Transposase
MM03048	COG2273 (G)	hypothetical protein
MM03049	COG0438 (M)	glycosyltransferase
MM03054	COG2843 (M)	conserved protein
MM03158	COG0326 (O)	hypothetical protein
MM03159	NOCOG	hypothetical protein
MM03160	COG3436 (L)	Transposase
MM03161	NOCOG	Transposase
MM03162	NOCOG	hypothetical protein
MM03165	NOCOG	hypothetical protein
MM03166	NOCOG	hypothetical protein
MM03167	NOCOG	Transposase
MM03168	COG3291 (R)	conserved protein
MM03169	COG3291 (R)	hypothetical protein
MM03171	NOCOG	hypothetical protein
MM03229	NOCOG	Transposase
MM03237	COG3335 (L)	Transposase
MM03242	NOCOG	hypothetical protein
MM03244	NOCOG	hypothetical protein
MM03245	COG3436 (L)	Transposase
MM03246	NOCOG	Transposase
MM03247	COG1357 (S)	conserved protein
MM03248	NOCOG	hypothetical protein
MM03249	COG1518 (L)	hypothetical protein
MM03250	NOCOG	hypothetical protein
MM03251	NOCOG	hypothetical protein
MM03252	NOCOG	hypothetical protein
MM03254	NOCOG	hypothetical protein
MM03255	NOCOG	hypothetical protein
MM03343	NOCOG	Transposase
MM03344	NOCOG	Transposase
MM03345	NOCOG	hypothetical protein
MM03347	NOCOG	conserved protein
MM03348	NOCOG	hypothetical protein
MM03351	NOCOG	hypothetical protein
MM03352	NOCOG	hypothetical protein
MM03353	NOCOG	conserved protein
MM03354	NOCOG	hypothetical protein
MM03355	COG1336 (L)	conserved protein
MM03356	NOCOG	hypothetical protein
MM03357	NOCOG	hypothetical protein
MM03358	NOCOG	hypothetical protein
MM03360	NOCOG	hypothetical protein
MM03361	NOCOG	hypothetical protein
MM03362	NOCOG	hypothetical protein
MM03363	NOCOG	hypothetical protein
MM03364	NOCOG	hypothetical protein
MM03365	NOCOG	hypothetical protein
MM03366	NOCOG	hypothetical protein
MM03367	NOCOG	hypothetical protein
MM03368	NOCOG	hypothetical protein
MM03369	NOCOG	hypothetical protein
MM03370	NOCOG	hypothetical protein

A.6. *T. thermophilus* Proteine ohne Homolog in *D. radiodurans*

Genname, COG, (COG-Gruppe), Annotation

```
TTC0001 NOCOG          hypothetical protein
TTC0007 COG0655 (R)    hypothetical conserved protein
TTC0008 COG1214 (O)    O-sialoglycoprotein endopeptidase
TTC0012 COG2414 (C)    probable tungsten-containing
                       aldehyde ferredoxin oxidoreductase
TTC0013 COG1721 (R)    hypothetical protein
TTC0015 COG0526 (OC)   hypothetical protein
TTC0022 COG1145 (C)    hypothetical protein
TTC0027 NOCOG          hypothetical membrane spanning protein
TTC0030 NOCOG          hypothetical membrane spanning protein
TTC0031 NOCOG          hypothetical protein
TTC0033 COG3255 (I)    hypothetical conserved protein
TTC0034 NOCOG          hypothetical protein
TTC0037 NOCOG          hypothetical protein
TTC0041 COG0863 (L)    methyltransferase
TTC0042 NOCOG          hypothetical protein
TTC0053 COG0457 (R)    hypothetical conserved protein
TTC0055 NOCOG          hypothetical protein
TTC0056 COG0611 (H)    thiamin-monophosphate kinase
TTC0058 NOCOG          hypothetical protein
TTC0059 NOCOG          hypothetical protein
TTC0065 NOCOG          hypothetical protein
TTC0066 NOCOG          hypothetical protein
TTC0069 NOCOG          hypothetical protein
TTC0077 COG0594 (J)    ribonuclease P protein subunit
TTC0082 COG0683 (E)    leucine-, isoleucine-, valine-,
                       threonine-, and alanine-binding protein
TTC0090 NOCOG          hypothetical protein
TTC0091 COG1051 (F)    hypothetical conserved protein
TTC0093 NOCOG          probable transmembrane protein
TTC0094 COG3557 (J)    probable cysteinyl-tRNA synthetase
TTC0095 NOCOG          hypothetical protein
TTC0105 COG2104 (H)    hypothetical protein
TTC0108 COG0500 (QR)    methyltransferase
TTC0113 COG3744 (S)    hypothetical conserved protein
TTC0114 COG4118 (D)    hypothetical protein
TTC0116 NOCOG          hypothetical cytosolic protein
TTC0117 COG0498 (E)    threonine synthase
TTC0118 COG0477 (GEPR) hypothetical protein
TTC0119 NOCOG          hypothetical protein
TTC0120 NOCOG          hypothetical protein
TTC0125 COG3742 (S)    hypothetical conserved protein
TTC0126 NOCOG          hypothetical protein
TTC0128 COG0277 (C)    (S)-2-hydroxy-acid oxidase subunit glcE
TTC0136 COG2152 (G)    hypothetical cytosolic protein
TTC0139 COG3329 (R)    sodium bicarbonate cotransporter
TTC0140 NOCOG          hypothetical conserved protein
TTC0141 COG0586 (S)    hypothetical protein
TTC0142 COG0861 (P)    hypothetical membrane spanning protein
TTC0145 NOCOG          hypothetical protein
TTC0146 NOCOG          hypothetical protein
TTC0153 NOCOG          hypothetical membrane spanning protein
TTC0154 COG0437 (C)    nrfC protein
TTC0156 NOCOG          hypothetical protein
TTC0163 COG0697 (GER)  hypothetical membrane spanning protein
TTC0167 COG2236 (R)    purine phosphoribosyltransferase
TTC0171 COG1568 (R)    hypothetical conserved protein
TTC0176 NOCOG          hypothetical protein
TTC0178 COG1189 (J)    hemolysin
TTC0179 NOCOG          hypothetical protein
TTC0180 NOCOG          hypothetical protein
TTC0186 COG2426 (S)    hypothetical conserved protein
TTC0194 COG0005 (F)    purine-nucleoside phosphorylase
TTC0196 COG3599 (D)    hypothetical conserved protein
TTC0197 COG1434 (S)    hypothetical conserved protein
TTC0212 NOCOG          hypothetical protein
TTC0214 COG4475 (S)    transcriptional regulator
TTC0215 NOCOG          hypothetical protein
TTC0223 NOCOG          hypothetical protein
TTC0224 COG0526 (OC)    thiol-disulfide isomerase and thioredoxins
TTC0225 COG4663 (Q)    probable mannitol-binding protein
                       with frameshift
TTC0226 COG0153 (G)    galactokinase
TTC0228 COG3044 (R)    isopentenyl-diphosphate delta-isomerase
TTC0229 COG3263 (P)    hypothetical membrane spanning protein
TTC0242 COG0502 (H)    biotin synthase
TTC0244 COG2519 (J)    tRNA (adenine-N(1)-)-methyltransferase
TTC0248 COG0607 (P)    hypothetical conserved protein
TTC0254 NOCOG          hypothetical conserved protein
```

TTC0265	COG5405 (O)	ATP-dependent protease hslV
TTC0266	COG1708 (R)	hypothetical protein
TTC0267	COG2250 (S)	hypothetical conserved protein
TTC0270	NOCOG	probable nucleotidyltransferase
TTC0272	NOCOG	hypothetical protein
TTC0273	NOCOG	transposase
TTC0274	NOCOG	hypothetical protein
TTC0277	NOCOG	hypothetical protein
TTC0278	NOCOG	hypothetical protein
TTC0279	COG0463 (M)	probable glycosyl transferase
TTC0280	COG0438 (M)	lipopolysaccharideN-acetylglucosaminyltransferase
TTC0281	NOCOG	hypothetical protein
TTC0282	COG0367 (E)	asparagine synthetase [glutamine-hydrolyzing]
TTC0283	NOCOG	hypothetical protein
TTC0288	COG0677 (M)	UDP-N-acetyl-D-mannosamine 6-dehydrogenase
TTC0290	NOCOG	transposase
TTC0292	NOCOG	hypothetical protein
TTC0293	NOCOG	hypothetical protein
TTC0294	NOCOG	hypothetical protein
TTC0299	NOCOG	hypothetical protein
TTC0300	COG0679 (R)	transporter
TTC0302	NOCOG	hypothetical protein
TTC0303	NOCOG	hypothetical protein
TTC0306	COG2141 (C)	putative oxidoreductase
TTC0309	NOCOG	hypothetical membrane spanning protein
TTC0311	COG1648 (H)	uroporphyrin-III C-methyltransferase (EC 2.1.1.107) / precorrin-2 oxidase (EC 1.-.-.-) / ferrochelatase
TTC0312	COG1587 (H)	uroporphyrinogen-III synthase
TTC0314	NOCOG	hypothetical protein
TTC0318	COG0665 (E)	putative oxidoreductase
TTC0323	NOCOG	hypothetical protein
TTC0328	COG1653 (G)	gucose-binding protein
TTC0344	NOCOG	hypothetical protein
TTC0346	NOCOG	hypothetical protein
TTC0361	NOCOG	hypothetical protein
TTC0362	NOCOG	hypothetical protein
TTC0365	COG0725 (P)	molybdate-binding protein
TTC0368	NOCOG	hypothetical protein
TTC0375	COG0599 (S)	hypothetical conserved protein
TTC0380	NOCOG	hypothetical protein
TTC0381	COG1846 (K)	transcriptional regulator, marR family
TTC0384	NOCOG	hypothetical protein
TTC0385	COG0329 (EM)	dihydrodipicolinate synthase
TTC0386	NOCOG	hypothetical protein
TTC0393	NOCOG	hypothetical protein
TTC0395	NOCOG	hypothetical protein
TTC0399	NOCOG	hypothetical protein
TTC0400	NOCOG	hypothetical protein
TTC0401	COG0365 (I)	hypothetical protein
TTC0405	NOCOG	hypothetical protein
TTC0412	COG0235 (G)	L-ribulose-phosphate 4-epimerase
TTC0414	COG4663 (Q)	mannitol-binding protein
TTC0415	COG4665 (Q)	tripartite transporter, small subunit
TTC0419	COG0600 (P)	ABC transporter permease protein
TTC0423	NOCOG	hypothetical protein
TTC0425	NOCOG	hypothetical conserved protein
TTC0429	NOCOG	hypothetical membrane spanning protein
TTC0430	COG0730 (R)	putative permease
TTC0432	COG0312 (R)	pmbA protein
TTC0433	COG0312 (R)	tldD protein
TTC0434	NOCOG	hypothetical cytosolic protein
TTC0442	NOCOG	hypothetical protein
TTC0444	NOCOG	hypothetical conserved protein
TTC0445	NOCOG	hypothetical protein
TTC0453	NOCOG	hypothetical protein
TTC0464	NOCOG	hypothetical protein
TTC0466	NOCOG	hypothetical protein
TTC0468	NOCOG	RNA polymerase sigma-32 factor
TTC0472	COG0421 (E)	spermine synthase
TTC0473	COG1586 (E)	S-adenosylmethionine decarboxylase proenzyme
TTC0476	NOCOG	hypothetical protein
TTC0478	NOCOG	hypothetical protein
TTC0482	COG0648 (L)	endonuclease IV
TTC0486	COG0526 (OC)	glutaredoxin-like protein
TTC0489	COG2719 (S)	stage V sporulation protein R
TTC0490	COG2718 (S)	glycosyl transferase
TTC0491	COG2766 (T)	prkA protein
TTC0492	NOCOG	hypothetical protein
TTC0494	COG0491 (R)	putative Zn-dependent hydrolase (beta-lactamase superfamily)
TTC0496	COG2137 (R)	regulatory protein recX
TTC0497	NOCOG	hypothetical conserved protein
TTC0501	NOCOG	hypothetical protein
TTC0512	COG1610 (S)	hypothetical protein
TTC0514	COG4401 (E)	chorismate mutase
TTC0516	NOCOG	hypothetical membrane spanning protein
TTC0539	COG0589 (T)	hypothetical conserved protein
TTC0541	COG3809 (S)	hypothetical conserved protein
TTC0542	COG1694 (R)	hypothetical cytosolic protein
TTC0544	NOCOG	hypothetical protein

```
TTC0547 COG0388 (R)    hypothetical conserved protein
TTC0548 COG1268 (R)    probable bioY protein
TTC0551 COG0858 (J)    hypothetical conserved protein
TTC0555 COG1355 (R)    hypothetical conserved protein
TTC0558 NOCOG          hypothetical protein
TTC0563 COG3263 (P)    Na(+)/H(+) antiporter
TTC0566 COG0517 (R)    hypothetical protein
TTC0567 COG4636 (S)    hypothetical cytosolic protein
TTC0571 NOCOG          probable transposase
TTC0574 NOCOG          hypothetical conserved protein
TTC0575 NOCOG          hypothetical conserved protein
TTC0576 NOCOG          hypothetical protein
TTC0578 NOCOG          probable transposase
TTC0579 NOCOG          hypothetical conserved protein
TTC0580 NOCOG          hypothetical protein
TTC0582 COG3239 (I)    omega-(3) fatty acid desaturase
TTC0587 NOCOG          hypothetical conserved protein
TTC0588 NOCOG          mannosyl-3-phoshogylcerate synthase
TTC0589 COG3769 (R)    mannosyl-3-phosphoglycerate phosphatase
TTC0591 COG0329 (EM)   dihydrodipicolinate synthase
TTC0614 NOCOG          hypothetical conserved protein
TTC0615 COG1653 (G)    sugar-binding protein
TTC0616 COG1980 (G)    dihydroorotate dehydrogenase
TTC0618 COG2193 (P)    bacterioferritin
TTC0620 COG0379 (H)    quinolinate synthetase A
TTC0621 COG0157 (H)    nicotinate-nucleotide pyrophosphorylase
                       [carboxylating]
TTC0628 NOCOG          hypothetical conserved protein
TTC0632 NOCOG          hypothetical protein
TTC0635 COG2062 (T)    hypothetical conserved protein
TTC0639 NOCOG          hypothetical protein
TTC0645 NOCOG          hypothetical protein
TTC0650 COG0842 (V)    probable transmembrane protein
TTC0651 NOCOG          hypothetical protein
TTC0653 NOCOG          hypothetical protein
TTC0654 NOCOG          hypothetical protein
TTC0655 NOCOG          hypothetical protein
TTC0656 NOCOG          hypothetical conserved protein
TTC0657 NOCOG          hypothetical conserved protein
TTC0658 NOCOG          hypothetical protein
TTC0659 NOCOG          hypothetical protein
TTC0660 NOCOG          hypothetical protein
TTC0661 NOCOG          hypothetical protein
TTC0662 NOCOG          hypothetical protein
TTC0663 NOCOG          hypothetical protein
TTC0664 NOCOG          hypothetical conserved protein
TTC0675 COG2050 (Q)    hypothetical conserved protein
TTC0681 NOCOG          hypothetical protein
TTC0685 NOCOG          hypothetical protein
TTC0686 COG4864 (S)    hypothetical conserved protein
TTC0687 COG1030 (O)    putative protease / transporter
TTC0695 COG2239 (P)    Mg2+ transporter mgtE
TTC0707 COG0182 (J)    translation initiation factor
                       aIF-2B alpha subunit
TTC0709 NOCOG          hypothetical protein
TTC0712 NOCOG          hypothetical protein
TTC0715 NOCOG          hypothetical protein
TTC0726 COG1839 (S)    hypothetical conserved protein
TTC0731 COG1351 (F)    thy1 protein
TTC0741 NOCOG          hypothetical protein
TTC0750 COG0861 (P)    hypothetical membrane spanning protein
TTC0751 COG0500 (QR)   methyltransferfase
TTC0753 NOCOG          hypothetical protein
TTC0762 NOCOG          hypothetical protein
TTC0768 NOCOG          cytochrome c oxidase polypeptide IIA
TTC0769 COG1622 (C)    subunit II of C(O/B)3-type
                       cytochrome c oxidase
TTC0771 NOCOG          hypothetical protein
TTC0782 NOCOG          hypothetical protein
TTC0786 NOCOG          hypothetical protein
TTC0787 NOCOG          hypothetical protein
TTC0790 NOCOG          hypothetical protein
TTC0796 NOCOG          hypothetical protein
TTC0797 COG1752 (R)    conserved hypothetical protein
TTC0798 NOCOG          putative excisionase
TTC0799 NOCOG          hypothetical protein
TTC0802 NOCOG          hypothetical protein
TTC0804 NOCOG          hypothetical protein
TTC0809 COG0168 (P)    trk system potassium uptake protein trkG
TTC0812 NOCOG          hypothetical protein
TTC0814 NOCOG          hypothetical protein
TTC0815 COG3001 (G)    phosphatidylserine decarboxylase
TTC0821 COG1764 (O)    hypothetical protein
TTC0824 COG1502 (I)    putative phosphoslipase
TTC0826 COG1792 (M)    rod shape-determining protein mreC
TTC0827 NOCOG          hypothetical protein
TTC0829 NOCOG          hypothetical protein
TTC0830 NOCOG          hypothetical protein
TTC0832 NOCOG          hypothetical protein
TTC0837 COG4636 (S)    conserved hypothetical protein
TTC0845 NOCOG          protein with frame Shift
TTC0846 NOCOG          protein with a frame Shift
```

TTC0854	NOCOG	prepilin-like protein
TTC0855	NOCOG	prepilin-like protein
TTC0856	NOCOG	prepilin-like protein
TTC0857	NOCOG	comZ
TTC0858	NOCOG	pilA
TTC0859	COG1196 (D)	hypothetical protein
TTC0861	NOCOG	hypothetical protein
TTC0863	NOCOG	hypothetical protein
TTC0868	COG4636 (S)	conserved hypothetical protein
TTC0869	COG4636 (S)	conserved hypothetical protein
TTC0872	NOCOG	cytochrome c-552 precursor
TTC0877	COG0772 (D)	rod shape-determining protein rodA
TTC0880	COG1733 (K)	conserved hypothetical protein
TTC0881	NOCOG	methylmalonyl-CoA mutase
TTC0883	NOCOG	fragment of protein kinase
TTC0889	NOCOG	hypothetical protein
TTC0892	COG0393 (S)	conserved hypothetical protein
TTC0895	COG2114 (T)	protein with frame shift
TTC0903	NOCOG	hypothetical protein
TTC0904	NOCOG	putative esterase
TTC0910	NOCOG	V-type ATP synthase subunit E
TTC0911	COG0636 (C)	V-type sodium ATP synthase subunit K Homolog
TTC0913	NOCOG	V-Type ATPase subunit
TTC0916	NOCOG	hypothetical protein
TTC0921	NOCOG	hypothetical protein
TTC0924	COG5438 (S)	hypothetical membrane spanning protein
TTC0930	NOCOG	hypothetical protein
TTC0935	COG1079 (R)	putative transport system permease ABC transporter protein
TTC0936	COG4603 (R)	protein with frame shift (permease)
TTC0940	NOCOG	putative oxidoreductase
TTC0941	NOCOG	hypothetical protein
TTC0943	COG3369 (S)	hypothetical conserved protein
TTC0950	COG1994 (R)	conserved hypothetical protein
TTC0954	NOCOG	hypothetical protein
TTC0955	COG0456 (R)	hypothetical protein
TTC0963	COG2258 (S)	hypothetical protein
TTC0982	COG1515 (L)	endonuclease V
TTC0992	NOCOG	hypothetical protein
TTC0993	NOCOG	hypothetical protein
TTC0995	COG2928 (S)	transporter
TTC0996	NOCOG	hypothetical protein
TTC0999	COG4127 (S)	hypothetical conserved protein
TTC1004	COG1752 (R)	serine protease
TTC1009	NOCOG	hypothetical protein
TTC1011	NOCOG	hypothetical protein
TTC1030	NOCOG	ribosomal protein
TTC1032	NOCOG	hypothetical protein
TTC1040	NOCOG	hypothetical protein
TTC1045	COG3258 (C)	cytochrome c
TTC1047	COG0446 (R)	sulfide dehydrogenase /flavocytochrome c)
TTC1048	NOCOG	hypothetical protein
TTC1050	NOCOG	putative cytochrome c, sox a-like
TTC1051	NOCOG	hypothetical protein
TTC1053	NOCOG	hypothetical protein
TTC1054	NOCOG	putative cytochrome c precursor
TTC1055	NOCOG	soxZ
TTC1056	NOCOG	soxY protein
TTC1057	COG0526 (OC)	thioredoxin
TTC1058	COG2010 (C)	cytochrome c-552 precursor
TTC1061	NOCOG	hypothetical protein
TTC1066	COG3360 (S)	hypothetical protein
TTC1067	COG5502 (S)	hypothetical protein
TTC1076	NOCOG	hypothetical protein
TTC1078	NOCOG	conserved hypothetical protein
TTC1083	COG0427 (C)	4-hydroxybutyrate coenzyme A transferase
TTC1085	COG1502 (I)	hypothetical membrane associated protein
TTC1086	NOCOG	hypothetical protein
TTC1093	COG1586 (E)	S-adenosylmethionine decarboxylase proenzyme
TTC1094	COG1470 (S)	ABC transporter substrate-binding protein
TTC1096	COG1277 (R)	ABC transporter permease protein
TTC1097	COG1917 (S)	conserved hypothetical protein
TTC1098	COG1926 (R)	putative phosphoribosyl transferase
TTC1103	COG2607 (R)	fragment of ATP-dependent protease subunit
TTC1105	COG0840 (NT)	conserved hypothetical protein
TTC1107	COG0741 (M)	conserved hypothetical protein
TTC1109	COG0517 (R)	conserved hypothetical protein
TTC1110	COG0705 (R)	hypothetical membrane spanning protein
TTC1111	COG0616 (OU)	periplasmic serine protease
TTC1112	COG0475 (P)	hypothetical protein
TTC1114	NOCOG	hypothetical protein
TTC1115	NOCOG	hypothetical protein
TTC1118	NOCOG	bacterioferritin
TTC1119	COG2129 (R)	heat-stable protein
TTC1121	NOCOG	hypothetical protein
TTC1122	NOCOG	heat-stable protein
TTC1124	NOCOG	dnaJ/dnaK assembly factor dafA
TTC1131	COG1714 (S)	hypothetical membrane spanning protein
TTC1141	COG0755 (O)	hypothetical protein
TTC1146	COG2353 (S)	hypothetical protein

```
TTC1147 COG3485 (Q)    conserved hypothetical protein
TTC1150 COG4636 (S)    hypothetical cytosolic protein
TTC1159 NOCOG          hypothetical protein
TTC1162 COG0119 (E)    hydroxymethylglutaryl-CoA lyase
TTC1165 NOCOG          hypothetical protein
TTC1166 COG0727 (R)    hypothetical protein
TTC1169 COG3547 (L)    probable DNA-binding protein
TTC1170 COG1518 (L)    hypothetical protein
TTC1173 COG0665 (E)    fragment of putative oxidoreductase
TTC1174 NOCOG          conserved hypothetical protein
TTC1175 COG0665 (E)    fragment of putative oxidoreductase
TTC1177 NOCOG          hypothetical protein
TTC1178 COG0596 (R)    hypothetical protein
TTC1179 COG0596 (R)    hypothetical protein
TTC1183 NOCOG          hypothetical protein
TTC1189 COG4274 (S)    hypothetical protein
TTC1199 COG0350 (L)    putative methylated-DNA--protein-cysteine
                       methyltransferase
TTC1204 NOCOG          hypothetical protein
TTC1205 COG1899 (O)    deoxyhypusine synthase
TTC1206 NOCOG          hypothetical protein
TTC1207 NOCOG          hypothetical protein
TTC1208 NOCOG          hypothetical protein
TTC1216 NOCOG          transposase
TTC1218 NOCOG          transcriptional regulator
TTC1220 COG0827 (L)    modification methylase
TTC1221 NOCOG          hypothetical protein
TTC1226 NOCOG          hypothetical protein
TTC1227 NOCOG          hypothetical protein
TTC1228 COG0642 (T)    two component system histidine kinase
TTC1232 NOCOG          hypothetical protein
TTC1237 NOCOG          hypothetical protein
TTC1239 NOCOG          hypothetical protein
TTC1240 COG0374 (C)    conserved hypothetical protein
TTC1244 COG1430 (S)    conserved hypothetical protein
TTC1245 COG1430 (S)    hypothetical protein
TTC1251 COG1259 (S)    conserved hypothetical protein
TTC1252 COG1409 (R)    acid phosphatase
TTC1258 COG2897 (P)    thiosulfate sulfurtransferase
TTC1262 COG3374 (S)    conserved hypothetical protein
TTC1264 COG1840 (P)    iron(III)-binding protein
TTC1269 NOCOG          hypothetical protein
TTC1276 NOCOG          hypothetical protein
TTC1289 NOCOG          hypothetical protein
TTC1292 NOCOG          hypothetical protein
TTC1294 NOCOG          hypothetical protein
TTC1295 COG0457 (R)    hypothetical protein
TTC1297 COG0697 (GER)  transporter
TTC1298 NOCOG          hypothetical protein
TTC1310 COG1841 (J)    LSU ribosomal protein L30P
TTC1320 COG0255 (J)    50S ribosomal protein L29
TTC1335 COG2606 (S)    ebsC protein
TTC1337 NOCOG          hypothetical protein
TTC1338 NOCOG          hypothetical protein
TTC1340 NOCOG          hypothetical protein
TTC1341 COG1647 (R)    carboxylesterase
TTC1343 COG1489 (R)    sugar fermentation stimulation protein
TTC1344 COG1835 (I)    hypothetical protein
TTC1346 NOCOG          hypothetical protein
TTC1347 COG1180 (O)    pyruvate formate-lyase activating enzyme
TTC1348 COG1355 (R)    hypothetical conserved protein
TTC1349 NOCOG          hypothetical conserved protein
TTC1350 COG0038 (P)    hypothetical protein
TTC1352 COG3253 (S)    hypothetical cytosolic protein Cluster
                       function: Polyketide synthase curD
TTC1359 COG3462 (S)    hypothetical protein
TTC1362 NOCOG          hypothetical conserved protein
TTC1363 NOCOG          S-layer repressor
TTC1364 NOCOG          hypothetical protein
TTC1365 COG5428 (S)    hypothetical conserved protein
TTC1367 NOCOG          hypothetical protein
TTC1368 NOCOG          hypothetical membrane associated protein
TTC1369 COG3439 (S)    hypothetical conserved protein
TTC1370 COG2132 (Q)    laccase
TTC1372 COG3462 (S)    hypothetical protein
TTC1374 COG0822 (C)    IscU protein
TTC1375 NOCOG          hypothetical protein
TTC1379 NOCOG          hypothetical protein
TTC1383 COG1371 (S)    hypothetical protein Cluster function:
                       PHOSPHORIBOSYLFORMYLGLYCINAMIDINE SYNTHASE
TTC1390 NOCOG          putative mannosyltransferase
TTC1391 COG0438 (M)    conserved hypothetical protein
TTC1392 NOCOG          hypothetical protein
TTC1394 COG1598 (S)    hypothetical protein
TTC1395 COG1724 (N)    hypothetical conserved protein
TTC1398 NOCOG          hypothetical protein
TTC1399 COG0446 (R)    sulfide dehydrogenase (flavocytochrome C)
TTC1400 COG1138 (O)    hypothetical protein
TTC1401 COG2010 (C)    high-affinity iron permease
TTC1402 NOCOG          hypothetical conserved protein
TTC1403 COG5557 (C)    polysulfide reductase chain CHAIN C
TTC1404 COG0437 (C)    anaerobic dimethyl sulfoxide
```

		reductase chain B
TTC1405	NOCOG	hypothetical protein
TTC1406	NOCOG	chaperone protein htpG
TTC1410	COG1664 (M)	hypothetical cytosolic protein
TTC1413	COG4636 (S)	hypothetical conserved protein
TTC1417	NOCOG	hypothetical protein
TTC1421	COG1057 (H)	nicotinate-nucleotide adenylyltransferase
TTC1425	COG1314 (U)	hypothetical protein
TTC1428	COG0286 (V)	hypothetical protein
TTC1429	NOCOG	hypothetical protein
TTC1430	COG0433 (R)	hypothetical protein
TTC1431	NOCOG	transposase
TTC1432	COG3328 (L)	transposase
TTC1433	NOCOG	hypothetical protein
TTC1434	NOCOG	hypothetical protein
TTC1435	NOCOG	hypothetical protein
TTC1440	NOCOG	hypothetical protein
TTC1441	COG0477 (GEPR)	transporter
TTC1443	COG1803 (G)	methylglyoxal synthase
TTC1444	COG1051 (F)	hypothetical protein
TTC1445	COG0301 (H)	thiamine biosynthesis protein thiI
TTC1447	COG1055 (P)	transporter
TTC1449	COG2234 (R)	putative aminopeptidase
TTC1450	NOCOG	hypothetical conserved protein
TTC1451	COG2154 (H)	pterin-4-alpha-carbinolamine dehydratase
TTC1453	NOCOG	hypothetical protein
TTC1458	NOCOG	hypothetical protein
TTC1459	COG0235 (G)	L-fuculose phosphate aldolase
TTC1470	NOCOG	hypothetical protein
TTC1472	NOCOG	hypothetical protein
TTC1475	NOCOG	hypothetical protein
TTC1479	NOCOG	hypothetical protein
TTC1481	NOCOG	hypothetical protein
TTC1495	COG0465 (O)	hypothetical protein
TTC1496	COG0457 (R)	tetratricopeptide repeat family protein
TTC1497	NOCOG	hypothetical protein
TTC1498	NOCOG	hypothetical protein
TTC1499	NOCOG	hypothetical protein
TTC1500	NOCOG	hypothetical protein
TTC1506	NOCOG	hypothetical protein
TTC1507	COG1853 (R)	flavoredoxin
TTC1511	NOCOG	hypothetical protein
TTC1512	NOCOG	hypothetical protein
TTC1513	COG1653 (G)	sugar-binding protein
TTC1515	COG3327 (K)	phaN
TTC1520	COG3342 (S)	major pilin protein fimA
TTC1522	NOCOG	hypothetical protein
TTC1524	NOCOG	hypothetical protein
TTC1537	NOCOG	hypothetical conserved protein
TTC1540	COG1543 (S)	conserved hypothetical protein
TTC1544	NOCOG	hypothetical conserved protein
TTC1545	NOCOG	hypothetical conserved protein
TTC1548	NOCOG	hypothetical protein
TTC1549	COG1724 (N)	hypothetical protein
TTC1561	NOCOG	hypothetical protein
TTC1562	NOCOG	hypothetical protein
TTC1566	NOCOG	hypothetical protein
TTC1571	COG1082 (G)	hypothetical conserved protein
TTC1581	COG0393 (S)	hypothetical conserved protein
TTC1582	NOCOG	hypothetical protein
TTC1587	NOCOG	hypothetical protein
TTC1591	COG0674 (C)	2-oxoacid-ferredoxin oxidoreductase
TTC1592	COG1013 (C)	2-oxoacid-ferredoxin oxidoreductase
TTC1593	COG0665 (E)	hypothetical protein
TTC1596	NOCOG	hypothetical protein
TTC1599	COG1405 (K)	hypothetical protein
TTC1600	COG0221 (C)	Inorganic pyrophosphatase
TTC1601	NOCOG	hypothetical protein
TTC1612	NOCOG	hypothetical protein
TTC1617	NOCOG	hypothetical conserved protein
TTC1633	COG0589 (T)	hypothetical protein
TTC1634	COG0517 (R)	inosine-5'-monophosphate dehydrogenase
TTC1635	NOCOG	hypothetical protein
TTC1644	COG0457 (R)	hypothetical protein
TTC1645	COG1912 (S)	hypothetical conserved protein
TTC1646	NOCOG	hypothetical protein
TTC1649	COG1993 (S)	hypothetical conserved protein
TTC1651	NOCOG	hypothetical protein
TTC1655	COG1335 (Q)	putative isochorismatase
TTC1656	COG0685 (E)	methylenetetrahydrofolate reductase
TTC1658	NOCOG	hypothetical protein
TTC1661	NOCOG	hypothetical protein
TTC1666	NOCOG	hypothetical protein
TTC1667	NOCOG	hypothetical protein
TTC1669	NOCOG	hypothetical protein
TTC1670	NOCOG	hypothetical protein
TTC1681	NOCOG	hypothetical protein
TTC1682	NOCOG	hypothetical membrane spanning protein
TTC1683	NOCOG	hypothetical protein
TTC1684	NOCOG	hypothetical protein
TTC1685	NOCOG	hypothetical protein
TTC1686	NOCOG	hypothetical protein

```
TTC1687 NOCOG          hypothetical protein
TTC1694 NOCOG          hypothetical protein
TTC1704 COG1598 (S)    hypothetical protein
TTC1705 NOCOG          hypothetical protein
TTC1707 COG2759 (F)    formate--tetrahydrofolate ligase
TTC1711 COG1042 (C)    6-carboxyhexanoate--CoA ligase
TTC1717 NOCOG          hypothetical conserved protein
TTC1726 COG0226 (P)    phosphate-binding protein
TTC1728 COG3872 (R)    hypothetical conserved protein
TTC1730 COG0735 (P)    ferric uptake regulation protein
TTC1731 NOCOG          hypothetical protein
TTC1732 NOCOG          hypothetical protein
TTC1748 NOCOG          hypothetical protein
TTC1750 NOCOG          methyltransferase
TTC1752 NOCOG          hypothetical protein
TTC1755 NOCOG          hypothetical conserved protein
TTC1759 NOCOG          hypothetical protein
TTC1760 NOCOG          hypothetical conserved protein
TTC1773 NOCOG          hypothetical protein
TTC1778 COG1282 (C)    NAD(P) transhydrogenase subunit beta
TTC1779 COG3288 (C)    NAD(P) transhydrogenase subunit alpha 2
TTC1781 COG0697 (GER)  permease
TTC1783 COG0477 (GEPR) transporter
TTC1784 COG3177 (S)    hypothetical conserved protein
TTC1790 COG4663 (Q)    C4-dicarboxylate-binding protein
TTC1792 COG4665 (Q)    tripartite transporter, small subunit
TTC1794 COG0589 (T)    universal stress protein family
TTC1803 COG0212 (H)    5-formyltetrahydrofolate cyclo-ligase
TTC1804 COG4113 (R)    hypothetical conserved protein
TTC1805 NOCOG          hypothetical protein
TTC1808 NOCOG          hypothetical conserved protein
TTC1817 NOCOG          hypothetical conserved protein
TTC1820 NOCOG          cyanophycinase
TTC1826 COG1341 (R)    hypothetical protein
TTC1827 NOCOG          hypothetical protein
TTC1828 COG1449 (G)    amylopullulanase
TTC1830 COG1846 (K)    transcriptional regulators, marR/emrR family
TTC1831 NOCOG          hypothetical membrane spanning protein
TTC1832 NOCOG          hypothetical protein
TTC1834 COG2414 (C)    tungsten-containing aldehyde
                       ferredoxin oxidoreductase
TTC1835 COG1977 (H)    molybdopterin converting factor,
                       small subunit
TTC1836 NOCOG          hypothetical protein
TTC1837 COG3307 (M)    hypothetical membrane spanning protein
TTC1838 NOCOG          general secretion pathway protein G
TTC1839 NOCOG          general secretion pathway protein G
TTC1840 NOCOG          hypothetical protein
TTC1841 NOCOG          hypothetical protein
TTC1842 NOCOG          hypothetical protein
TTC1843 NOCOG          hypothetical protein
TTC1846 NOCOG          hypothetical protein
TTC1853 NOCOG          hypothetical protein
TTC1864 COG0717 (F)    deoxycytidine triphosphate deaminase
TTC1868 NOCOG          hypothetical protein
TTC1869 NOCOG          hypothetical protein
TTC1872 COG3546 (P)    pseudocatalase
TTC1878 NOCOG          hypothetical protein
TTC1881 NOCOG          transposase
TTC1882 NOCOG          hypothetical protein
TTC1883 NOCOG          hypothetical conserved protein
TTC1884 COG1896 (R)    hydrolase (had superfamily)
TTC1885 NOCOG          hypothetical conserved protein
TTC1887 COG4636 (S)    hypothetical conserved protein
TTC1891 NOCOG          hypothetical conserved protein
TTC1897 NOCOG          hypothetical membrane spanning protein
TTC1900 NOCOG          hypothetical conserved protein
TTC1901 COG2070 (R)    2-nitropropane dioxygenase
TTC1902 NOCOG          hypothetical protein
TTC1911 COG0839 (C)    NADH-quinone oxidoreductase chain CHAIN J
TTC1924 NOCOG          hypothetical protein
TTC1925 NOCOG          hypothetical protein
TTC1926 NOCOG          hypothetical protein
TTC1927 COG5551 (S)    hypothetical conserved protein
TTC1932 NOCOG          hypothetical protein
TTC1935 COG4662 (H)    tungstate transport system permease protein
TTC1936 COG2998 (H)    tungstate-binding protein precursor
TTC1938 COG4448 (E)    L-asparaginase II
TTC1939 NOCOG          hypothetical protein
TTC1941 COG0477 (GEPR) multidrug resistance protein, putative
TTC1943 NOCOG          hypothetical protein
TTC1946 NOCOG          hypothetical protein
TTC1948 COG1420 (K)    heat-inducible transcription repressor hrcA
TTC1953 NOCOG          putative phosphohydrolases
TTC1958 NOCOG          probable nucleotidyltransferase
TTC1959 NOCOG          probable nucleotidyltransferase
TTC1961 COG3400 (S)    zrk system potassium uptake protein trkA
TTC1962 COG0475 (P)    trk system potassium uptake protein trkA
TTC1965 COG0739 (M)    membrane proteins related to
                       metalloendopeptidases
TTC1967 NOCOG          hypothetical protein
TTC1968 COG0477 (GEPR) probable macrolide-efflux transmembrane protein
```

TTC1970	COG3369 (S)	hypothetical conserved protein
TTC1971	NOCOG	hypothetical protein
TTC1974	NOCOG	hypothetical conserved protein
TTC1977	COG0778 (C)	glucose-1-phosphate adenylyltransferase
TTC1981	COG0457 (R)	tetratricopeptide repeat family protein
TTC1984	NOCOG	hypothetical protein
TTC1985	NOCOG	hypothetical protein
TTP0003	NOCOG	hypothetical protein
TTP0004	NOCOG	cobalamin [5'-phosphate] synthase
TTP0007	COG3376 (P)	hypothetical conserved protein
TTP0008	COG1903 (H)	cobalamin biosynthetic protein cbiD
TTP0009	COG2082 (H)	precorrin-8X methylmutase
TTP0010	COG2242 (H)	precorrin-6Y C5,15-methyltransferase [decarboxylating]
TTP0011	COG2243 (H)	precorrin-2 C20-methyltransferase
TTP0014	COG2073 (H)	cbiG protein
TTP0015	COG2138 (S)	cbiX protein
TTP0016	NOCOG	hypothetical protein
TTP0018	COG0778 (C)	nicotinate-nucleotide--dimethylbenzimidazolephosphoribosyltransferase
TTP0020	COG2250 (S)	hypothetical conserved protein
TTP0021	COG1708 (R)	hypothetical conserved protein
TTP0026	NOCOG	hypothetical protein
TTP0027	NOCOG	hypothetical conserved protein
TTP0029	COG1082 (G)	hypothetical conserved protein
TTP0030	COG0800 (G)	putative KHG/KDPG aldolase
TTP0031	COG1638 (G)	ABC transporter, periplasmic substrate-binding protein, putative
TTP0032	COG3090 (G)	probable transporter
TTP0035	COG2055 (C)	dehydrogenase
TTP0042	COG2723 (G)	beta-glycosidase
TTP0044	NOCOG	hypothetical conserved protein
TTP0046	NOCOG	hypothetical protein
TTP0047	NOCOG	probable phytoene dehydrogenase
TTP0048	COG0702 (MG)	nucleoside-diphosphate-sugar epimerase
TTP0049	NOCOG	hypothetical protein
TTP0050	COG1853 (R)	hypothetical conserved protein
TTP0051	NOCOG	hypothetical protein
TTP0053	NOCOG	hypothetical protein
TTP0056	COG0789 (K)	probable transcriptional regulator
TTP0058	COG0415 (L)	deoxyribodipyrimidine photolyase
TTP0060	NOCOG	hypothetical conserved membrane protein
TTP0063	NOCOG	hypothetical conserved protein
TTP0064	NOCOG	hypothetical protein
TTP0068	COG1708 (R)	hypothetical conserved protein
TTP0069	COG1683 (S)	hypothetical conserved protein
TTP0071	COG1085 (C)	galactose-1-phosphate uridylyltransferase
TTP0072	COG3345 (G)	alpha-galactosidase
TTP0073	NOCOG	hypothetical protein
TTP0075	COG0154 (J)	hypothetical protein
TTP0076	NOCOG	hypothetical protein
TTP0077	NOCOG	hypothetical protein
TTP0080	NOCOG	hypothetical conserved protein
TTP0081	COG1373 (R)	probable ATPase
TTP0082	NOCOG	hypothetical conserved protein
TTP0086	COG0243 (C)	hypothetical protein
TTP0087	NOCOG	hypothetical protein
TTP0088	NOCOG	hypothetical membrane spanning protein
TTP0089	NOCOG	hypothetical protein
TTP0090	NOCOG	hypothetical conserved protein
TTP0091	NOCOG	probable cytochrome
TTP0092	NOCOG	hypothetical conserved protein
TTP0093	NOCOG	hypothetical conserved protein
TTP0094	NOCOG	hypothetical protein
TTP0096	NOCOG	hypothetical conserved protein
TTP0097	COG3744 (S)	hypothetical conserved protein
TTP0098	NOCOG	hypothetical protein
TTP0100	NOCOG	probable transposase
TTP0101	COG1343 (L)	hypothetical protein
TTP0102	COG1353 (R)	putative hydrolase of the HD superfamily (permuted catalytic motifs)
TTP0103	NOCOG	hypothetical protein
TTP0104	COG1337 (L)	hypothetical conserved protein
TTP0105	NOCOG	hypothetical conserved protein
TTP0106	COG1332 (L)	hypothetical protein
TTP0107	NOCOG	hypothetical protein
TTP0108	NOCOG	hypothetical protein
TTP0109	NOCOG	hypothetical protein
TTP0110	NOCOG	hypothetical protein
TTP0111	NOCOG	hypothetical protein
TTP0112	NOCOG	hypothetical protein
TTP0113	NOCOG	hypothetical protein
TTP0114	NOCOG	hypothetical protein
TTP0115	COG1353 (R)	hypothetical protein
TTP0116	NOCOG	hypothetical conserved protein
TTP0117	COG1367 (L)	hypothetical conserved protein
TTP0118	COG1336 (L)	hypothetical conserved protein
TTP0119	NOCOG	hypothetical protein
TTP0120	COG1604 (L)	hypothetical protein
TTP0121	NOCOG	hypothetical conserved protein
TTP0122	COG2414 (C)	probable tungsten-containing aldehyde ferredoxin oxidoreductase

```
TTP0123 COG1266 (R)    hypothetical membrane spanning protein
TTP0124 COG1110 (L)    hypothetical protein
TTP0125 NOCOG          hypothetical protein
TTP0126 NOCOG          hypothetical protein
TTP0127 NOCOG          hypothetical protein
TTP0132 COG1203 (R)    hypothetical conserved protein
TTP0133 NOCOG          hypothetical conserved protein
TTP0134 NOCOG          hypothetical conserved protein
TTP0135 NOCOG          hypothetical conserved protein
TTP0136 COG1468 (L)    hypothetical conserved protein
TTP0139 NOCOG          hypothetical protein
TTP0140 COG1373 (R)    hypothetical conserved protein
TTP0141 NOCOG          hypothetical protein
TTP0143 NOCOG          hypothetical conserved protein
TTP0144 COG1178 (P)    hypothetical protein
TTP0145 NOCOG          hypothetical protein
TTP0146 NOCOG          hypothetical protein
TTP0147 NOCOG          hypothetical protein
TTP0148 NOCOG          hypothetical protein
TTP0149 NOCOG          hypothetical protein
TTP0151 NOCOG          hypothetical conserved protein
TTP0152 NOCOG          hypothetical protein
TTP0153 NOCOG          hypothetical protein
TTP0154 NOCOG          hypothetical protein
TTP0155 NOCOG          hypothetical protein
TTP0156 NOCOG          hypothetical protein
TTP0157 COG5433 (L)    hypothetical protein
TTP0158 NOCOG          hypothetical protein
TTP0159 NOCOG          hypothetical conserved protein
TTP0161 COG0208 (F)    ribonucleoside-diphosphate reductase
                       beta chain
TTP0163 NOCOG          hypothetical protein
TTP0166 NOCOG          hypothetical protein
TTP0169 NOCOG          hypothetical conserved protein
TTP0173 COG3488 (C)    hypothetical conserved protein
TTP0174 NOCOG          hypothetical protein
TTP0176 NOCOG          hypothetical protein
TTP0180 NOCOG          hypothetical conserved protein
TTP0182 COG3328 (L)    transposase
TTP0184 COG2002 (K)    hypothetical protein
TTP0185 NOCOG          hypothetical protein
TTP0186 COG2963 (L)    hypothetical conserved protein
TTP0188 NOCOG          hypothetical protein
TTP0189 COG4993 (G)    hypothetical protein
TTP0190 NOCOG          pilA
TTP0191 COG0210 (L)    putative helicase superfamily I
TTP0192 NOCOG          hypothetical protein
TTP0193 NOCOG          hypothetical protein
TTP0195 COG1343 (L)    hypothetical conserved protein
TTP0196 COG1518 (L)    hypothetical conserved protein
TTP0197 COG1468 (L)    hypothetical conserved protein
TTP0198 COG2378 (K)    hypothetical protein
TTP0199 COG0515 (RTKL) hypothetical protein
TTP0200 NOCOG          hypothetical protein
TTP0201 NOCOG          hypothetical protein
TTP0202 COG1203 (R)    ATP-dependent helicase with frameshift
TTP0203 NOCOG          ATP-dependent helicase with frameshift
TTP0204 NOCOG          hypothetical conserved protein
TTP0205 COG3547 (L)    probable DNA-binding protein
TTP0206 NOCOG          hypothetical protein
TTP0207 NOCOG          hypothetical protein
TTP0208 NOCOG          hypothetical protein
TTP0209 NOCOG          hypothetical protein
TTP0210 NOCOG          transposase
TTP0211 COG1373 (R)    ATPase
TTP0212 COG2414 (C)    hypothetical protein
TTP0213 COG3547 (L)    transposase
TTP0214 NOCOG          hypothetical protein
TTP0215 COG1518 (L)    hypothetical protein
TTP0217 NOCOG          hypothetical conserved protein
TTP0218 COG3394 (S)    hypothetical conserved protein
TTP0219 COG2211 (G)    hypothetical conserved protein
TTP0220 COG3250 (G)    probable beta-galactosidase protein
TTP0221 COG1501 (G)    alpha-glucosidase
TTP0222 COG3250 (G)    probable beta-galactosidase
TTP0223 COG2211 (G)    hypothetical conserved protein
TTP0224 NOCOG          hypothetical protein
TTP0225 NOCOG          hypothetical protein
TTP0226 NOCOG          hypothetical protein
TTP0227 NOCOG          hypothetical protein
TTP0228 NOCOG          hypothetical conserved protein
TTP0229 NOCOG          hypothetical protein
TTP0230 NOCOG          hypothetical protein
```

A.7. Fremdgene in *T. thermophilus*

Genname, COG, (COG-Gruppe), Annotation

```
TTC0270 NOCOG       probable nucleotidyltransferase
TTC0272 NOCOG       hypothetical protein
TTC0273 NOCOG       transposase
TTC0277 NOCOG       hypothetical protein
TTC0278 NOCOG       hypothetical protein
TTC0279 COG0463 (M) probable glycosyl transferase
TTC0280 COG0438 (M) lipopolysaccharideN-acetylglucosaminyltransferase
TTC0281 NOCOG       hypothetical protein
TTC0282 COG0367 (E) asparagine synthetase [glutamine-hydrolyzing]
TTC0283 NOCOG       hypothetical protein
TTC0284 COG0399 (M) pleiotropic regulatory protein
TTC0285 COG0381 (M) UDP-N-acetylglucosamine 2-epimerase
TTC0286 COG0110 (R) acetyltransferase
TTC0649 COG0577 (V) ABC transporter permease protein
TTC0651 NOCOG       hypothetical protein
TTC0653 NOCOG       hypothetical protein
TTC0654 NOCOG       hypothetical protein
TTC0655 NOCOG       hypothetical protein
TTC0658 NOCOG       hypothetical protein
TTC0659 NOCOG       hypothetical protein
TTC0660 NOCOG       hypothetical protein
TTC0661 NOCOG       hypothetical protein
TTC0662 NOCOG       hypothetical protein
TTC0664 NOCOG       hypothetical conserved protein
TTC0665 COG0582 (L) transposase
TTC0858 NOCOG       pilA
TTC0859 COG1196 (D) hypothetical protein
TTC0860 COG4636 (S) conserved hypothetical protein
TTC0861 NOCOG       hypothetical protein
TTC0862 COG4636 (S) conserved hypothetical protein
TTC1421 COG1057 (H) nicotinate-nucleotide adenylyltransferase
TTC1423 COG0211 (J) LSU ribosomal protein L27P
TTC1424 COG0261 (J) LSU ribosomal protein L21P
TTC1425 COG1314 (U) hypothetical protein
TTC1428 COG0286 (V) hypothetical protein
TTC1429 NOCOG       hypothetical protein
TTC1430 COG0433 (R) hypothetical protein
TTC1431 NOCOG       transposase
TTC1432 COG3328 (L) transposase
TTC1434 NOCOG       hypothetical protein
TTC1868 NOCOG       hypothetical protein
TTC1869 NOCOG       hypothetical protein
TTC1871 COG0583 (K) transcriptional regulatory protein, lysR family
                    (hydrogen peroxide-inducible genes activator)
TTC1876 COG0582 (L) DNA integration/recombination/invertion protein
TTC1877 COG1002 (V) hypothetical conserved protein
TTC1878 NOCOG       hypothetical protein
TTC1879 COG0433 (R) hypothetical cytosolic protein
TTC1880 COG0863 (L) hypothetical protein
TTC1881 NOCOG       transposase
TTC1882 NOCOG       hypothetical protein
TTC1883 NOCOG       hypothetical conserved protein
TTC1884 COG1896 (R) hydrolase (had superfamily)
TTC1885 NOCOG       hypothetical conserved protein
TTP0064 NOCOG       hypothetical protein
TTP0068 COG1708 (R) hypothetical conserved protein
TTP0070 NOCOG       hypothetical conserved protein
TTP0073 NOCOG       hypothetical protein
TTP0075 COG0154 (J) hypothetical protein
TTP0078 NOCOG       hypothetical conserved protein
TTP0079 COG2206 (T) hypothetical conserved protein
TTP0080 NOCOG       hypothetical conserved protein
TTP0082 NOCOG       hypothetical conserved protein
TTP0092 NOCOG       hypothetical conserved protein
TTP0093 NOCOG       hypothetical conserved protein
TTP0094 NOCOG       hypothetical protein
TTP0099 COG3335 (L) probable transposase
TTP0109 NOCOG       hypothetical protein
TTP0127 NOCOG       hypothetical protein
TTP0131 NOCOG       hypothetical protein
TTP0132 COG1203 (R) hypothetical conserved protein
TTP0133 NOCOG       hypothetical conserved protein
TTP0134 NOCOG       hypothetical conserved protein
TTP0135 NOCOG       hypothetical conserved protein
TTP0136 COG1468 (L) hypothetical conserved protein
TTP0139 NOCOG       hypothetical protein
TTP0140 COG1373 (R) hypothetical conserved protein
TTP0141 NOCOG       hypothetical protein
TTP0142 COG1487 (R) putative plasmid stability protein Y4JK
TTP0143 NOCOG       hypothetical conserved protein
TTP0145 NOCOG       hypothetical protein
TTP0146 NOCOG       hypothetical protein
TTP0147 NOCOG       hypothetical protein
```

```
TTP0148 NOCOG       hypothetical protein
TTP0149 NOCOG       hypothetical protein
TTP0151 NOCOG       hypothetical conserved protein
TTP0152 NOCOG       hypothetical protein
TTP0153 NOCOG       hypothetical protein
TTP0154 NOCOG       hypothetical protein
TTP0155 NOCOG       hypothetical protein
TTP0156 NOCOG       hypothetical protein
TTP0157 COG5433 (L) hypothetical protein
TTP0159 NOCOG       hypothetical conserved protein
TTP0176 NOCOG       hypothetical protein
TTP0179 COG4558 (P) hemin-binding periplasmic protein
                    hmuT precursor
TTP0180 NOCOG       hypothetical conserved protein
TTP0182 COG3328 (L) transposase
TTP0184 COG2002 (K) hypothetical protein
TTP0186 COG2963 (L) hypothetical conserved protein
TTP0187 COG2801 (L) putative transposase
TTP0188 NOCOG       hypothetical protein
TTP0189 COG4993 (G) hypothetical protein
TTP0190 NOCOG       pilA
TTP0195 COG1343 (L) hypothetical conserved protein
TTP0196 COG1518 (L) hypothetical conserved protein
TTP0197 COG1468 (L) hypothetical conserved protein
TTP0198 COG2378 (K) hypothetical protein
TTP0199 COG0515 (L) hypothetical protein
TTP0201 NOCOG       hypothetical protein
TTP0203 NOCOG       ATP-dependent helicase with frameshift
TTP0204 NOCOG       hypothetical conserved protein
TTP0205 COG3547 (L) probable DNA-binding protein
TTP0206 NOCOG       hypothetical protein
TTP0209 NOCOG       hypothetical protein
TTP0210 NOCOG       transposase
TTP0212 COG2414 (C) hypothetical protein
TTP0214 NOCOG       hypothetical protein
TTP0215 COG1518 (L) hypothetical protein
TTP0216 COG0739 (M) hypothetical protein
TTP0217 NOCOG       hypothetical conserved protein
TTP0224 NOCOG       hypothetical protein
TTP0226 NOCOG       hypothetical protein
TTP0228 NOCOG       hypothetical conserved protein
```

A.8. *T. thermophilus*-spezifische Proteine (COG-Gruppe C)

Genname, COG, (COG-Gruppe), Annotation

```
TTC0012 COG2414 (C)     probable tungsten-containing aldehyde ferredoxin oxidoreductase
TTC0022 COG1145 (C)     hypothetical protein
TTC0128 COG0277 (C)     (S)-2-hydroxy-acid oxidase subunit glcE
TTC0154 COG0437 (C)     nrfC protein
TTC0306 COG2141 (C)     putative oxidoreductase
TTC0769 COG1622 (C)     subunit II of C(O/B)3-type cytochrome c oxidase
TTC0911 COG0636 (C)     V-type sodium ATP synthase subunit K Homolog
TTC1045 COG3258 (C)     cytochrome c
TTC1058 COG2010 (C)     cytochrome c-552 precursor
TTC1083 COG0427 (C)     4-hydroxybutyrate coenzyme A transferase
TTC1240 COG0374 (C)     conserved hypothetical protein
TTC1374 COG0822 (C)     IscU protein
TTC1401 COG2010 (C)     high-affinity iron permease
TTC1403 COG5557 (C)     polysulfide reductase chain CHAIN C
TTC1404 COG0437 (C)     anaerobic dimethyl sulfoxide reductase chain B
TTC1591 COG0674 (C)     2-oxoacid-ferredoxin oxidoreductase
TTC1592 COG1013 (C)     2-oxoacid-ferredoxin oxidoreductase
TTC1600 COG0221 (C)     Inorganic pyrophosphatase
TTC1711 COG1042 (C)     6-carboxyhexanoate--CoA ligase
TTC1778 COG1282 (C)     NAD(P) transhydrogenase subunit beta
TTC1779 COG3288 (C)     NAD(P) transhydrogenase subunit alpha 2
TTC1834 COG2414 (C)     tungsten-containing aldehyde ferredoxin oxidoreductase
TTC1911 COG0839 (C)     NADH-quinone oxidoreductase chain CHAIN J
TTC1977 COG0778 (C)     glucose-1-phosphate adenylyltransferase
TTP0018 COG0778 (C)     nicotinate-nucleotide--dimethylbenzimidazolephosphoribosyltransferase
TTP0035 COG2055 (C)     dehydrogenase
TTP0071 COG1085 (C)     galactose-1-phosphate uridylyltransferase
TTP0086 COG0243 (C)     hypothetical protein
```

TTP0122	COG2414 (C)	probable tungsten-containing aldehyde ferredoxin oxidoreductase
TTP0173	COG3488 (C)	hypothetical conserved protein
TTP0212	COG2414 (C)	hypothetical protein

A.9. Gecoggte *D. radiodurans*-spezifische Proteine

Genname, COG, (COG-Gruppe), Annotation

NP-293731.1	COG1378 (K)	conserved hypothetical protein
NP-293736.1	COG3385 (L)	putative transposase
NP-293737.1	COG0702 (MG)	conserved hypothetical protein
NP-293759.1	COG0251 (J)	conserved hypothetical protein
NP-293771.1	COG3644 (S)	hypothetical protein
NP-293774.1	COG2323 (S)	conserved hypothetical protein
NP-293776.1	COG2318 (S)	hypothetical protein
NP-293777.1	COG0316 (S)	HesB/YadR/YfhF family protein
NP-293778.1	COG3246 (S)	hypothetical protein
NP-293780.1	COG1416 (S)	conserved hypothetical protein
NP-293781.1	COG2378 (K)	repressor protein, putative
NP-293782.1	COG3135 (Q)	benzoate membrane transport protein, putative
NP-293783.1	COG0526 (OC)	conserved hypothetical protein
NP-293792.1	COG3231 (J)	aminoglycoside 3‘-phosphotransferase
NP-293799.1	COG0857 (R)	phosphate acetyltransferase
NP-293801.1	COG1511 (S)	conserved hypothetical protein
NP-293805.1	COG0494 (LR)	MutT/nudix family protein
NP-293817.1	COG2324 (S)	conserved hypothetical protein
NP-293819.1	COG1233 (Q)	phytoene dehydrogenase, putative
NP-293823.1	COG4293 (S)	conserved hypothetical protein
NP-293833.1	COG0534 (V)	conserved hypothetical protein
NP-293834.1	COG2409 (R)	conserved hypothetical protein
NP-293835.1	COG0346 (E)	hypothetical protein
NP-293838.1	COG3823 (O)	glutamine cyclotransferase
NP-293847.1	COG0457 (R)	hypothetical protein
NP-293858.1	COG2326 (S)	conserved hypothetical protein
NP-293861.1	COG0697 (GER)	conserved hypothetical protein
NP-293862.1	COG1393 (P)	conserved hypothetical protein
NP-293863.1	COG2810 (V)	hypothetical protein
NP-293864.1	COG3296 (S)	conserved hypothetical protein
NP-293873.1	COG0494 (LR)	MutT/nudix family protein
NP-293890.1	COG4281 (I)	acyl-CoA-binding protein
NP-293891.1	COG2856 (E)	hypothetical protein
NP-293900.1	COG2027 (M)	D-alanyl-D-alanine carboxypeptidase, putative
NP-293901.1	COG1943 (L)	transposase, putative
NP-293902.1	COG0675 (L)	transposase, putative
NP-293904.1	COG1595 (K)	RNA polymerase sigma-E factor
NP-293909.1	COG0239 (D)	hypothetical protein
NP-293910.1	COG1570 (L)	exodeoxyribonuclease VII large subunit, putative
NP-293911.1	COG2949 (S)	sanA protein
NP-293912.1	COG0454 (KR)	conserved hypothetical protein
NP-293914.1	COG0501 (O)	conserved hypothetical protein
NP-293915.1	COG3408 (G)	glycogen debranching enzyme-related protein
NP-293916.1	COG1051 (F)	MutT/nudix family protein
NP-293924.1	COG1522 (K)	transcriptional regulator, Lrp/AsnC family
NP-293926.1	COG0563 (F)	hypothetical protein
NP-293934.1	COG1434 (S)	conserved hypothetical protein
NP-293945.1	COG2852 (S)	conserved hypothetical protein
NP-293952.1	COG2967 (P)	apaG protein
NP-293953.1	COG1363 (G)	endoglucanase, putative
NP-293957.1	COG0640 (K)	transcriptional regulator, ArsR family
NP-293958.1	COG3860 (S)	hypothetical protein
NP-293961.1	COG0652 (O)	peptidyl-prolyl cis-trans isomerase, cyclophilin-type
NP-293966.1	COG3502 (S)	conserved hypothetical protein
NP-293968.1	COG0454 (KR)	hypothetical protein
NP-293973.1	COG0667 (C)	oxidoreductase, putative
NP-293983.1	COG2807 (P)	CynX-related transport protein, putative
NP-293984.1	COG1051 (F)	MutT/nudix family protein
NP-293988.1	COG2188 (K)	histidine utilization repressor HutC, putative
NP-293993.1	COG1234 (R)	ElaC family protein
NP-293995.1	COG2329 (R)	conserved hypothetical protein
NP-293996.1	COG2258 (S)	hypothetical protein
NP-293999.1	COG2384 (R)	conserved hypothetical protein
NP-294000.1	COG3386 (G)	gluconolactonase, putative
NP-294017.1	COG1488 (H)	pre-B cell enhancing factor-related protein
NP-294018.1	COG0639 (T)	phosphoprotein phosphatase
NP-294021.1	COG1428 (F)	deoxyguanosine kinase/deoxyadenosine kinase subunit
NP-294022.1	COG1428 (F)	deoxyguanosine kinase/deoxyadenosine kinase subunit
NP-294047.1	COG3404 (E)	serine cycle enzyme, putative
NP-294057.1	COG1075 (R)	lipase, putative

NP-294062.1	COG1473 (R)	N-acyl-L-amino acid amidohydrolase, putative
NP-294074.1	COG4291 (S)	hypothetical protein
NP-294077.1	COG0708 (L)	exodeoxyribonuclease III
NP-294080.1	COG1678 (K)	conserved hypothetical protein
NP-294091.1	COG2013 (S)	conserved hypothetical protein
NP-294096.1	COG0387 (P)	cation exchanger, putative
NP-294097.1	COG1695 (K)	conserved hypothetical protein
NP-294115.1	COG2823 (R)	hypothetical protein
NP-294117.1	COG2334 (R)	conserved hypothetical protein
NP-294122.1	COG2335 (M)	osteoblast specific factor 2-related protein
NP-294126.1	COG1957 (F)	inosine-uridine preferring nucleoside hydrolase
NP-294128.1	COG1523 (G)	alpha-dextran endo-1,6-alpha-glucosidase
NP-294133.1	COG0221 (C)	inorganic pyrophosphatase-related protein
NP-294139.1	COG2336 (T)	ppGpp-regulated growth inhibitor suppressor ChpR/MazE, putative
NP-294140.1	COG2337 (T)	ppGpp-regulated growth inhibitor ChpA/MazF, putative
NP-294143.1	COG1643 (L)	ATP-dependent helicase
NP-294151.1	COG3695 (L)	conserved hypothetical protein
NP-294152.1	COG2010 (C)	cytochrome-related protein
NP-294164.1	COG2339 (S)	conserved hypothetical protein
NP-294168.1	COG5504 (O)	hypothetical protein
NP-294169.1	COG2329 (R)	conserved hypothetical protein
NP-294176.1	COG4290 (F)	guanyl-specific ribonuclease SA
NP-294178.1	COG3570 (V)	streptomycin 3-kinase
NP-294185.1	COG0614 (P)	hypothetical protein
NP-294187.1	COG0296 (G)	maltooligosyltrehalose trehalohydrolase, putative
NP-294191.1	COG0500 (QR)	hypothetical protein
NP-294205.1	COG0330 (O)	B-cell receptor associated protein-related protein
NP-294206.1	COG3012 (S)	conserved hypothetical protein
NP-294207.1	COG0790 (R)	hypothetical protein
NP-294215.1	COG1359 (S)	conserved hypothetical protein
NP-294217.1	COG0563 (F)	DNA topology modulation protein FlaR-related protein
NP-294218.1	COG0259 (H)	pyridoxamine 5-phosphate oxidase
NP-294219.1	COG0493 (ER)	ferredoxin/ferredoxin--NADP reductase, putative
NP-294226.1	COG1520 (S)	hypothetical protein
NP-294231.1	COG1715 (V)	mrr restriction system protein
NP-294233.1	COG2202 (T)	sensory box protein
NP-294264.1	COG1801 (S)	conserved hypothetical protein
NP-294265.1	COG0584 (C)	glycerophosphoryl diester phosphodiesterase, putative
NP-294309.1	COG2127 (S)	conserved hypothetical protein
NP-294310.1	COG1715 (V)	mrr restriction system protein
NP-294322.1	COG2746 (V)	aminoglycoside N3-acetyltransferase
NP-294326.1	COG0590 (FJ)	cytidine/deoxycytidylate deaminase/nudix/methyltransferase domains protein
NP-294331.1	COG0454 (KR)	hypothetical protein
NP-294332.1	COG0645 (R)	hypothetical protein
NP-294334.1	COG2141 (C)	LuxA-related protein
NP-294335.1	COG4187 (E)	arginine utilization protein RocB, putative
NP-294336.1	COG3665 (S)	conserved hypothetical protein
NP-294339.1	COG2018 (R)	conserved hypothetical protein
NP-294340.1	COG1026 (R)	metalloprotease, putative
NP-294343.1	COG1305 (E)	conserved hypothetical protein
NP-294347.1	COG0431 (R)	conserved hypothetical protein
NP-294348.1	COG2013 (S)	conserved hypothetical protein
NP-294355.1	COG3021 (S)	hypothetical protein
NP-294357.1	COG4392 (S)	hypothetical protein
NP-294375.1	COG0025 (P)	Na+/H+ antiporter, putative
NP-294376.1	COG1670 (J)	conserved hypothetical protein
NP-294379.1	COG1301 (C)	proton/sodium-glutamate symport protein
NP-294380.1	COG0400 (R)	serine esterase, putative
NP-294382.1	COG2761 (Q)	frnE protein
NP-294385.1	COG2337 (T)	conserved hypothetical protein
NP-294389.1	COG0675 (L)	transposase, putative
NP-294390.1	COG1943 (L)	transposase, putative
NP-294393.1	COG0346 (E)	hypothetical protein
NP-294398.1	COG0454 (KR)	hypothetical protein
NP-294399.1	COG0454 (KR)	acetyltransferase, putative
NP-294400.1	COG0454 (KR)	acetyltransferase, putative
NP-294404.1	COG0454 (KR)	hypothetical protein
NP-294405.1	COG0454 (KR)	hypothetical protein
NP-294412.1	COG0692 (L)	uracil-DNA N-glycosylase
NP-294413.1	COG3569 (L)	type I topoisomerase, putative
NP-294419.1	COG0636 (C)	v-type ATP synthase, K subunit
NP-294426.1	COG4297 (S)	conserved hypothetical protein
NP-294427.1	COG1896 (R)	hypothetical protein
NP-294428.1	COG2342 (G)	conserved hypothetical protein
NP-294438.1	COG3663 (L)	G/U mismatch-specific DNA glycosylase
NP-294443.1	COG1309 (K)	transcriptional regulator, TetR family
NP-294445.1	COG4891 (S)	hypothetical protein
NP-294470.1	COG2017 (G)	aldose epimerase family protein
NP-294473.1	COG1863 (P)	Multisubunit Na+/H+ antiporter, MnhE subunit
NP-294484.1	COG1680 (V)	hypothetical protein
NP-294487.1	COG3981 (R)	hypothetical protein
NP-294490.1	COG1520 (S)	hypothetical protein
NP-294503.1	COG2819 (R)	conserved hypothetical protein
NP-294516.1	COG0534 (V)	conserved hypothetical protein
NP-294517.1	COG0262 (H)	conserved hypothetical protein
NP-294520.1	COG1670 (J)	acetyltransferase, putative
NP-294521.1	COG0454 (KR)	hypothetical protein
NP-294523.1	COG0673 (R)	glucose-fructose oxidoreductase
NP-294528.1	COG1595 (K)	sigma factor, putative
NP-294533.1	COG0382 (H)	conserved hypothetical protein
NP-294534.1	COG1233 (Q)	dehydrogenase, putative
NP-294535.1	COG3169 (S)	hypothetical protein

NP-294540.1	COG1285 (S)	Mg(2+) transport ATPase-related protein
NP-294553.1	COG2319 (R)	WD-repeat family protein
NP-294554.1	COG1226 (P)	ion transporter, putative
NP-294556.1	COG0451 (MG)	conserved hypothetical protein
NP-294566.1	COG3467 (R)	NimA-related protein
NP-294567.1	COG1305 (E)	conserved hypothetical protein
NP-294585.1	COG1233 (Q)	phytoene dehydrogenase
NP-294586.1	COG1562 (I)	phytoene synthase
NP-294594.1	COG3385 (L)	transposase, putative
NP-294599.1	COG0308 (E)	zinc metalloprotease, putative
NP-294605.1	COG2111 (P)	cation transport system protein, putative
NP-294606.1	COG1006 (P)	cation transport system protein, putative
NP-294607.1	COG0651 (CP)	cation transport system protein, putative
NP-294608.1	COG1863 (P)	conserved hypothetical protein
NP-294609.1	COG2212 (P)	cation transport system protein, putative
NP-294610.1	COG1320 (P)	conserved hypothetical protein
NP-294612.1	COG1652 (S)	hypothetical protein
NP-294617.1	COG0670 (R)	conserved hypothetical protein
NP-294618.1	COG3881 (S)	hypothetical protein
NP-294643.1	COG1721 (R)	conserved hypothetical protein
NP-294647.1	COG2343 (S)	conserved hypothetical protein
NP-294648.1	COG1392 (P)	conserved hypothetical protein
NP-294649.1	COG0306 (P)	low-affinity inorganic phosphate transporter, putative
NP-294654.1	COG1735 (R)	phosphotriesterase, putative
NP-294659.1	COG0639 (T)	phosphatase, putative
NP-294669.1	COG3545 (R)	conserved hypothetical protein
NP-294670.1	COG1363 (G)	conserved hypothetical protein
NP-294671.1	COG1335 (Q)	hypothetical protein
NP-294672.1	COG0526 (OC)	hypothetical protein
NP-294684.1	COG2319 (R)	hypothetical protein
NP-294686.1	COG0431 (R)	hypothetical protein
NP-294688.1	COG2939 (E)	carboxypeptidase-related protein
NP-294689.1	COG2303 (E)	GMC oxidoreductase
NP-294692.1	COG0685 (E)	5,10-methylenetetrahydrofolate reductase-related protein
NP-294702.1	COG0675 (L)	transposase, putative
NP-294703.1	COG1943 (L)	transposase, putative
NP-294726.1	COG2346 (R)	conserved hypothetical protein
NP-294727.1	COG1309 (K)	transcriptional regulator
NP-294738.1	COG0671 (I)	vanadium chloroperoxidase-related protein
NP-294743.1	COG0578 (C)	glycerol-3-phosphate dehydrogenase
NP-294753.1	COG1396 (K)	Transcription regulator
NP-294766.1	COG1695 (K)	Pex-related protein
NP-294773.1	COG4367 (S)	hypothetical protein
NP-294775.1	COG2258 (S)	conserved hypothetical protein
NP-294776.1	COG0421 (E)	hypothetical protein
NP-294777.1	COG1073 (R)	hydrolase, putative
NP-294781.1	COG0454 (KR)	phosphinothricin acetyltransferase, putative
NP-294782.1	COG0454 (KR)	hypothetical protein
NP-294794.1	COG3340 (E)	peptidase E, putative
NP-294795.1	COG2350 (S)	conserved hypothetical protein
NP-294800.1	COG0707 (M)	cell wall synthesis protein, putative
NP-294802.1	COG0500 (QR)	antibiotic biosynthesis protein LmbJ, putative
NP-294823.1	COG3544 (S)	conserved hypothetical protein
NP-294824.1	COG1054 (R)	conserved hypothetical protein
NP-294835.1	COG0789 (K)	transcriptional regulator, MerR family
NP-294836.1	COG3238 (S)	hypothetical protein
NP-294837.1	COG3238 (S)	hypothetical protein
NP-294844.1	COG3426 (C)	kinase, acetokinase family
NP-294851.1	COG3853 (P)	hypothetical protein
NP-294861.1	COG0454 (KR)	hypothetical protein
NP-294868.1	COG3680 (S)	hypothetical protein
NP-294870.1	COG3871 (R)	general stress protein 26, putative
NP-294876.1	COG3257 (R)	conserved hypothetical protein
NP-294878.1	COG0624 (E)	N-carbamyl-L-amino acid amidohydrolase
NP-294882.1	COG3195 (S)	conserved hypothetical protein
NP-294883.1	COG1846 (K)	transcriptional regulator, MarR family
NP-294884.1	COG3648 (Q)	uricase
NP-294885.1	COG2351 (R)	conserved hypothetical protein
NP-294892.1	COG2816 (L)	MutT/nudix family protein
NP-294895.1	COG0714 (R)	conserved hypothetical protein
NP-294902.1	COG0384 (R)	hypothetical protein
NP-294905.1	COG3871 (R)	hypothetical protein
NP-294914.1	COG3548 (S)	hypothetical protein
NP-294916.1	COG4850 (S)	conserved hypothetical protein
NP-294941.1	COG1414 (K)	transcriptional regulator, IclR family
NP-294944.1	COG1918 (P)	ferrous iron transport protein A
NP-294956.1	COG2165 (NU)	pilin, type IV, putative
NP-294957.1	COG2165 (NU)	pilin, type IV, putative
NP-294961.1	COG4762 (S)	hypothetical protein
NP-294964.1	COG2301 (G)	citrate lyase, beta subunit
NP-294976.1	COG1748 (E)	conserved hypothetical protein
NP-294981.1	COG2133 (G)	L-sorbosone dehydrogenase
NP-294982.1	COG0553 (KL)	SNF2/Rad54 helicase-related protein
NP-294983.1	COG0553 (KL)	DNA helicase, SNF2/RAD54 family
NP-294987.1	COG3236 (S)	conserved hypothetical protein
NP-294991.1	COG0457 (R)	hypothetical protein
NP-294995.1	COG0778 (C)	conserved hypothetical protein
NP-294997.1	COG4292 (S)	conserved hypothetical protein
NP-295019.1	COG5006 (R)	PecM-related protein
NP-295020.1	COG3293 (L)	transposase, putative
NP-295021.1	COG2354 (S)	conserved hypothetical protein
NP-295022.1	COG1032 (C)	conserved hypothetical protein

NP-295036.1	COG1403 (V)	conserved hypothetical protein
NP-295038.1	COG3861 (S)	conserved hypothetical protein
NP-295057.1	COG1472 (G)	glycosyl hyrolase, family 3
NP-295059.1	COG3135 (Q)	transport protein, putative
NP-295063.1	COG3631 (R)	conserved hypothetical protein
NP-295064.1	COG0346 (E)	hypothetical protein
NP-295078.1	COG0671 (I)	phosphatidylglycerophosphatase B-related protein
NP-295080.1	COG2011 (P)	ABC transporter, permease protein
NP-295081.1	COG1464 (P)	outer membrane protein
NP-295082.1	COG1464 (P)	outer membrane protein
NP-295085.1	COG0673 (R)	streptomycin biosynthesis protein StrI-related protein
NP-295101.1	COG0229 (O)	MsrA-related protein
NP-295102.1	COG1309 (K)	transcriptional repressor, TetR family
NP-295104.1	COG0675 (L)	transposase, putative
NP-295105.1	COG1943 (L)	transposase, putative
NP-295108.1	COG3652 (S)	hypothetical protein
NP-295113.1	COG3358 (S)	hypothetical protein
NP-295114.1	COG2501 (S)	conserved hypothetical protein
NP-295119.1	COG0454 (KR)	conserved hypothetical protein
NP-295126.1	COG0596 (R)	hydrolase-related protein
NP-295133.1	COG0019 (E)	carboxynorspermidine decarboxylase
NP-295172.1	COG1346 (M)	conserved hypothetical protein
NP-295173.1	COG1380 (R)	conserved hypothetical protein
NP-295180.1	COG2832 (S)	hypothetical protein
NP-295192.1	COG0860 (M)	hypothetical protein
NP-295204.1	COG3253 (S)	conserved hypothetical protein
NP-295208.1	COG0657 (I)	lipase, putative
NP-295212.1	COG0604 (CR)	conserved hypothetical protein
NP-295246.1	COG3415 (L)	putative transposase
NP-295253.1	COG1487 (R)	hypothetical protein
NP-295256.1	COG3440 (V)	conserved hypothetical protein
NP-295269.1	COG2032 (P)	superoxide dismutase (sodC), Cu-Zn family
NP-295295.1	COG3973 (R)	helicase-related protein
NP-295297.1	COG0688 (I)	conserved hypothetical protein
NP-295315.1	COG1943 (L)	transposase, putative
NP-295316.1	COG0675 (L)	transposase, putative
NP-295319.1	COG0364 (G)	glucose-6-phosphate 1-dehydrogenase
NP-295320.1	COG3429 (G)	glucose 6-phosphate dehydrogenase assembly protein OpcA, putative
NP-295322.1	COG0265 (O)	conserved hypothetical protein
NP-295323.1	COG2356 (L)	conserved hypothetical protein
NP-295324.1	COG1814 (S)	nodulin 21-related protein
NP-295335.1	COG2018 (R)	conserved hypothetical protein
NP-295341.1	COG3385 (L)	transposase, putative
NP-295354.1	COG2357 (S)	conserved hypothetical protein
NP-295355.1	COG0860 (M)	N-acetylmuramoyl-L-alanine amidase, putative
NP-295374.1	COG0675 (L)	transposase, putative
NP-295375.1	COG1943 (L)	transposase, putative
NP-295382.1	COG0339 (E)	oligopeptidase A
NP-295390.1	COG0168 (P)	potassium uptake protein KtrB
NP-295391.1	COG0168 (P)	potassium uptake protein KtrB
NP-295399.1	COG0500 (QR)	hypothetical protein
NP-295401.1	COG0454 (KR)	conserved hypothetical protein
NP-295406.1	COG2808 (K)	transcriptional regulator, putative
NP-295408.1	COG1192 (D)	chromosome partitioning protein, ParA family
NP-295418.1	COG0346 (E)	lactoylglutathione lyase-related protein
NP-295432.1	COG1914 (P)	integral membrane protein, NRAMP family
NP-295434.1	COG1473 (R)	N-acyl-L-amino acid amidohydrolase
NP-295436.1	COG2360 (O)	leu/phe-tRNA-protein transferase
NP-295438.1	COG3603 (S)	conserved hypothetical protein
NP-295446.1	COG2124 (Q)	cytochrome P450
NP-295447.1	COG1233 (Q)	hypothetical protein
NP-295448.1	COG2319 (R)	WD-repeat family protein
NP-295449.1	COG3183 (V)	hypothetical protein
NP-295451.1	COG1806 (S)	conserved hypothetical protein
NP-295462.1	COG1695 (K)	conserved hypothetical protein
NP-295468.1	COG1755 (S)	conserved hypothetical protein
NP-295476.1	COG0850 (D)	septum site-determining protein, putative
NP-295481.1	COG0019 (E)	diaminopimelate decarboxylase
NP-295485.1	COG3385 (L)	putative transposase
NP-295488.1	COG2128 (S)	conserved hypothetical protein
NP-295499.1	COG0494 (LR)	MutT/nudix family protein
NP-295513.1	COG3386 (G)	yellow-related protein
NP-295523.1	COG1670 (J)	acetyltransferase, putative
NP-295540.1	COG0722 (E)	phospho-2-dehydro-3-deoxyheptonate aldolase
NP-295542.1	COG4294 (L)	UV damage endonuclease, putative
NP-295546.1	COG0491 (R)	conserved hypothetical protein
NP-295562.1	COG0545 (O)	peptidyl-prolyl cis-trans isomerase, FKBP-type
NP-295566.1	COG2362 (E)	peptide ABC transporter, putative
NP-295567.1	COG2388 (R)	conserved hypothetical protein
NP-295569.1	COG1695 (K)	conserved hypothetical protein
NP-295571.1	COG0296 (G)	1,4-alpha-glucan branching enzyme, putative
NP-295572.1	COG0225 (O)	peptide methionine sulfoxide reductase
NP-295581.1	COG0730 (R)	conserved hypothetical protein
NP-295587.1	COG2363 (S)	conserved hypothetical protein
NP-295598.1	COG2364 (S)	conserved hypothetical protein
NP-295600.1	COG1787 (V)	conserved hypothetical protein
NP-295606.1	COG2258 (S)	conserved hypothetical protein
NP-295616.1	COG0652 (O)	peptidyl-prolyl cis-trans isomerase, cyclophilin-type
NP-295617.1	COG1522 (K)	transcriptional regulator, AsnC family
NP-295625.1	COG0507 (L)	exodeoxyribonuclease V, subunit RecD, putative
NP-295629.1	COG1620 (C)	L-lactate permease, putative
NP-295633.1	COG3265 (G)	thermoresistant gluconokinase

NP-295635.1	COG2365 (T)	protein-tyrosine phosphatase, putative
NP-295638.1	COG1073 (R)	conserved hypothetical protein
NP-295641.1	COG0454 (KR)	aminoglycoside acetyltransferase (6') type I
NP-295650.1	COG1943 (L)	transposase
NP-295651.1	COG0554 (C)	glycerol kinase
NP-295652.1	COG0580 (G)	glycerol uptake facilitator protein
NP-295654.1	COG1073 (R)	conserved hypothetical protein
NP-295655.1	COG1943 (L)	transposase, putative
NP-295656.1	COG0675 (L)	transposase, putative
NP-295658.1	COG2390 (K)	transcriptional regulator
NP-295663.1	COG3187 (O)	conserved hypothetical protein
NP-295677.1	COG1695 (K)	conserved hypothetical protein
NP-295682.1	COG3575 (S)	conserved hypothetical protein
NP-295684.1	COG0666 (R)	ankyrin-related protein
NP-295694.1	COG0477 (GEPR)	hypothetical protein
NP-295698.1	COG0454 (KR)	hypothetical protein
NP-295701.1	COG0456 (R)	acetyltransferase, putative
NP-295707.1	COG1435 (F)	thymidine kinase
NP-295715.1	COG0705 (R)	conserved hypothetical protein
NP-295716.1	COG1305 (E)	conserved hypothetical protein
NP-295721.1	COG0753 (P)	catalase
NP-295722.1	COG1280 (E)	conserved hypothetical protein
NP-295723.1	COG1619 (V)	mccF protein, putative
NP-295747.1	COG4270 (S)	hypothetical protein
NP-295752.1	COG2306 (S)	hypothetical protein
NP-295754.1	COG0477 (GEPR)	hypothetical protein
NP-295757.1	COG2746 (V)	aminoglycoside N3'-acetyltransferase, type IV
NP-295758.1	COG4762 (S)	conserved hypothetical protein
NP-295764.1	COG2307 (S)	conserved hypothetical protein
NP-295765.1	COG2308 (S)	conserved hypothetical protein
NP-295779.1	COG3187 (O)	hypothetical protein
NP-295780.1	COG3597 (S)	hypothetical protein
NP-295785.1	COG2319 (R)	WD-repeat family protein
NP-295795.1	COG3214 (S)	conserved hypothetical protein
NP-295797.1	COG2094 (L)	3-methyladenine glycosidase
NP-295798.1	COG0633 (C)	ferredoxin
NP-295801.1	COG1075 (R)	lipase, putative
NP-295802.1	COG1695 (K)	conserved hypothetical protein
NP-295807.1	COG0584 (C)	glycerophosphoryl diester phosphodiesterase
NP-295811.1	COG4269 (S)	conserved hypothetical protein
NP-295818.1	COG2010 (C)	c-type cytochrome, putative
NP-295821.1	COG0477 (GEPR)	conserved hypothetical protein
NP-295823.1	COG1359 (S)	conserved hypothetical protein
NP-295825.1	COG1487 (R)	plasmid stability protein, putative
NP-295827.1	COG0346 (E)	conserved hypothetical protein
NP-295853.1	COG0826 (O)	protease, putative
NP-295862.1	COG0645 (R)	hypothetical protein
NP-295864.1	COG3224 (S)	conserved hypothetical protein
NP-295865.1	COG1585 (OU)	conserved hypothetical protein
NP-295866.1	COG0330 (O)	conserved hypothetical protein
NP-295867.1	COG0406 (G)	hypothetical protein
NP-295870.1	COG0561 (R)	hypothetical protein
NP-295875.1	COG0230 (J)	ribosomal protein L34
NP-295882.1	COG3409 (M)	hypothetical protein
NP-295885.1	COG1515 (L)	conserved hypothetical protein
NP-295896.1	COG3878 (S)	conserved hypothetical protein
NP-295903.1	COG2018 (R)	conserved hypothetical protein
NP-295904.1	COG2229 (R)	conserved hypothetical protein
NP-295910.1	COG2230 (M)	cyclopropane-fatty-acyl-phospholipid synthase, putative
NP-295912.1	COG2802 (R)	conserved hypothetical protein
NP-295918.1	COG5339 (S)	hypothetical protein
NP-295927.1	COG1289 (S)	conserved hypothetical protein
NP-295928.1	COG2301 (G)	citrate lyase, beta subunit
NP-295931.1	COG1718 (TD)	SudD-related protein
NP-295932.1	COG3755 (S)	hypothetical protein
NP-295938.1	COG2013 (S)	hypothetical protein
NP-295939.1	COG2310 (T)	tellurium resistance protein TerE
NP-295941.1	COG2301 (G)	hypothetical protein
NP-295942.1	COG3793 (P)	tellurium resistance protein TerB (terB) putative
NP-295943.1	COG2310 (T)	tellurium resistance protein TerD
NP-295944.1	COG3385 (L)	transposase, putative
NP-295945.1	COG4110 (R)	tellurium resistance protein TerA
NP-295946.1	COG2310 (T)	tellurium resistance protein TerZ
NP-295947.1	COG2310 (T)	tellurium resistance protein TerD
NP-295953.1	COG4696 (S)	conserved hypothetical protein
NP-295957.1	COG0613 (R)	conserved hypothetical protein
NP-295960.1	COG0288 (P)	carbonic anhydrase
NP-295963.1	COG2138 (S)	ribosomal protein S2-related protein
NP-295969.1	COG3427 (S)	hypothetical protein
NP-295971.1	COG0639 (T)	protein serine-threonine phosphatase, putative
NP-295972.1	COG1233 (Q)	methoxyneurosporene dehydrogenase
NP-295974.1	COG2311 (S)	conserved hypothetical protein
NP-295975.1	COG0454 (KR)	hypothetical protein
NP-295976.1	COG0454 (KR)	hypothetical protein
NP-295978.1	COG2312 (R)	erythromycin esterase, putative
NP-295980.1	COG1476 (K)	transcriptional regulator, HTH
NP-295984.1	COG0783 (P)	DNA-binding stress response protein, Dps family
NP-295988.1	COG1002 (V)	DNA modification methyltransferase-related protein
NP-295993.1	COG0494 (LR)	MutT/nudix family protein
NP-295994.1	COG0563 (F)	DNA topology modulation protein FlaR-related protein
NP-296002.1	COG0571 (K)	hypothetical protein
NP-296014.1	COG3021 (S)	conserved hypothetical protein

NP-296028.1	COG0477 (GEPR)	multidrug-efflux transporter, putative
NP-296033.1	COG0524 (G)	carbohydrate kinase, PfkB family
NP-296036.1	COG1277 (R)	hypothetical protein
NP-296044.1	COG1943 (L)	transposase, putative
NP-296045.1	COG0675 (L)	transposase, putative
NP-296046.1	COG1404 (O)	serine protease, subtilase family, N-terminal fragment
NP-296055.1	COG0454 (KR)	hypothetical protein
NP-296066.1	COG1408 (R)	conserved hypothetical protein
NP-296074.1	COG0252 (EJ)	L-asparaginase
NP-296076.1	COG0454 (KR)	acetyltransferase, putative
NP-296077.1	COG0494 (LR)	MutT/nudix family protein
NP-296080.1	COG2957 (E)	conserved hypothetical protein
NP-296110.1	COG2261 (S)	transglycosylase associated protein
NP-296119.1	COG3576 (R)	conserved hypothetical protein
NP-296120.1	COG0598 (P)	conserved hypothetical protein
NP-296121.1	COG2315 (S)	conserved hypothetical protein
NP-296129.1	COG0523 (R)	cobW protein, putative
NP-296133.1	COG1649 (S)	hypothetical protein
NP-296134.1	COG2059 (P)	chromate transport protein
NP-296145.1	COG0675 (L)	putative transposase
NP-296148.1	COG1051 (F)	conserved hypothetical protein
NP-296152.1	COG0454 (KR)	conserved hypothetical protein
NP-296159.1	COG2318 (S)	conserved hypothetical protein
NP-296163.1	COG1670 (J)	ribosomal protein N-acetyltransferase, putative
NP-296171.1	COG3544 (S)	conserved hypothetical protein
NP-296174.1	COG1476 (K)	transcriptional regulator, PbsX family
NP-296180.1	COG0727 (R)	hypothetical protein
NP-296183.1	COG0598 (P)	transport protein, putative
NP-296186.1	COG0327 (S)	conserved hypothetical protein
NP-296187.1	COG0645 (R)	hypothetical protein
NP-296188.1	COG0619 (P)	conserved hypothetical protein
NP-296193.1	COG2124 (Q)	cytochrome P450
NP-296198.1	COG1073 (R)	hypothetical protein
NP-296202.1	COG0568 (K)	hypothetical protein
NP-296204.1	COG2319 (R)	WD-repeat family protein
NP-296207.1	COG2863 (C)	cytochrome C4, putative
NP-296210.1	COG1451 (R)	conserved hypothetical protein
NP-296222.1	COG0477 (GEPR)	transport protein, putative
NP-296223.1	COG1505 (E)	prolyl endopeptidase
NP-296239.1	COG0789 (K)	transcriptional regulator, MerR family
NP-296241.1	COG1625 (C)	conserved hypothetical protein
NP-296242.1	COG2267 (I)	lipase, putative
NP-296245.1	COG1301 (C)	C4-dicarboxylate transport protein
NP-296254.1	COG2320 (S)	grpB protein
NP-296257.1	COG2717 (S)	conserved hypothetical protein
NP-296258.1	COG2124 (Q)	cytochrome P450
NP-296262.1	COG0652 (O)	peptidyl-prolyl cis-trans isomerase, cyclophilin-type
NP-296265.1	COG2070 (R)	conserved hypothetical protein
NP-296266.1	COG4248 (R)	conserved hypothetical protein
NP-296271.1	COG1073 (R)	conserved hypothetical protein
NP-296280.1	COG3668 (R)	hypothetical protein
NP-296284.1	COG0406 (G)	hypothetical protein
NP-296294.1	COG1396 (K)	transcriptional regulator, HTH
NP-296304.1	COG1853 (R)	conserved hypothetical protein
NP-296305.1	COG1722 (L)	hypothetical protein
NP-296310.1	COG3411 (C)	hypothetical protein
NP-296315.1	COG2321 (R)	conserved hypothetical protein
NP-296316.1	COG4362 (PE)	nitric oxide synthase-related protein
NP-296320.1	COG3698 (S)	conserved hypothetical protein
NP-296321.1	COG0282 (C)	acetate kinase
NP-296328.1	COG1295 (S)	brkB protein, putative
NP-296331.1	COG0586 (S)	DedA family protein
NP-296341.1	COG0500 (QR)	conserved hypothetical protein
NP-296348.1	COG1695 (K)	conserved hypothetical protein
NP-296349.1	COG0207 (F)	thymidylate synthase
NP-296351.1	COG0262 (H)	dihydrofolate reductase
NP-296352.1	COG1357 (S)	OxrA-related protein
NP-285332.1	COG0842 (V)	conserved hypothetical protein
NP-285336.1	COG2138 (S)	hypothetical protein
NP-285338.1	COG0529 (P)	adenylylsulfate kinase
NP-285343.1	COG0454 (KR)	hypothetical protein
NP-285344.1	COG2087 (H)	cobinamide kinase/cobinamide phosphate guanylyltransferase
NP-285347.1	COG2843 (M)	conserved hypothetical protein
NP-285351.1	COG1509 (E)	conserved hypothetical protein
NP-285359.1	COG2244 (R)	O-antigen transporter RfbX, putative
NP-285361.1	COG0463 (M)	glycosyltransferase
NP-285367.1	COG1898 (M)	lipopolysaccharide biosynthesis protein, putative
NP-285368.1	COG1091 (M)	dTDP-4-rhamnose reductase-related protein
NP-296358.1	COG3230 (P)	Heme oxygenase
NP-285377.1	COG2057 (I)	succinyl-CoA:3-ketoacid-CoA transferase, putative
NP-285382.1	COG1695 (K)	conserved hypothetical protein
NP-285383.1	COG0596 (R)	conserved hypothetical protein
NP-285389.1	COG1820 (G)	N-acetylglucosamine-6-phosphate deacetylase
NP-285390.1	COG2057 (I)	CoA transferase, subunit B
NP-285391.1	COG1788 (I)	CoA transferase, subunit A
NP-285395.1	COG0695 (O)	hypothetical protein
NP-285398.1	COG2801 (L)	transposase, putative
NP-285399.1	COG2842 (R)	hypothetical protein
NP-285420.1	COG2369 (S)	head morphogenesis protein, putative
NP-285431.1	COG5280 (S)	minor tail protein gp26-related protein
NP-285440.1	COG3179 (R)	glycohydrolase, putative
NP-285443.1	COG1002 (V)	conserved hypothetical protein

NP-285446.1	COG3971 (Q)	2-oxo-hepta-3-ene-1,7-dioate hydratase
NP-285451.1	COG2303 (E)	GMC oxidoreductase
NP-285452.1	COG1396 (K)	hypothetical protein
NP-285453.1	COG1975 (O)	conserved hypothetical protein
NP-285454.1	COG0500 (QR)	methyltransferase, putative
NP-285455.1	COG1061 (KL)	conserved hypothetical protein
NP-285459.1	COG1732 (M)	ABC transporter, periplasmic substrate-binding protein, putative
NP-285462.1	COG1174 (E)	amino acid ABC transporter, permease protein
NP-285466.1	COG1309 (K)	transcriptional regulator, TetR family
NP-285469.1	COG2837 (P)	peroxidase, putative
NP-285471.1	COG2986 (E)	histidine ammonia-lyase
NP-285472.1	COG1228 (Q)	imidazolonepropionase
NP-285475.1	COG2987 (E)	urocanate hydratase
NP-285481.1	COG0226 (P)	phosphate ABC transporter, periplasmic phosphate-binding protein
NP-285486.1	COG2391 (R)	hypothetical protein
NP-285487.1	COG2391 (R)	conserved hypothetical protein
NP-285491.1	COG1628 (S)	conserved hypothetical protein
NP-285493.1	COG0725 (P)	molybdenum ABC transporter, periplasmic molybdate-binding protein, putative
NP-296359.1	COG4630 (F)	xanthine dehydrogenase, N-terminal subunit
NP-285502.1	COG4631 (F)	xanthine dehydrogenase, C-terminal subunit
NP-285503.1	COG1975 (O)	conserved hypothetical protein
NP-285507.1	COG2252 (R)	conserved hypothetical protein
NP-285508.1	COG2240 (H)	pyridoxamine kinase
NP-296360.1	COG2124 (Q)	cytochrome P450, putative
NP-285516.1	COG0406 (G)	phosphoglycerate mutase-related protein
NP-285517.1	COG3173 (R)	conserved hypothetical protein
NP-285530.1	COG1649 (S)	conserved hypothetical protein
NP-285534.1	COG2188 (K)	transcriptional regulator, GntR family
NP-296362.1	COG2377 (O)	conserved hypothetical protein
NP-285536.1	COG2103 (R)	conserved hypothetical protein
NP-285537.1	COG0655 (R)	trp repressor binding protein WrbA, putative
NP-285539.1	COG1680 (V)	esterase, putative
NP-285544.1	COG4282 (G)	glucan synthase 1-related protein
NP-285550.1	COG3232 (E)	5-carboxymethyl-2-hydroxymuconate isomerase
NP-285554.1	COG1529 (C)	oxidoreductase
NP-285555.1	COG1319 (C)	oxidoreductase
NP-285556.1	COG2080 (C)	oxidoreductase, iron-sulfur subunit
NP-285558.1	COG2080 (C)	oxidoreductase, iron-sulfur subunit
NP-285559.1	COG1319 (C)	oxidoreductase
NP-285560.1	COG1529 (C)	oxidoreductase
NP-285562.1	COG0368 (H)	cobalamin synthase
NP-285563.1	COG2038 (H)	nicotinate-nucleotide--dimethylbenzimidazolephosphoribosyltransferase
NP-285566.1	COG1018 (C)	flavohemoprotein
NP-285575.1	COG0789 (K)	transcriptional regulator, MerR family
NP-296363.1	COG3385 (L)	putative transposase
NP-285580.1	COG0666 (R)	ankyrin-related protein
NP-285582.1	COG0753 (P)	catalase
NP-285589.1	COG1401 (V)	hypothetical protein
NP-285590.1	COG2191 (C)	conserved hypothetical protein
NP-285593.1	COG1001 (F)	adenine deaminase
NP-285596.1	COG1056 (H)	conserved hypothetical protein
NP-285597.1	COG1231 (E)	flavin monoamine oxidase-related protein
NP-285602.1	COG0577 (V)	hypothetical protein
NP-285607.1	COG1540 (R)	conserved hypothetical protein
NP-285610.1	COG0463 (M)	hypothetical protein
NP-285614.1	COG3836 (G)	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase
NP-285616.1	COG2373 (R)	hypothetical protein
NP-285621.1	COG1276 (P)	hypothetical protein
NP-285622.1	COG2372 (R)	copper resistance protein, putative
NP-285624.1	COG1858 (P)	methylamine utilization protein
NP-285634.1	COG0829 (O)	urease accessory protein UreD
NP-285635.1	COG0378 (OK)	urease accessory protein UreG
NP-285636.1	COG0830 (O)	urease accessory protein UreF
NP-285637.1	COG2371 (O)	urease accessory protein UreE
NP-285638.1	COG0378 (OK)	hydrogenase expression/formation HypB-related protein
NP-285639.1	COG0375 (R)	hydrogenase expression/formation HypA-related protein
NP-285640.1	COG2370 (O)	urease accessory protein UreJ
NP-285641.1	COG0804 (E)	urease, alpha subunit
NP-285642.1	COG0831 (E)	urease, beta/gamma subunit
NP-285645.1	COG4177 (E)	urea/short chain-amide ABC transporter, permease protein
NP-285649.1	COG3424 (Q)	chalcone synthase, putative
NP-285650.1	COG2227 (H)	hypothetical protein
NP-285651.1	COG0654 (HC)	oxidoreductase, putative
NP-285652.1	COG1819 (GC)	glycosyltransferase, putative
NP-285654.1	COG2304 (R)	conserved hypothetical protein
NP-285661.1	COG3844 (E)	kynureninase, putative
NP-285662.1	COG3483 (E)	tryptophan 2,3-dioxygenase, putative
NP-285663.1	COG0657 (I)	conserved hypothetical protein
NP-285667.1	COG1974 (KT)	lexA repressor
NP-285668.1	COG2312 (R)	hypothetical protein
NP-285675.1	COG0840 (NT)	methyl-accepting chemotaxis protein
NP-285676.1	COG0840 (NT)	methyl-accepting chemotaxis-related protein
NP-285677.1	COG0840 (NT)	methyl-accepting chemotaxis protein
NP-285681.1	COG4850 (S)	hypothetical protein
NP-285683.1	COG0498 (E)	threonine synthase
NP-285684.1	COG0598 (P)	cation transporter, putative
NP-285686.1	COG1233 (Q)	P49 secreted protein
NP-285691.1	COG0438 (M)	conserved hypothetical protein
NP-051551.1	COG2109 (H)	cob(I)alamin adenosyltransferase
NP-051552.1	COG1797 (H)	cobyrinic acid a,c-diamide synthase
NP-051553.1	COG1270 (H)	cobalamin biosynthesis protein CobD, putative
NP-051554.1	COG0079 (E)	histidinol-phosphate aminotransferase

NP-051555.1	COG1492 (H)	cobyric acid synthase
NP-051560.1	COG2375 (P)	iron-chelator utilization protein, putative
NP-284927.1	COG3385 (L)	putative transposase
NP-284928.1	COG2801 (L)	transposase
NP-284929.1	COG2963 (L)	transposase
NP-051564.1	COG1073 (R)	hypothetical protein
NP-051566.1	COG1366 (T)	sigma-B regulator RsbS
NP-051567.1	COG2172 (T)	sigma-B regulator RsbT
NP-051583.1	COG0477 (GEPR)	integral membrane protein LmrP
NP-284931.1	COG1785 (P)	alkaline phosphatase
NP-051589.1	COG0577 (V)	hypothetical protein
NP-051591.1	COG3412 (S)	conserved hypothetical protein
NP-284932.1	COG2376 (G)	dihydroxyacetone kinase
NP-284934.1	COG3385 (L)	putative transposase
NP-051595.1	COG3385 (L)	transposase, putative
NP-051598.1	COG3121 (NU)	hypothetical protein
NP-051603.1	COG2374 (R)	extracellular nuclease, putative
NP-051608.1	COG0654 (HC)	salicylate monooxygenase-related protein
NP-051609.1	COG1299 (G)	PTS system, fructose-specific IIBC component
NP-051610.1	COG1105 (G)	1-phosphofructokinase
NP-051611.1	COG1349 (KG)	glycerol-3-phosphate regulon repressor
NP-051619.1	COG2060 (P)	potassium-transporting ATPase, A subunit
NP-051620.1	COG2156 (P)	potassium-transporting ATPase, C subunit
NP-051621.1	COG2205 (T)	KdpD-related protein
NP-051625.1	COG0783 (P)	DNA-binding stress response protein, Dps family
NP-051631.1	COG0617 (J)	hypothetical protein
NP-284938.1	COG3733 (Q)	amine oxidase-related protein
NP-051635.1	COG3385 (L)	transposase, putative
NP-051639.1	COG1780 (F)	ribonucleotide reductase, NrdI family
NP-051640.1	COG0209 (F)	ribonucleoside-diphosphate reductase, alpha chain
NP-051641.1	COG0208 (F)	ribonucleoside-diphosphate reductase, beta chain
NP-051642.1	COG0526 (OC)	thioredoxin-related protein
NP-051643.1	COG0584 (C)	glycerophosphoryl diester phosphodiesterase
NP-284939.1	COG3385 (L)	putative transposase [Deinococcus radiodurans]
NP-051645.1	COG3385 (L)	transposase
NP-051650.1	COG0025 (P)	Na+/H+ antiporter, putative
NP-051653.1	COG0609 (P)	iron ABC transporter, permease protein
NP-051654.1	COG2375 (P)	iron-chelator utilization protein, putative
NP-051656.1	COG1309 (K)	transcriptional regulator, TetR family
NP-051662.1	COG3385 (L)	transposase
NP-051666.1	COG0553 (KL)	ATP-dependent helicase HepA
NP-051667.1	COG1002 (V)	conserved hypothetical protein
NP-051668.1	COG1002 (V)	conserved hypothetical protein
NP-051670.1	COG4226 (S)	HicB-related protein
NP-051672.1	COG1401 (V)	McrB-related protein
NP-051673.1	COG4268 (V)	hypothetical protein
NP-051675.1	COG2124 (Q)	cytochrome P450-related protein
NP-051678.1	COG1961 (L)	resolvase
NP-051682.1	COG4928 (R)	hypothetical protein
NP-051687.1	COG1520 (S)	hypothetical protein
NP-277100.1	COG4644 (L)	putative transposase
NP-051698.1	COG3436 (L)	transposase, putative
NP-051701.1	COG3293 (L)	transposase, putative
NP-051704.1	COG1032 (C)	oxidative cyclase, putative
NP-051705.1	COG2192 (O)	nodulation protein-related protein
NP-051706.1	COG0312 (R)	conserved hypothetical protein
NP-051708.1	COG0477 (GEPR)	multidrug-efflux transporter, putative

A.10. *D. radiodurans*-spezifische Proteine (COG-Gruppe K)

Genname, COG, (COG-Gruppe), Annotation

NP-293781.1	COG2378 (K)	repressor protein, putative
NP-293904.1	COG1595 (K)	RNA polymerase sigma-E factor
NP-293924.1	COG1522 (K)	transcriptional regulator, Lrp/AsnC family
NP-293957.1	COG0640 (K)	transcriptional regulator, ArsR family
NP-293988.1	COG2188 (K)	histidine utilization repressor HutC, putative
NP-294443.1	COG1309 (K)	transcriptional regulator, TetR family
NP-294528.1	COG1595 (K)	sigma factor, putative
NP-294727.1	COG1309 (K)	transcriptional regulator
NP-294753.1	COG1396 (K)	Transcription regulator
NP-294766.1	COG1695 (K)	Pex-related protein
NP-294835.1	COG0789 (K)	transcriptional regulator, MerR family
NP-294883.1	COG1846 (K)	transcriptional regulator, MarR family

NP-294941.1	COG1414 (K)	transcriptional regulator, IclR family
NP-295102.1	COG1309 (K)	transcriptional repressor, TetR family
NP-295406.1	COG2808 (K)	transcriptional regulator, putative
NP-295617.1	COG1522 (K)	transcriptional regulator, AsnC family
NP-295658.1	COG2390 (K)	transcriptional regulator
NP-295980.1	COG1476 (K)	transcriptional regulator, HTH
NP-296174.1	COG1476 (K)	transcriptional regulator, PbsX family
NP-296239.1	COG0789 (K)	transcriptional regulator, MerR family
NP-296294.1	COG1396 (K)	transcriptional regulator, HTH
NP-285466.1	COG1309 (K)	transcriptional regulator, TetR family
NP-285534.1	COG2188 (K)	transcriptional regulator, GntR family
NP-285575.1	COG0789 (K)	transcriptional regulator, MerR family
NP-051656.1	COG1309 (K)	transcriptional regulator, TetR family

A.11. *D. radiodurans*-spezifische Proteine (COG-Gruppe V)

Genname, COG, (COG-Gruppe), Annotation

NP-293833.1	COG0534 (V)	conserved hypothetical protein
NP-293863.1	COG2810 (V)	hypothetical protein
NP-294178.1	COG3570 (V)	streptomycin 3-kinase
NP-294231.1	COG1715 (V)	mrr restriction system protein
NP-294310.1	COG1715 (V)	mrr restriction system protein
NP-294322.1	COG2746 (V)	aminoglycoside N3-acetyltransferase
NP-294484.1	COG1680 (V)	hypothetical protein
NP-294516.1	COG0534 (V)	conserved hypothetical protein
NP-295036.1	COG1403 (V)	conserved hypothetical protein
NP-295256.1	COG3440 (V)	conserved hypothetical protein
NP-295449.1	COG3183 (V)	hypothetical protein
NP-295600.1	COG1787 (V)	conserved hypothetical protein
NP-295723.1	COG1619 (V)	mccF protein, putative
NP-295757.1	COG2746 (V)	aminoglycoside N3‘-acetyltransferase, type IV
NP-295988.1	COG1002 (V)	DNA modification methyltransferase-related protein
NP-285332.1	COG0842 (V)	conserved hypothetical protein
NP-285443.1	COG1002 (V)	conserved hypothetical protein
NP-285539.1	COG1680 (V)	esterase, putative
NP-285589.1	COG1401 (V)	hypothetical protein
NP-285602.1	COG0577 (V)	hypothetical protein
NP-051589.1	COG0577 (V)	hypothetical protein
NP-051667.1	COG1002 (V)	conserved hypothetical protein
NP-051668.1	COG1002 (V)	conserved hypothetical protein
NP-051672.1	COG1401 (V)	McrB-related protein
NP-051673.1	COG4268 (V)	hypothetical protein

A.12. Gecoggte *T. thermophilus*-spezifische Proteine

Genname, COG (COG-Gruppe), Annotation

TTC0007	COG0655 (R)	hypothetical conserved protein
TTC0008	COG1214 (O)	O-sialoglycoprotein endopeptidase
TTC0012	COG2414 (C)	probable tungsten-containing aldehyde ferredoxin oxidoreductase
TTC0013	COG1721 (R)	hypothetical protein
TTC0015	COG0526 (OC)	hypothetical protein
TTC0022	COG1145 (C)	hypothetical protein
TTC0033	COG3255 (I)	hypothetical conserved protein
TTC0041	COG0863 (L)	methyltransferase

```
TTC0053 COG0457 (R)     hypothetical conserved protein
TTC0056 COG0611 (H)     thiamin-monophosphate kinase
TTC0077 COG0594 (J)     ribonuclease P protein subunit
TTC0082 COG0683 (E)     leucine-, isoleucine-, valine-, threonine-, and alanine-binding protein
TTC0091 COG1051 (F)     hypothetical conserved protein
TTC0094 COG3557 (J)     probable cysteinyl-tRNA synthetase
TTC0105 COG2104 (H)     hypothetical protein
TTC0108 COG0500 (QR)    methyltransferase
TTC0113 COG3744 (S)     hypothetical conserved protein
TTC0114 COG4118 (D)     hypothetical protein
TTC0117 COG0498 (E)     threonine synthase
TTC0118 COG0477 (GEPR)  hypothetical protein
TTC0125 COG3742 (S)     hypothetical conserved protein
TTC0128 COG0277 (C)     (S)-2-hydroxy-acid oxidase subunit glcE
TTC0136 COG2152 (G)     hypothetical cytosolic protein
TTC0139 COG3329 (R)     sodium bicarbonate cotransporter
TTC0141 COG0586 (S)     hypothetical protein
TTC0142 COG0861 (P)     hypothetical membrane spanning protein
TTC0154 COG0437 (C)     nrfC protein
TTC0163 COG0697 (GER)   hypothetical membrane spanning protein
TTC0167 COG2236 (R)     purine phosphoribosyltransferase
TTC0171 COG1568 (R)     hypothetical conserved protein
TTC0178 COG1189 (J)     hemolysin
TTC0186 COG2426 (S)     hypothetical conserved protein
TTC0194 COG0005 (F)     purine-nucleoside phosphorylase
TTC0196 COG3599 (D)     hypothetical conserved protein
TTC0197 COG1434 (S)     hypothetical conserved protein
TTC0214 COG4475 (S)     transcriptional regulator
TTC0224 COG0526 (OC)    thiol-disulfide isomerase and thioredoxins
TTC0225 COG4663 (Q)     probable mannitol-binding protein with frameshift
TTC0226 COG0153 (G)     galactokinase
TTC0228 COG3044 (R)     isopentenyl-diphosphate delta-isomerase
TTC0229 COG3263 (P)     hypothetical membrane spanning protein
TTC0242 COG0502 (H)     biotin synthase
TTC0244 COG2519 (J)     tRNA (adenine-N(1)-)-methyltransferase
TTC0248 COG0607 (P)     hypothetical conserved protein
TTC0265 COG5405 (O)     ATP-dependent protease hslV
TTC0266 COG1708 (R)     hypothetical protein
TTC0267 COG2250 (S)     hypothetical conserved protein
TTC0279 COG0463 (M)     probable glycosyl transferase
TTC0280 COG0438 (M)     lipopolysaccharideN-acetylglucosaminyltransferase
TTC0282 COG0367 (E)     asparagine synthetase [glutamine-hydrolyzing]
TTC0288 COG0677 (M)     UDP-N-acetyl-D-mannosamine 6-dehydrogenase
TTC0300 COG0679 (R)     transporter
TTC0306 COG2141 (C)     putative oxidoreductase
TTC0311 COG1648 (H)     uroporphyrin-III C-methyltransferase (EC 2.1.1.107) /
                        precorrin-2 oxidase (EC 1.-.-.-) / ferrochelatase
TTC0312 COG1587 (H)     uroporphyrinogen-III synthase
TTC0318 COG0665 (E)     putative oxidoreductase
TTC0328 COG1653 (G)     gucose-binding protein
TTC0365 COG0725 (P)     molybdate-binding protein
TTC0375 COG0599 (S)     hypothetical conserved protein
TTC0381 COG1846 (K)     transcriptional regulator, marR family
TTC0385 COG0329 (EM)    dihydrodipicolinate synthase
TTC0401 COG0365 (I)     hypothetical protein
TTC0412 COG0235 (G)     L-ribulose-phosphate 4-epimerase
TTC0414 COG4663 (Q)     mannitol-binding protein
TTC0415 COG4665 (Q)     tripartite transporter, small subunit
TTC0419 COG0600 (P)     ABC transporter permease protein
TTC0430 COG0730 (R)     putative permease
TTC0432 COG0312 (R)     pmbA protein
TTC0433 COG0312 (R)     tldD protein
TTC0472 COG0421 (E)     spermine synthase
TTC0473 COG1586 (E)     S-adenosylmethionine decarboxylase proenzyme
TTC0482 COG0648 (L)     endonuclease IV
TTC0486 COG0526 (OC)    glutaredoxin-like protein
TTC0489 COG2719 (S)     stage V sporulation protein R
TTC0490 COG2718 (S)     glycosyl transferase
TTC0491 COG2766 (T)     prkA protein
TTC0494 COG0491 (R)     putative Zn-dependent hydrolase (beta-lactamase superfamily)
TTC0496 COG2137 (R)     regulatory protein recX
TTC0512 COG1610 (S)     hypothetical protein
TTC0514 COG4401 (E)     chorismate mutase
TTC0539 COG0589 (T)     hypothetical conserved protein
TTC0541 COG3809 (S)     hypothetical conserved protein
TTC0542 COG1694 (R)     hypothetical cytosolic protein
TTC0547 COG0388 (R)     hypothetical conserved protein
TTC0548 COG1268 (R)     probable bioY protein
TTC0551 COG0858 (J)     hypothetical conserved protein
TTC0555 COG1355 (R)     hypothetical conserved protein
TTC0563 COG3263 (P)     Na(+)/H(+) antiporter
TTC0566 COG0517 (R)     hypothetical protein
TTC0567 COG4636 (S)     hypothetical cytosolic protein
TTC0582 COG3239 (I)     omega-(3) fatty acid desaturase
TTC0589 COG3769 (R)     mannosyl-3-phosphoglycerate phosphatase
TTC0591 COG0329 (EM)    dihydrodipicolinate synthase
TTC0615 COG1653 (G)     sugar-binding protein
TTC0616 COG1980 (G)     dihydroorotate dehydrogenase
TTC0618 COG2193 (P)     bacterioferritin
TTC0620 COG0379 (H)     quinolinate synthetase A
TTC0621 COG0157 (H)     nicotinate-nucleotide pyrophosphorylase [carboxylating]
TTC0635 COG2062 (T)     hypothetical conserved protein
```

TTC0650	COG0842	(V)	probable transmembrane protein
TTC0675	COG2050	(Q)	hypothetical conserved protein
TTC0686	COG4864	(S)	hypothetical conserved protein
TTC0687	COG1030	(O)	putative protease / transporter
TTC0695	COG2239	(P)	Mg2+ transporter mgtE
TTC0707	COG0182	(J)	translation initiation factor aIF-2B alpha subunit
TTC0726	COG1839	(S)	hypothetical conserved protein
TTC0731	COG1351	(F)	thy1 protein
TTC0750	COG0861	(P)	hypothetical membrane spanning protein
TTC0751	COG0500	(QR)	methyltransferfase
TTC0769	COG1622	(C)	subunit II of C(O/B)3-type cytochrome c oxidase
TTC0797	COG1752	(R)	conserved hypothetical protein
TTC0809	COG0168	(P)	trk system potassium uptake protein trkG
TTC0815	COG3001	(G)	phosphatidylserine decarboxylase
TTC0821	COG1764	(O)	hypothetical protein
TTC0824	COG1502	(I)	putative phosphoslipase
TTC0826	COG1792	(M)	rod shape-determining protein mreC
TTC0837	COG4636	(S)	conserved hypothetical protein
TTC0859	COG1196	(D)	hypothetical protein
TTC0868	COG4636	(S)	conserved hypothetical protein
TTC0869	COG4636	(S)	conserved hypothetical protein
TTC0877	COG0772	(D)	rod shape-determining protein rodA
TTC0880	COG1733	(K)	conserved hypothetical protein
TTC0892	COG0393	(S)	conserved hypothetical protein
TTC0895	COG2114	(T)	protein with frame shift
TTC0911	COG0636	(C)	V-type sodium ATP synthase subunit K Homolog
TTC0924	COG5438	(S)	hypothetical membrane spanning protein
TTC0935	COG1079	(R)	putative transport system permease ABC transporter protein
TTC0936	COG4603	(R)	protein with frame shift (permease)
TTC0943	COG3369	(S)	hypothetical conserved protein
TTC0950	COG1994	(R)	conserved hypothetical protein
TTC0955	COG0456	(R)	hypothetical protein
TTC0963	COG2258	(S)	hypothetical protein
TTC0982	COG1515	(L)	endonuclease V
TTC0995	COG2928	(S)	transporter
TTC0999	COG4127	(S)	hypothetical conserved protein
TTC1004	COG1752	(R)	serine protease
TTC1045	COG3258	(C)	cytochrome c
TTC1047	COG0446	(R)	sulfide dehydrogenase /flavocytochrome c)
TTC1057	COG0526	(OC)	thioredoxin
TTC1058	COG2010	(C)	cytochrome c-552 precursor
TTC1066	COG3360	(S)	hypothetical protein
TTC1067	COG5502	(S)	hypothetical protein
TTC1083	COG0427	(C)	4-hydroxybutyrate coenzyme A transferase
TTC1085	COG1502	(I)	hypothetical membrane associated protein
TTC1093	COG1586	(E)	S-adenosylmethionine decarboxylase proenzyme
TTC1094	COG1470	(S)	ABC transporter substrate-binding protein
TTC1096	COG1277	(R)	ABC transporter permease protein
TTC1097	COG1917	(S)	conserved hypothetical protein
TTC1098	COG1926	(R)	putative phosphoribosyl transferase
TTC1103	COG2607	(R)	fragment of ATP-dependent protease subunit
TTC1105	COG0840	(NT)	conserved hypothetical protein
TTC1107	COG0741	(M)	conserved hypothetical protein
TTC1109	COG0517	(R)	conserved hypothetical protein
TTC1110	COG0705	(R)	hypothetical membrane spanning protein
TTC1111	COG0616	(OU)	periplasmic serine protease
TTC1112	COG0475	(P)	hypothetical protein
TTC1119	COG2129	(R)	heat-stable protein
TTC1131	COG1714	(S)	hypothetical membrane spanning protein
TTC1141	COG0755	(O)	hypothetical protein
TTC1146	COG2353	(S)	hypothetical protein
TTC1147	COG3485	(Q)	conserved hypothetical protein
TTC1150	COG4636	(S)	hypothetical cytosolic protein
TTC1162	COG0119	(E)	hydroxymethylglutaryl-CoA lyase
TTC1166	COG0727	(R)	hypothetical protein
TTC1169	COG3547	(L)	probable DNA-binding protein
TTC1170	COG1518	(L)	hypothetical protein
TTC1173	COG0665	(E)	fragment of putative oxidoreductase
TTC1175	COG0665	(E)	fragment of putative oxidoreductase
TTC1178	COG0596	(R)	hypothetical protein
TTC1179	COG0596	(R)	hypothetical protein
TTC1189	COG4274	(S)	hypothetical protein
TTC1199	COG0350	(L)	putative methylated-DNA--protein-cysteine methyltransferase
TTC1205	COG1899	(O)	deoxyhypusine synthase
TTC1207	COG1569	(R)	hypothetical protein
TTC1220	COG0827	(L)	modification methylase
TTC1228	COG0642	(T)	two component system histidine kinase
TTC1240	COG0374	(C)	conserved hypothetical protein
TTC1244	COG1430	(S)	conserved hypothetical protein
TTC1245	COG1430	(S)	hypothetical protein
TTC1251	COG1259	(S)	conserved hypothetical protein
TTC1252	COG1409	(R)	acid phosphatase
TTC1258	COG2897	(P)	thiosulfate sulfurtransferase
TTC1262	COG3374	(S)	conserved hypothetical protein
TTC1264	COG1840	(P)	iron(III)-binding protein
TTC1295	COG0457	(R)	hypothetical protein
TTC1297	COG0697	(GER)	transporter
TTC1310	COG1841	(J)	LSU ribosomal protein L30P
TTC1320	COG0255	(J)	50S ribosomal protein L29
TTC1335	COG2606	(S)	ebsC protein
TTC1341	COG1647	(R)	carboxylesterase
TTC1343	COG1489	(R)	sugar fermentation stimulation protein

```
TTC1344 COG1835 (I)     hypothetical protein
TTC1347 COG1180 (O)     pyruvate formate-lyase activating enzyme
TTC1348 COG1355 (R)     hypothetical conserved protein
TTC1350 COG0038 (P)     hypothetical protein
TTC1352 COG3253 (S)     hypothetical cytosolic protein Cluster function: Polyketide synthase curD
TTC1359 COG3462 (S)     hypothetical protein
TTC1365 COG5428 (S)     hypothetical conserved protein
TTC1369 COG3439 (S)     hypothetical conserved protein
TTC1370 COG2132 (Q)     laccase
TTC1372 COG3462 (S)     hypothetical protein
TTC1374 COG0822 (C)     IscU protein
TTC1383 COG1371 (S)     hypothetical protein              Cluster function: PHOSPHORIBOSYLFORMYLGLYCINAMIDINE SYNTHASE
TTC1391 COG0438 (M)     conserved hypothetical protein
TTC1394 COG1598 (S)     hypothetical protein
TTC1395 COG1724 (N)     hypothetical conserved protein
TTC1399 COG0446 (R)     sulfide dehydrogenase (flavocytochrome C)
TTC1400 COG1138 (O)     hypothetical protein
TTC1401 COG2010 (C)     high-affinity iron permease
TTC1403 COG5557 (C)     polysulfide reductase chain CHAIN C
TTC1404 COG0437 (C)     anaerobic dimethyl sulfoxide reductase chain B
TTC1410 COG1664 (M)     hypothetical cytosolic protein
TTC1413 COG4636 (S)     hypothetical conserved protein
TTC1421 COG1057 (H)     nicotinate-nucleotide adenylyltransferase
TTC1425 COG1314 (U)     hypothetical protein
TTC1428 COG0286 (V)     hypothetical protein
TTC1430 COG0433 (R)     hypothetical protein
TTC1432 COG3328 (L)     transposase
TTC1441 COG0477 (GEPR)  transporter
TTC1443 COG1803 (G)     methylglyoxal synthase
TTC1444 COG1051 (F)     hypothetical protein
TTC1445 COG0301 (H)     thiamine biosynthesis protein thiI
TTC1447 COG1055 (P)     transporter
TTC1449 COG2234 (R)     putative aminopeptidase
TTC1451 COG2154 (H)     pterin-4-alpha-carbinolamine dehydratase
TTC1459 COG0235 (G)     L-fuculose phosphate aldolase
TTC1495 COG0465 (O)     hypothetical protein
TTC1496 COG0457 (R)     tetratricopeptide repeat family protein
TTC1507 COG1853 (R)     flavoredoxin
TTC1513 COG1653 (G)     sugar-binding protein
TTC1515 COG3327 (K)     phaN
TTC1520 COG3342 (S)     major pilin protein fimA
TTC1540 COG1543 (S)     conserved hypothetical protein
TTC1549 COG1724 (N)     hypothetical protein
TTC1571 COG1082 (G)     hypothetical conserved protein
TTC1581 COG0393 (S)     hypothetical conserved protein
TTC1591 COG0674 (C)     2-oxoacid-ferredoxin oxidoreductase
TTC1592 COG1013 (C)     2-oxoacid-ferredoxin oxidoreductase
TTC1593 COG0665 (E)     hypothetical protein
TTC1599 COG1405 (K)     hypothetical protein
TTC1600 COG0221 (C)     Inorganic pyrophosphatase
TTC1633 COG0589 (T)     hypothetical protein
TTC1634 COG0517 (R)     inosine-5'-monophosphate dehydrogenase
TTC1644 COG0457 (R)     hypothetical protein
TTC1645 COG1912 (S)     hypothetical conserved protein
TTC1649 COG1993 (S)     hypothetical conserved protein
TTC1655 COG1335 (Q)     putative isochorismatase
TTC1656 COG0685 (E)     methylenetetrahydrofolate reductase
TTC1704 COG1598 (S)     hypothetical protein
TTC1707 COG2759 (F)     formate--tetrahydrofolate ligase
TTC1711 COG1042 (C)     6-carboxyhexanoate--CoA ligase
TTC1726 COG0226 (P)     phosphate-binding protein
TTC1728 COG3872 (R)     hypothetical conserved protein
TTC1730 COG0735 (P)     ferric uptake regulation protein
TTC1778 COG1282 (C)     NAD(P) transhydrogenase subunit beta
TTC1779 COG3288 (C)     NAD(P) transhydrogenase subunit alpha 2
TTC1781 COG0697 (GER)   permease
TTC1783 COG0477 (GEPR)  transporter
TTC1784 COG3177 (S)     hypothetical conserved protein
TTC1790 COG4663 (Q)     C4-dicarboxylate-binding protein
TTC1792 COG4665 (Q)     tripartite transporter, small subunit
TTC1794 COG0589 (T)     universal stress protein family
TTC1803 COG0212 (H)     5-formyltetrahydrofolate cyclo-ligase
TTC1804 COG4113 (R)     hypothetical conserved protein
TTC1826 COG1341 (R)     hypothetical protein
TTC1828 COG1449 (G)     amylopullulanase
TTC1830 COG1846 (K)     transcriptional regulators, marR/emrR family
TTC1834 COG2414 (C)     tungsten-containing aldehyde ferredoxin oxidoreductase
TTC1835 COG1977 (H)     molybdopterin converting factor, small subunit
TTC1837 COG3307 (M)     hypothetical membrane spanning protein
TTC1864 COG0717 (F)     deoxycytidine triphosphate deaminase
TTC1872 COG3546 (P)     pseudocatalase
TTC1884 COG1896 (R)     hydrolase (had superfamily)
TTC1887 COG4636 (S)     hypothetical conserved protein
TTC1901 COG2070 (R)     2-nitropropane dioxygenase
TTC1911 COG0839 (C)     NADH-quinone oxidoreductase chain CHAIN J
TTC1927 COG5551 (S)     hypothetical conserved protein
TTC1935 COG4662 (H)     tungstate transport system permease protein
TTC1936 COG2998 (H)     tungstate-binding protein precursor
TTC1938 COG4448 (E)     L-asparaginase II
TTC1941 COG0477 (GEPR)  multidrug resistance protein, putative
TTC1948 COG1420 (K)     heat-inducible transcription repressor hrcA
TTC1961 COG3400 (S)     zrk system potassium uptake protein trkA
```

```
TTC1962 COG0475 (P)     trk system potassium uptake protein trkA
TTC1965 COG0739 (M)     membrane proteins related to metalloendopeptidases
TTC1968 COG0477 (GEPR)  probable macrolide-efflux transmembrane protein
TTC1970 COG3369 (S)     hypothetical conserved protein
TTC1977 COG0778 (C)     glucose-1-phosphate adenylyltransferase
TTC1981 COG0457 (R)     tetratricopeptide repeat family protein
TTP0007 COG3376 (P)     hypothetical conserved protein
TTP0008 COG1903 (H)     cobalamin biosynthetic protein cbiD
TTP0009 COG2082 (H)     precorrin-8X methylmutase
TTP0010 COG2242 (H)     precorrin-6Y C5,15-methyltransferase [decarboxylating]
TTP0011 COG2243 (H)     precorrin-2 C20-methyltransferase
TTP0014 COG2073 (H)     cbiG protein
TTP0015 COG2138 (S)     cbiX protein
TTP0018 COG0778 (C)     nicotinate-nucleotide--dimethylbenzimidazolephosphoribosyltransferase
TTP0020 COG2250 (S)     hypothetical conserved protein
TTP0021 COG1708 (R)     hypothetical conserved protein
TTP0029 COG1082 (G)     hypothetical conserved protein
TTP0030 COG0800 (G)     putative KHG/KDPG aldolase
TTP0031 COG1638 (G)     ABC transporter, periplasmic substrate-binding protein, putative
TTP0032 COG3090 (G)     probable transporter
TTP0035 COG2055 (C)     dehydrogenase
TTP0042 COG2723 (G)     beta-glycosidase
TTP0048 COG0702 (MG)    nucleoside-diphosphate-sugar epimerase
TTP0050 COG1853 (R)     hypothetical conserved protein
TTP0056 COG0789 (K)     probable transcriptional regulator
TTP0058 COG0415 (L)     deoxyribodipyrimidine photolyase
TTP0068 COG1708 (R)     hypothetical conserved protein
TTP0069 COG1683 (S)     hypothetical conserved protein
TTP0071 COG1085 (C)     galactose-1-phosphate uridylyltransferase
TTP0072 COG3345 (G)     alpha-galactosidase
TTP0075 COG0154 (J)     hypothetical protein
TTP0081 COG1373 (R)     probable ATPase
TTP0086 COG0243 (C)     hypothetical protein
TTP0097 COG3744 (S)     hypothetical conserved protein
TTP0101 COG1343 (L)     hypothetical protein
TTP0102 COG1353 (R)     putative hydrolase of the HD superfamily (permuted catalytic motifs)
TTP0104 COG1337 (L)     hypothetical conserved protein
TTP0106 COG1332 (L)     hypothetical protein
TTP0115 COG1353 (R)     hypothetical protein
TTP0117 COG1367 (L)     hypothetical conserved protein
TTP0118 COG1336 (L)     hypothetical conserved protein
TTP0120 COG1604 (L)     hypothetical protein
TTP0122 COG2414 (C)     probable tungsten-containing aldehyde ferredoxin oxidoreductase
TTP0123 COG1266 (R)     hypothetical membrane spanning protein
TTP0124 COG1110 (L)     hypothetical protein
TTP0132 COG1203 (R)     hypothetical conserved protein
TTP0136 COG1468 (L)     hypothetical conserved protein
TTP0140 COG1373 (R)     hypothetical conserved protein
TTP0144 COG1178 (P)     hypothetical protein
TTP0157 COG5433 (L)     hypothetical protein
TTP0161 COG0208 (F)     ribonucleoside-diphosphate reductase beta chain
TTP0173 COG3488 (C)     hypothetical conserved protein
TTP0182 COG3328 (L)     transposase
TTP0184 COG2002 (K)     hypothetical protein
TTP0186 COG2963 (L)     hypothetical conserved protein
TTP0189 COG4993 (G)     hypothetical protein
TTP0191 COG0210 (L)     putative helicase superfamily I
TTP0195 COG1343 (L)     hypothetical conserved protein
TTP0196 COG1518 (L)     hypothetical conserved protein
TTP0197 COG1468 (L)     hypothetical conserved protein
TTP0198 COG2378 (K)     hypothetical protein
TTP0199 COG0515 (RTKL)  hypothetical protein
TTP0202 COG1203 (R)     ATP-dependent helicase with frameshift
TTP0205 COG3547 (L)     probable DNA-binding protein
TTP0211 COG1373 (R)     ATPase
TTP0212 COG2414 (C)     hypothetical protein
TTP0213 COG3547 (L)     transposase
TTP0215 COG1518 (L)     hypothetical protein
TTP0218 COG3394 (S)     hypothetical conserved protein
TTP0219 COG2211 (G)     hypothetical conserved protein
TTP0220 COG3250 (G)     probable beta-galactosidase protein
TTP0221 COG1501 (G)     alpha-glucosidase
TTP0222 COG3250 (G)     probable beta-galactosidase
TTP0223 COG2211 (G)     hypothetical conserved protein
```

A.13. Fremdgene in *E. blattae*

Genname, COG, (COG-Gruppe), Annotation

```
REB003928 COG2932 (K)   regulatory protein cI
REB004723 NOCOG         unnamed protein product
REB004724 NOCOG         phage-related transcriptional phage related transcriptional regulator
REB004725 COG2183 (K)   Hypothetical Protein
REB004726 NOCOG         O (DNA replication;299)
REB000238 NOCOG         DNA replication protein
REB000240 COG0270 (L)   cytosine methylase
REB000241 NOCOG         Hypothetical Protein
REB004665 NOCOG         hypothetical protein HK022p34
REB004727 NOCOG         conserved hypothetical protein
REB000244 NOCOG         hypothetical phage-related protein
REB004728 NOCOG         hypothetical protein
REB000245 NOCOG         hypothetical protein sb49
REB004729 NOCOG         Hypothetical Protein
REB000248 NOCOG         Rz lysis protein
REB004730 NOCOG         hypothetical protein sb54
REB004732 NOCOG         Orf-129
REB004733 COG0358 (L)   Hypothetical Protein
REB004734 NOCOG         terminase small subunit
REB004735 NOCOG         gp2
REB002881 COG2801 (L)   IS911 orfA
REB004736 COG2963 (L)   Transposase
REB003572 NOCOG         DNA stabilization protein
REB002876 NOCOG         packaged DNA stabilization protein
REB002875 COG1196 (D)   head completion protein
REB002874 NOCOG         head assembly protein
REB002873 NOCOG         DNA transfer protein
REB002872 NOCOG         DNA transfer protein
REB002871 COG0741 (M)   DNA transfer protein
REB002870 NOCOG         tail spike protein
REB003573 COG3274 (S)   hypothetical membrane protein
REB002869 COG0389 (L)   DNA polymerase V subunit
REB002868 COG1961 (L)   putative recombinase
REB004233 COG5302 (R)   putative antitoxin of gyrase inhibiting toxin-antitoxin system
REB002823 NOCOG         Putative toxin of gyrase inhibiting toxin-antitoxin system
REB003580 NOCOG         conserved hypothetical protein
REB003585 NOCOG         hypothetical protein
REB003586 NOCOG         hypothetical protein
REB003587 NOCOG         hypothetical protein
REB003588 COG0454 (KR)  possible acetyltransferase
REB003589 NOCOG         hypothetical protein
REB004235 NOCOG         resolvase
REB003590 COG1598 (S)   conserved hypothetical protein
REB004740 NOCOG         hypothetical protein
REB003591 NOCOG         hypothetical protein
REB003592 NOCOG         hypothetical protein
REB003593 NOCOG         hypothetical protein
REB003595 NOCOG         putative regulator of cell division encoded by prophage CP-933O
REB002807 NOCOG         unknown protein encoded by cryptic prophage CP-933M
REB003597 COG0270 (L)   DNA-cytosine methyltransferase (EC 2.1.1.73)
REB003598 NOCOG         hypothetical histidine kinase
REB003599 NOCOG         conserved hypothetical protein
REB003600 NOCOG         hypothetical protein
REB001985 NOCOG         conserved hypothetical protein (prohage related)
REB001986 COG0863 (L)   putative DNA methylase
REB003601 COG2801 (L)   Transposase
REB004742 COG2801 (L)   ISSod2, transposase OrfA
REB003603 NOCOG         hypothetical protein
REB003604 COG0863 (L)   putative DNA adenine methyltransferase encoded by prophage CP-933O
REB003605 NOCOG         hypothetical protein
REB003114 COG1598 (S)   conserved hypothetical protein (phage related)
REB003115 NOCOG         putative bacteriophage protein
REB003116 COG3179 (R)   lytic enzyme (putative glycohydrolase)
REB003117 NOCOG         Gifsy-1 prophage protein
REB003118 NOCOG         hypothetical protein
REB003607 NOCOG         hypothetical protein
REB003125 NOCOG         gp7
REB003126 NOCOG         conserved hypothetical protein (phage related)
REB003608 NOCOG         hypothetical protein
REB003128 NOCOG         conserved hypothetical protein (phage related)
REB003129 NOCOG         conserved hypothetical protein (phage related)
REB003130 NOCOG         conserved hypothetical protein (phage related)
REB003131 COG5281 (S)   tape measure protein
REB003611 COG4718 (S)   Gp17
REB003133 COG1310 (R)   gp19
REB003134 NOCOG         putative Chromosome segregation ATPase
REB003612 NOCOG         hypothetical protein
REB003613 COG3680 (S)   conserved hypothetical protein
REB003135 COG4723 (S)   putative tail component protein
REB003614 NOCOG         hypothetical protein
REB003136 COG4723 (S)   gp20
REB003615 NOCOG         hypothetical protein
REB003616 NOCOG         hypothetical protein
REB003617 NOCOG         hypothetical protein
REB003213 COG2731 (G)   conserved hypothetical protein
REB003626 NOCOG         conserved hypothetical protein (Yersinia)
REB003214 COG1396 (K)   transcriptional regulator
REB003215 COG3550 (R)   conserved hypothetical protein
REB003216 COG3568 (R)   conserved hypothetical protein
REB003218 COG3204 (S)   hypothetical membrane protein
REB003627 NOCOG         hypothetical protein
REB001420 COG0845 (M)   putative Chromosome segregation ATPase
```

```
REB001421 COG4618 (R)   ABC-type transporter, ATPase component
REB001422 COG2931 (Q)   secreted hemolysin-type calcium-binding bacteriocin, putative
REB004783 NOCOG         hypothetical protein
REB001423 COG1278 (K)   Cold shock-like protein cspH
REB004214 COG3448 (T)   putative membrane protein
REB001425 COG0679 (R)   putative permease
REB004784 NOCOG         hypothetical protein
REB002127 NOCOG         hypothetical phosphatase
REB004303 NOCOG         hypothetical protein
REB002128 COG1596 (M)   CAPSULE POLYSACCHARIDE EXPORT PROTEIN BEXD
REB004304 COG0394 (T)   putative acid phosphatase Wzb
REB002130 COG3206 (M)   putative ATPase
REB002131 COG2148 (M)   Undecaprenyl-phosphate galactosephosphotransferase (EC 2.7.8.6)
REB002132 COG0438 (M)   putative glycosyltransferase
REB002133 COG0438 (M)   putative glycosyltransferase
REB004810 COG2981 (E)   hypothetical protein
REB004817 NOCOG         hypothetical protein
REB003714 NOCOG         hypothetical protein
REB003715 COG3518 (S)   conserved hypothetical protein
REB003716 NOCOG         hypothetical protein
REB003717 NOCOG         hypothetical protein
REB003718 NOCOG         conserved hypothetical protein
REB004818 NOCOG         hypothetical protein
REB001797 COG1134 (GM)  O-antigen export system ATP-binding protein rfbB
REB004820 COG1682 (GM)  O-antigen export system permease protein
REB004116 COG0562 (M)   Probable UDP-galactopyranose mutase (EC 5.4.99.9)
REB000536 COG1216 (R)   putative glycosyl transferase
REB000535 COG0438 (M)   galactosyl transferase
REB004822 NOCOG         hypothetical protein
REB004823 COG1835 (I)   conserved hypothetical protein
REB000531 COG0463 (M)   putative glycosyltransferase
REB000530 COG0463 (M)   putative glycosyltransferase
REB000529 COG0463 (M)   putative glycosyltransferase
REB004824 NOCOG         hypothetical protein
REB000528 COG1835 (I)   hypothetical acyltransferase
REB003808 NOCOG         hypothetical protein
REB004829 COG2801 (L)   conserved hypothetical protein
REB004278 COG3547 (L)   putative transposase
REB004830 COG2801 (L)   Transposase
REB003796 COG2963 (L)   Transposase
REB002522 COG0684 (H)   4-hydroxy-4-methyl-2-oxoglutarate aldolase, putative
REB003795 NOCOG         hypothetical protein
REB002523 COG0477 (GEPR)        conserved hypothetical protein
REB003794 NOCOG         hypothetical protein
REB003793 COG2801 (L)   ISxac2 transposase
REB002525 COG2801 (L)   Transposase
REB002526 COG1292 (M)   secondary glycine betaine transporter BetU
REB002527 COG2801 (L)   putative transposase
REB004831 NOCOG         conserved hypothetical protein
REB004832 NOCOG         hypothetical protein
REB004833 COG2872 (R)   Predicted metal-dependent hydrolases related to alanyl-tRNA synthetase HxxxH domain
REB004834 COG0697 (GER) permease
REB004835 COG1396 (K)   putative transcriptional regulator
REB003877 COG3549 (R)   conserved hypothetical protein
REB003879 COG0642 (T)   putative tetrathionate reductase complex, sensory transduction histidine kinase
REB003880 COG4566 (T)   putative ATPase
REB002766 COG0437 (C)   Anaerobic dimethyl sulfoxide reductase chain B
REB003881 COG3302 (R)   Anaerobic dimethyl sulfoxide reductase chain C
REB002765 COG0243 (C)   Anaerobic dimethyl sulfoxide reductase chain A precursor (EC 1.8.99.-)
REB003882 COG3381 (R)   conserved hypothetical protein
REB002764 COG1737 (K)   transcriptional regulator
REB002763 COG1263 (G)   PTS enzyme (EC 2.7.1.69), maltose- and glucose-specific, factor II homologue
REB003883 NOCOG         hypothetical protein
REB004846 COG2963 (L)   insertion sequence IS2
REB002741 COG2801 (L)   Transposase
REB000481 COG3177 (S)   conserved hypothetical protein
REB000482 COG2801 (L)   conserved hypothetical protein
REB004847 COG2963 (L)   putative transposase
REB004375 NOCOG         phage tail assembly protein
REB003318 NOCOG         probable tail fiber protein GP37
REB003317 COG3778 (S)   Mu-like prophage FluMu protein gp48
REB003316 COG3299 (S)   hypothetical protein Mup47
REB003527 COG4384 (S)   prophage MuSo2, baseplate assembly protein V
REB003313 NOCOG         probable bacteriophage tail protein
REB003312 NOCOG         Mu-like prophage FluMu protein gp41
REB003311 NOCOG         Mu-like prophage FluMu tail tube protein
REB003310 COG4386 (R)   putative prophage MuSo2, tail sheath protein
REB003306 NOCOG         Mu-like prophage FluMu protein gp36
REB004376 NOCOG         hypothetical protein
REB003526 NOCOG         hypothetical protein
REB003305 NOCOG         putative prophage MuSo2, major head subunit
REB003304 COG4388 (R)   putative prophage MuSo2, protein Gp32
REB003303 COG2369 (S)   prophage MuSo2, F protein
REB003300 NOCOG         Mu-like prophage FluMu protein gp27
REB003525 NOCOG         hypothetical protein
REB003299 NOCOG         Mu-like phage gp26
REB003524 COG1734 (T)   phage protein gp34
REB003523 NOCOG         hypothetical protein
REB004377 NOCOG         hypothetical protein
REB003521 NOCOG         hypothetical protein
REB003293 NOCOG         conserved hypothetical protein
REB003292 NOCOG         conserved hypothetical protein
```

```
REB003520 NOCOG         hypothetical protein gp9
REB003291 NOCOG         host nuclease inhibitor protein
REB003519 NOCOG         Cim protein (Kil protein)
REB003518 COG3423 (K)   possible DNA-binding protein
REB003517 COG2932 (K)   putative transcriptional regulator
REB003516 NOCOG         glycerol dehydratase 21kDa subunit (partial)
REB003515 NOCOG         hypothetical protein
REB001884 COG1442 (M)   UDP-glucose:(glucosyl) LPS alpha1,3-glucosyltransferase WaaO
REB001883 COG1442 (M)   UDP-galactose:(glucosyl) LPS alpha1,2-galactosyltransferase WaaT
REB001882 NOCOG         Lipopolysaccharide core biosynthesis protein rfaY
REB001881 NOCOG         Lipopolysaccharide core biosynthesis protein rfaZ
REB001880 COG3306 (M)   putative beta1,4-galactosyltransferase WaaX
REB001879 COG3307 (M)   lipid A-core:surface polymer ligase WaaL
REB004870 NOCOG         hypothetical protein
REB004871 NOCOG         hypothetical protein
REB001258 COG0582 (L)   Putative integrase
REB001261 COG0732 (V)   subunit S of type I restriction-modification system
REB001262 COG1205 (R)   helicase related protein
REB004872 NOCOG         conserved hypothetical protein
REB004873 COG2801 (L)   Transposase
REB004874 COG2801 (L)   ISSod2, transposase OrfA
REB004875 NOCOG         hypothetical protein
REB004876 COG2116 (P)   Probable formate transporter 1
REB000951 COG3604 (KT)  NtrC family transcriptional regulator
REB000950 NOCOG         hypothetical formate hydrogenlyase maturation protein HycH
REB004880 NOCOG         hypothetical protein
REB001650 NOCOG         hypothetical protein
REB004118 NOCOG         hypothetical protein
REB004120 COG3710 (K)   putative transcriptional regulatory protein
REB004881 NOCOG         hypothetical protein
REB001651 NOCOG         conserved hypothetical protein
REB004882 NOCOG         hypothetical protein
REB003730 NOCOG         conserved hypothetical protein
REB003732 NOCOG         conserved hypothetical protein
REB003733 NOCOG         hypothetical protein
REB003734 NOCOG         hypothetical protein
REB003737 NOCOG         hypothetical protein
REB003738 NOCOG         conserved hypothetical protein
REB003739 NOCOG         hypothetical protein
REB003741 COG0457 (R)   hypothetical protein
REB003743 NOCOG         hypothetical protein
REB003744 COG3179 (R)   conserved hypothetical protein
REB003745 NOCOG         conserved hypothetical protein
REB003748 NOCOG         hypothetical protein
REB003749 NOCOG         conserved hypothetical protein
REB003750 NOCOG         hypothetical protein
REB003751 COG2801 (L)   ISxac2 transposase
REB004909 COG2801 (L)   ISSod2, transposase OrfA
REB003753 NOCOG         hypothetical protein
REB004910 NOCOG         hypothetical protein
REB003755 NOCOG         hypothetical protein
REB004911 NOCOG         hypothetical protein
REB003756 COG3755 (S)   conserved hypothetical protein
REB003757 NOCOG         conserved hypothetical protein
REB003761 NOCOG         hypothetical protein
REB004937 NOCOG         hypothetical protein
REB000176 COG1186 (J)   Bacterial Peptide Chain Release Factor 2 (RF-2)
REB000174 COG1943 (L)   Transposase for insertion sequence element IS200
REB004938 COG1802 (K)   putative transcriptional regulator
REB003634 NOCOG         conserved hypothetical protein
REB000808 COG0499 (H)   Adenosylhomocysteinase (EC 3.3.1.1)
REB000809 COG0477 (GEPR)        putative xanthosine permease
REB004939 NOCOG         hypothetical protein
REB000810 COG1609 (K)   Bacterial regulatory proteins, lacI family
REB000811 COG1957 (F)   Inosine-uridine preferring nucleoside hydrolase (EC 3.2.2.1)
REB000106 COG0746 (H)   Molybdopterin-guanine dinucleotide biosynthesis protein A
REB000107 COG1763 (H)   Molybdopterin-guanine dinucleotide biosynthesis protein B
REB003641 COG0464 (O)   putative AAA+ class ATPase
REB003640 NOCOG         putative Phage-like element PBSX protein xkdO
REB000480 COG2801 (L)   putative integrase
REB004950 COG2963 (L)   insertion sequence IS2
REB004951 NOCOG         hypothetical protein
REB004525 COG2378 (K)   putative Bacterial regulatory proteins, deoR family
REB004952 COG2801 (L)   transposase
REB003638 NOCOG         hypothetical protein
REB004953 NOCOG         hypothetical protein
REB004954 NOCOG         hypothetical protein
REB004648 NOCOG         conserved hypothetical protein
REB000426 COG0347 (E)   Nitrogen regulatory protein P-II
REB004955 NOCOG         hypothetical protein
REB003512 COG3126 (S)   glycoprotein/polysaccharide metabolism
REB004000 COG2801 (L)   IS1400 transposase B
REB004005 NOCOG         hypothetical protein
REB000376 COG1702 (T)   PhoH protein (phosphate starvation-inducible protein PsiH)
REB004006 NOCOG         hypothetical protein
REB004957 COG1802 (K)   transcriptional regulator
REB004958 NOCOG         conserved hypothetical protein
REB000171 COG2801 (L)   putative transposase
REB004959 COG3547 (L)   putative IS402 family transposase
REB004960 COG2801 (L)   IS911 orfB
REB004961 COG2963 (L)   transposase insN for insertion sequence element IS911
REB000167 COG0582 (L)   Prophage P4 integrase
```

REB004962	NOCOG	hypothetical protein
REB004173	COG0491 (R)	Hydroxyacylglutathione hydrolase (EC 3.1.2.6)
REB004174	COG0741 (M)	Membrane-bound lytic murein transglycosylase D precursor (EC 3.2.1.-)
REB004963	NOCOG	hypothetical protein
REB004966	COG0656 (R)	hypothetical 2,5-diketo-D-gluconate reductase B
REB000122	COG2244 (R)	Polysaccharide biosynthesis protein
REB000121	NOCOG	conserved hypothetical protein
REB000119	COG0438 (M)	putative glycosyltransferase
REB004690	NOCOG	hypothetical protein
REB000117	COG0836 (M)	Mannose-1-phosphate guanylyltransferase (EC 2.7.7.22)
REB000116	NOCOG	hypothetical protein
REB000115	NOCOG	hypothetical protein
REB004967	COG2801 (L)	Transposase
REB004016	NOCOG	conserved hypothetical protein
REB004017	NOCOG	conserved hypothetical protein
REB004018	NOCOG	hypothetical protein
REB004019	NOCOG	hypothetical protein
REB004020	NOCOG	hypothetical protein
REB004022	NOCOG	conserved hypothetical protein
REB004023	NOCOG	hypothetical protein
REB004024	NOCOG	hypothetical protein
REB000565	COG1238 (S)	conserved hypothetical protein
REB004588	NOCOG	hypothetical protein
REB000570	COG1846 (K)	putative transcriptional regulator
REB003497	COG1296 (E)	conserved hypothetical protein
REB004591	NOCOG	hypothetical protein
REB004969	NOCOG	hypothetical protein
REB000577	COG1780 (F)	NrdI protein
REB000578	NOCOG	conserved hypothetical protein
REB000581	COG2916 (R)	DNA-binding protein stpA (H-NS homolog stpA)
REB004978	NOCOG	terminase small subunit
REB004979	NOCOG	DNA packaging protein gp2
REB004980	COG3547 (L)	putative IS402 family transposase
REB000169	COG2801 (L)	IS911 orfB
REB004981	COG2963 (L)	Transposase
REB000990	COG2801 (L)	Transposase
REB004616	NOCOG	conserved hypothetical protein
REB004983	NOCOG	hypothetical protein
REB004463	NOCOG	hypothetical protein
REB000957	COG3547 (L)	Transposase
REB000956	COG2801 (L)	Transposase
REB003978	COG2801 (L)	Transposase
REB003909	COG2801 (L)	Transposase
REB004106	COG2801 (L)	Transposase
REB004107	NOCOG	putative adhesin
REB000026	COG1428 (F)	hypothethical deoxyguanosine/deoxyadenosine kinase
REB000025	COG3256 (P)	Nitric-oxide reductase subunit B (EC 1.7.99.7)
REB004113	COG2801 (L)	ISxac2 transposase
REB004112	COG3256 (P)	nitric oxide reductase, subunit B
REB004986	COG2801 (L)	ISxac2 transposase
REB003545	NOCOG	conserved hypothetical protein
REB004637	NOCOG	hypothetical protein
REB003983	NOCOG	hypothetical protein
REB004988	NOCOG	hypothetical protein
REB003979	NOCOG	hypothetical protein
REB004989	NOCOG	hypothetical protein
REB000409	COG0201 (U)	Preprotein translocase secY subunit
REB000396	COG0758 (LU)	Smf protein
REB004643	COG0551 (L)	DNA topoisomerase I (EC 5.99.1.2)
REB004644	NOCOG	TRANSCRIPTIONAL REGULATORY PROTEIN
REB004645	NOCOG	hypothetical protein

A.14. *E. blattae*-spezifische Proteine

Genname, COG, (COG-Gruppe), Annotation

REB000661	COG3593 (L)	hypothetical overcoming lysogenization defect protein
REB000687	COG1896 (R)	hypothetical hydrolase
REB000808	COG0499 (H)	Adenosylhomocysteinase (EC 3.3.1.1)
REB000844	NOCOG	conserved hypothetical protein
REB000857	NOCOG	hypothetical protein
REB000874	NOCOG	conserved hypothetical protein
REB000875	NOCOG	conserved hypothetical protein
REB000912	NOCOG	hypothetical protein
REB000918	COG2244 (R)	conserved hypothetical protein
REB000920	NOCOG	conserved hypothetical protein
REB000935	COG3090 (G)	N-acetylneuraminate transporter small subunit; N-acetylneuraminate transporter large subunit
REB000957	COG3547 (L)	Transposase

REB000989	NOCOG	hypothetical protein
REB000995	NOCOG	conserved hypothetical protein
REB001003	NOCOG	conserved hypothetical protein
REB001006	COG0477 (GEPR)	conserved hypothetical protein
REB001007	COG0463 (M)	putative glycosyltransferase
REB001010	NOCOG	conserved hypothetical protein
REB001053	COG1670 (J)	putative acetyltransferase
REB001058	COG0819 (K)	putative transcription activator
REB001061	COG0715 (P)	Thiamine-binding protein
REB001063	NOCOG	hypothetical protein
REB001069	NOCOG	conserved hypothetical protein
REB001087	NOCOG	hypothetical nonspecific acid phosphatase precursor
REB001088	NOCOG	conserved hypothetical protein
REB001089	NOCOG	conserved hypothetical protein
REB001091	COG2304 (R)	putative membrane protein
REB001092	NOCOG	hypothetical protein
REB001093	COG1721 (R)	conserved hypothetical protein
REB001223	COG2159 (R)	conserved hypothetical protein
REB001225	NOCOG	conserved hypothetical protein
REB001241	NOCOG	hypothetical protein
REB001343	NOCOG	conserved hypothetical protein
REB001410	COG3672 (S)	conserved hypothetical protein
REB001412	NOCOG	conserved hypothetical protein
REB001422	COG2931 (Q)	secreted hemolysin-type calcium-binding bacteriocin, putative
REB001437	COG2991 (S)	conserved hypothetical protein
REB001497	NOCOG	MdcA
REB001499	NOCOG	Malonate decarboxylase delta subunit
REB001501	NOCOG	gamma subunit of malonate decarboxylase
REB001503	NOCOG	HOLO-ACYL CARRIER PROTEIN, MdcG/MdcF
REB001539	NOCOG	Transposase
REB001540	NOCOG	conserved hypothetical protein
REB001599	COG1853 (R)	conserved hypothetical protein
REB001650	NOCOG	hypothetical protein
REB001653	NOCOG	putative salicylate-5-hydroxylase small oxygenase component oxidoreductase protein (EC 1.-.-.-)
REB001658	COG2329 (R)	conserved hypothetical protein
REB001659	COG0640 (K)	putative arsR-family transcriptional regulatory protein
REB001660	NOCOG	conserved hypothetical protein
REB001700	NOCOG	conserved hypothetical protein
REB001701	COG4638 (PR)	putative ferredoxin subunit
REB001704	COG0346 (E)	hypothetical lactoylglutathione lyase
REB001707	NOCOG	conserved hypothetical protein
REB001710	COG1712 (R)	putative dinucleotide-utilizing enzyme
REB001753	NOCOG	conserved hypothetical protein
REB001831	COG1989 (NOU)	hypothetical protein
REB001880	COG3306 (M)	putative beta1,4-galactosyltransferase WaaX
REB001965	COG3911 (R)	conserved hypothetical protein
REB001966	NOCOG	conserved hypothetical protein
REB002005	COG3485 (Q)	putative dioxygenase
REB002007	NOCOG	hypothetical protein
REB002051	COG1914 (P)	putative Na+/protine symporter
REB002053	COG4336 (S)	conserved hypothetical protein
REB002058	NOCOG	conserved hypothetical protein
REB002059	NOCOG	hypothetical protein
REB002060	NOCOG	hypothetical protein
REB002061	NOCOG	conserved hypothetical protein
REB002062	NOCOG	hypothetical protein
REB002095	NOCOG	hypothetical protein
REB002127	NOCOG	hypothetical phosphatase
REB002132	COG0438 (M)	putative glycosyltransferase
REB002259	COG3238 (S)	conserved hypothetical protein
REB002260	COG0477 (GEPR)	conserved hypothetical protein
REB002272	COG0596 (R)	3-oxoadipate enol-lactonase II (Enol-lactone hydrolase II) (Beta-ketoadipate enol-lactone hydrolase II)
REB002274	COG4829 (Q)	Muconolactone delta-isomerase
REB002275	COG3485 (Q)	Catechol 1,2-dioxygenase (EC 1.13.11.1)
REB002277	COG5517 (Q)	hypothetical 2-halobenzoate 1,2-dioxygenase small subunit
REB002280	NOCOG	hypothetical protein
REB002281	NOCOG	hypothetical protein
REB002284	NOCOG	hypothetical protein
REB002288	COG3659 (M)	Carbohydrate-selective porin, OprB family
REB002299	COG2608 (P)	putative mercuric transport periplasmic protein MerP
REB002300	COG1305 (E)	transglutaminase family protein
REB002361	NOCOG	hypothetical protein
REB002378	NOCOG	unknown protein encoded by prophage CP-933N
REB002382	NOCOG	conserved hypothetical protein
REB002457	COG2159 (R)	putative amidohydrolase
REB002485	COG3837 (S)	conserved hypothetical protein
REB002494	COG4313 (C)	conserved hypothetical protein
REB002544	NOCOG	Arsenical resistance operon trans-acting repressor arsD
REB002545	COG0003 (P)	Arsenical pump-driving ATPase (EC 3.6.3.16)
REB002550	NOCOG	conserved hypothetical protein
REB002561	NOCOG	conserved hypothetical protein
REB002567	COG0730 (R)	putative permease
REB002570	COG1227 (C)	Inorganic pyrophosphatase/exopolyphosphatase
REB002571	NOCOG	hypothetical protein
REB002581	COG1402 (R)	conserved hypothetical protein
REB002584	NOCOG	hypothetical protein
REB002620	COG3662 (S)	conserved hypothetical protein
REB002621	COG2070 (R)	2-nitropropane dioxygenase (EC 1.13.11.32)
REB002622	NOCOG	hypothetical protein
REB002623	NOCOG	hypothetical protein
REB002633	NOCOG	conserved hypothetical protein

REB002635	COG1846 (K)	transcriptional regulatory protein
REB002636	COG3473 (Q)	maleate cis-trans isomerase
REB002637	COG0388 (R)	putative nitrilase
REB002641	COG0824 (R)	putative 4-hydroxybenzoyl-Coa thioesterase protein
REB002642	COG1014 (C)	indolepyruvate ferredoxin oxidoreductase, beta chain
REB002643	COG4231 (C)	Indolepyruvate oxidoreductase subunit iorA (EC 1.-.-.-)
REB002646	COG1878 (R)	conserved hypothetical protein
REB002647	COG1846 (K)	transcriptional regulator
REB002688	NOCOG	conserved hypothetical protein
REB002807	NOCOG	unknown protein encoded by cryptic prophage CP-933M
REB002808	NOCOG	hypothetical protein
REB002814	NOCOG	putative glucuronyl hydrolase
REB002818	COG3218 (R)	hypothetical membrane protein
REB002819	COG1463 (Q)	ABC-type transport system involved in resistance to organic solvents, periplasmic component
REB002822	NOCOG	hypothetical protein
REB002823	NOCOG	Putative toxin of gyrase inhibiting toxin-antitoxin system
REB002858	NOCOG	conserved hypothetical protein
REB002870	NOCOG	tail spike protein
REB002877	NOCOG	capsid protein
REB003035	NOCOG	hypothetical protein
REB003037	NOCOG	hypothetical protein
REB003118	NOCOG	hypothetical protein
REB003119	NOCOG	hypothetical protein
REB003120	COG3747 (L)	Phage terminase-like protein, small subunit
REB003123	COG3740 (R)	Phage Prohead Protease
REB003124	NOCOG	major head protein
REB003125	NOCOG	gp7
REB003126	NOCOG	conserved hypothetical protein (phage related)
REB003127	COG0486 (R)	conserved hypothetical protein (phage related)
REB003128	NOCOG	conserved hypothetical protein (phage related)
REB003129	NOCOG	conserved hypothetical protein (phage related)
REB003130	NOCOG	conserved hypothetical protein (phage related)
REB003134	NOCOG	putative Chromosome segregation ATPase
REB003141	NOCOG	hypothetical protein
REB003154	COG1670 (J)	Acetyltransferases, including N-acetylases of ribosomal proteins
REB003187	COG2010 (C)	cytochrome c subunit II
REB003188	COG2303 (E)	dehydrogenase subunit I
REB003189	NOCOG	dehydrogluconate dehydrogenase, small subunit (EC 1.1.99.4)
REB003204	NOCOG	conserved hypothetical protein
REB003215	COG3550 (R)	conserved hypothetical protein
REB003228	COG2358 (R)	immunogenic protein
REB003231	COG2315 (S)	conserved hypothetical protein
REB003233	COG3193 (R)	hypothetical protein
REB003237	COG1960 (I)	Acyl-CoA dehydrogenase (EC 1.3.99.-)
REB003242	NOCOG	hypothetical protein
REB003252	NOCOG	conserved hypothetical protein
REB003266	COG2010 (C)	gluconate 2-dehydrogenase, cytochrome c subunit (EC 1.1.99.3)
REB003276	NOCOG	conserved hypothetical protein
REB003285	NOCOG	putative glycerol dehydratase reactivation factor small subunit
REB003290	COG2842 (R)	DNA transposition protein
REB003291	NOCOG	host nuclease inhibitor protein
REB003292	NOCOG	conserved hypothetical protein
REB003297	NOCOG	conserved hypothetical protein
REB003298	NOCOG	conserved hypothetical protein
REB003299	NOCOG	Mu-like phage gp26
REB003304	COG4388 (R)	putative prophage MuSo2, protein Gp32
REB003305	NOCOG	putative prophage MuSo2, major head subunit
REB003306	NOCOG	Mu-like prophage FluMu protein gp36
REB003307	NOCOG	putative prophage MuSo2, virion morphogenesis protein
REB003308	NOCOG	conserved hypothetical protein
REB003309	NOCOG	hypothetical protein
REB003311	NOCOG	Mu-like prophage FluMu tail tube protein
REB003312	NOCOG	Mu-like prophage FluMu protein gp41
REB003313	NOCOG	probable bacteriophage tail protein
REB003322	NOCOG	hypothetical protein
REB003324	COG0640 (K)	Transcriptional regulator, ArsR family
REB003327	NOCOG	hypothetical protein
REB003340	NOCOG	hypothetical protein
REB003360	NOCOG	heavy metal RND efflux outer membrane protein, CzcC family
REB003368	NOCOG	putative type 3 fimbrial protein mrkA precursor
REB003377	NOCOG	hypothetical protein
REB003381	NOCOG	conserved hypothetical protein
REB003446	NOCOG	conserved hypothetical protein
REB003467	NOCOG	conserved hypothetical protein
REB003468	NOCOG	hypothetical protein
REB003470	NOCOG	hypothetical protein
REB003472	NOCOG	hypothetical protein
REB003473	NOCOG	hypothetical protein
REB003474	NOCOG	hypothetical protein
REB003476	NOCOG	hypothetical protein
REB003477	NOCOG	hypothetical protein
REB003478	NOCOG	conserved hypothetical protein
REB003479	NOCOG	hypothetical protein
REB003481	NOCOG	hypothetical protein
REB003482	NOCOG	hypothetical protein
REB003483	NOCOG	conserved hypothetical protein
REB003484	NOCOG	hypothetical protein
REB003487	NOCOG	hypothetical protein
REB003489	NOCOG	conserved hypothetical protein
REB003490	NOCOG	hypothetical protein
REB003492	NOCOG	conserved hypothetical protein
REB003494	NOCOG	hypothetical protein

REB003500	NOCOG	hypothetical protein
REB003504	NOCOG	hypothetical protein
REB003505	NOCOG	hypothetical protein
REB003509	NOCOG	hypothetical protein
REB003513	NOCOG	hypothetical protein
REB003514	NOCOG	hypothetical protein
REB003515	NOCOG	hypothetical protein
REB003516	NOCOG	glycerol dehydratase 21kDa subunit (partial)
REB003519	NOCOG	Cim protein (Kil protein)
REB003520	NOCOG	hypothetical protein gp9
REB003521	NOCOG	hypothetical protein
REB003523	NOCOG	hypothetical protein
REB003525	NOCOG	hypothetical protein
REB003526	NOCOG	hypothetical protein
REB003531	NOCOG	hypothetical protein
REB003533	NOCOG	hypothetical protein
REB003534	NOCOG	hypothetical protein
REB003538	NOCOG	hypothetical protein
REB003541	NOCOG	hypothetical protein
REB003557	NOCOG	hypothetical protein
REB003558	NOCOG	hypothetical protein
REB003559	NOCOG	hypothetical protein
REB003560	NOCOG	hypothetical protein
REB003562	NOCOG	conserved hypothetical protein
REB003571	NOCOG	unknown phage protein
REB003575	NOCOG	conserved hypothetical protein
REB003578	NOCOG	hypothetical protein
REB003581	NOCOG	hypothetical protein
REB003583	NOCOG	hypothetical protein
REB003584	NOCOG	hypothetical protein
REB003585	NOCOG	hypothetical protein
REB003586	NOCOG	hypothetical protein
REB003587	NOCOG	hypothetical protein
REB003588	COG0454 (KR)	possible acetyltransferase
REB003589	NOCOG	hypothetical protein
REB003591	NOCOG	hypothetical protein
REB003592	NOCOG	hypothetical protein
REB003593	NOCOG	hypothetical protein
REB003598	NOCOG	hypothetical histidine kinase
REB003599	NOCOG	conserved hypothetical protein
REB003603	NOCOG	hypothetical protein
REB003605	NOCOG	hypothetical protein
REB003607	NOCOG	hypothetical protein
REB003608	NOCOG	hypothetical protein
REB003609	NOCOG	conserved hypothetical protein (phage related)
REB003610	NOCOG	hypothetical protein
REB003612	NOCOG	hypothetical protein
REB003614	NOCOG	hypothetical protein
REB003615	NOCOG	hypothetical protein
REB003616	NOCOG	hypothetical protein
REB003617	NOCOG	hypothetical protein
REB003618	NOCOG	hypothetical protein
REB003619	NOCOG	hypothetical protein
REB003620	NOCOG	conserved hypothetical protein
REB003621	NOCOG	hypothetical protein
REB003623	NOCOG	hypothetical protein
REB003624	COG5475 (S)	hypothetical protein
REB003627	NOCOG	hypothetical protein
REB003628	NOCOG	hypothetical protein
REB003629	NOCOG	hypothetical protein
REB003631	NOCOG	hypothetical protein
REB003632	NOCOG	hypothetical protein
REB003633	NOCOG	hypothetical protein
REB003634	NOCOG	conserved hypothetical protein
REB003638	NOCOG	hypothetical protein
REB003640	NOCOG	putative Phage-like element PBSX protein xkd0
REB003648	NOCOG	hypothetical protein
REB003651	NOCOG	hypothetical protein
REB003652	NOCOG	hypothetical prophage protein
REB003653	NOCOG	hypothetical protein
REB003654	NOCOG	hypothetical protein
REB003655	NOCOG	hypothetical protein
REB003656	NOCOG	hypothetical protein
REB003657	NOCOG	hypothetical protein
REB003658	NOCOG	hypothetical protein
REB003660	NOCOG	hypothetical protein
REB003669	NOCOG	hypothetical protein
REB003675	NOCOG	hypothetical protein
REB003688	NOCOG	hypothetical protein
REB003689	NOCOG	hypothetical protein
REB003692	NOCOG	hypothetical protein
REB003695	NOCOG	hypothetical protein
REB003696	NOCOG	hypothetical protein
REB003699	NOCOG	hypothetical protein
REB003701	NOCOG	Transposase
REB003702	NOCOG	hypothetical protein
REB003703	NOCOG	hypothetical protein
REB003704	NOCOG	hypothetical protein
REB003706	NOCOG	hypothetical protein
REB003714	NOCOG	hypothetical protein
REB003716	NOCOG	hypothetical protein
REB003717	NOCOG	hypothetical protein

REB003718	NOCOG	conserved hypothetical protein
REB003721	NOCOG	conserved hypothetical protein
REB003733	NOCOG	hypothetical protein
REB003734	NOCOG	hypothetical protein
REB003735	NOCOG	conserved hypothetical protein
REB003737	NOCOG	hypothetical protein
REB003739	NOCOG	hypothetical protein
REB003741	COG0457 (R)	hypothetical protein
REB003743	NOCOG	hypothetical protein
REB003745	NOCOG	conserved hypothetical protein
REB003748	NOCOG	hypothetical protein
REB003749	NOCOG	conserved hypothetical protein
REB003750	NOCOG	hypothetical protein
REB003753	NOCOG	hypothetical protein
REB003755	NOCOG	hypothetical protein
REB003757	NOCOG	conserved hypothetical protein
REB003761	NOCOG	hypothetical protein
REB003768	NOCOG	hypothetical protein
REB003770	NOCOG	hypothetical protein
REB003781	COG2010 (C)	hypothetical Cytochrome c
REB003785	NOCOG	hypothetical protein
REB003786	NOCOG	hypothetical protein
REB003788	NOCOG	hypothetical protein
REB003794	NOCOG	hypothetical protein
REB003795	NOCOG	hypothetical protein
REB003801	NOCOG	hypothetical protein
REB003807	NOCOG	hypothetical protein
REB003808	NOCOG	hypothetical protein
REB003810	NOCOG	hypothetical protein
REB003813	NOCOG	hypothetical protein
REB003814	NOCOG	hypothetical protein
REB003818	NOCOG	hypothetical protein
REB003819	NOCOG	hypothetical protein
REB003820	NOCOG	hypothetical protein
REB003825	NOCOG	conserved hypothetical protein
REB003831	NOCOG	hypothetical protein
REB003832	NOCOG	conserved hypothetical protein
REB003837	NOCOG	conserved hypothetical protein
REB003852	COG1396 (K)	putative transcriptional regulator
REB003853	COG4683 (S)	conserved hypothetical protein
REB003859	NOCOG	hypothetical protein
REB003860	NOCOG	hypothetical protein
REB003863	NOCOG	hypothetical protein
REB003870	NOCOG	hypothetical protein
REB003872	NOCOG	hypothetical protein
REB003873	NOCOG	hypothetical protein
REB003874	NOCOG	hypothetical protein
REB003875	NOCOG	hypothetical protein
REB003876	NOCOG	hypothetical protein
REB003877	COG3549 (R)	conserved hypothetical protein
REB003883	NOCOG	hypothetical protein
REB003884	NOCOG	hypothetical protein
REB003891	NOCOG	hypothetical protein
REB003892	NOCOG	hypothetical protein
REB003973	NOCOG	hypothetical protein
REB003974	NOCOG	hypothetical protein
REB003979	NOCOG	hypothetical protein
REB003983	NOCOG	hypothetical protein
REB003990	NOCOG	hypothetical protein
REB003993	NOCOG	hypothetical protein
REB004005	NOCOG	hypothetical protein
REB004006	NOCOG	hypothetical protein
REB004011	NOCOG	hypothetical protein
REB004013	NOCOG	hypothetical protein
REB004016	NOCOG	conserved hypothetical protein
REB004017	NOCOG	conserved hypothetical protein
REB004018	NOCOG	hypothetical protein
REB004019	NOCOG	hypothetical protein
REB004020	NOCOG	hypothetical protein
REB004022	NOCOG	conserved hypothetical protein
REB004023	NOCOG	hypothetical protein
REB004024	NOCOG	hypothetical protein
REB004030	NOCOG	hypothetical protein
REB004032	NOCOG	hypothetical protein
REB004033	COG3547 (L)	putative transposase
REB004042	NOCOG	hypothetical protein
REB004045	NOCOG	hypothetical protein
REB004055	NOCOG	Pre-neck appendage protein
REB004069	NOCOG	hypothetical protein
REB004073	NOCOG	hypothetical protein
REB004091	NOCOG	hypothetical protein
REB004095	NOCOG	hypothetical protein
REB004112	COG3256 (P)	nitric oxide reductase, subunit B
REB004114	NOCOG	hypothetical protein
REB004118	NOCOG	hypothetical protein
REB004123	NOCOG	hypothetical protein
REB004125	NOCOG	conserved hypothetical protein
REB004128	NOCOG	hypothetical protein
REB004150	COG3657 (S)	conserved hypothetical protein
REB004161	NOCOG	putative type II intron maturase
REB004183	NOCOG	hypothetical protein
REB004184	NOCOG	hypothetical protein

REB004191	NOCOG	hypothetical protein
REB004214	COG3448 (T)	putative membrane protein
REB004215	NOCOG	hypothetical protein
REB004217	NOCOG	hypothetical protein
REB004220	NOCOG	hypothetical protein
REB004222	NOCOG	hypothetical protein
REB004223	COG3053 (C)	hypothetical protein
REB004224	NOCOG	hypothetical protein
REB004226	NOCOG	hypothetical protein
REB004227	NOCOG	hypothetical protein
REB004233	COG5302 (R)	putative antitoxin of gyrase inhibiting toxin-antitoxin system
REB004234	NOCOG	hypothetical protein
REB004235	NOCOG	resolvase
REB004238	NOCOG	hypothetical protein sb50
REB004239	NOCOG	hypothetical protein
REB004241	NOCOG	hypothetical protein
REB004242	NOCOG	hypothetical protein
REB004243	NOCOG	hypothetical protein
REB004246	NOCOG	hypothetical protein
REB004250	NOCOG	hypothetical protein
REB004255	NOCOG	hypothetical protein
REB004257	NOCOG	hypothetical protein
REB004259	NOCOG	hypothetical protein
REB004260	NOCOG	hypothetical protein
REB004261	NOCOG	hypothetical protein
REB004262	NOCOG	hypothetical protein
REB004263	NOCOG	hypothetical protein
REB004264	NOCOG	hypothetical protein
REB004267	NOCOG	hypothetical protein
REB004269	NOCOG	hypothetical protein
REB004270	NOCOG	hypothetical protein
REB004271	NOCOG	hypothetical protein
REB004273	COG1917 (S)	conserved hypothetical protein
REB004278	COG3547 (L)	putative transposase
REB004281	NOCOG	conserved hypothetical protein
REB004285	NOCOG	hypothetical protein
REB004286	NOCOG	hypothetical protein
REB004287	NOCOG	hypothetical protein
REB004289	NOCOG	hypothetical protein
REB004290	NOCOG	hypothetical protein
REB004294	NOCOG	hypothetical protein
REB004299	NOCOG	hypothetical protein
REB004300	NOCOG	hypothetical protein
REB004303	NOCOG	hypothetical protein
REB004308	NOCOG	hypothetical protein
REB004309	NOCOG	hypothetical protein
REB004312	NOCOG	hypothetical protein
REB004323	NOCOG	hypothetical protein
REB004325	NOCOG	hypothetical protein
REB004327	NOCOG	hypothetical protein
REB004329	NOCOG	hypothetical protein
REB004330	NOCOG	hypothetical protein
REB004331	NOCOG	hypothetical protein
REB004332	NOCOG	hypothetical protein
REB004333	NOCOG	hypothetical protein
REB004352	COG0629 (L)	single-strand DNA-binding protein (partial)
REB004356	NOCOG	hypothetical protein
REB004357	NOCOG	hypothetical protein
REB004359	NOCOG	hypothetical protein
REB004367	NOCOG	hypothetical protein
REB004369	NOCOG	hypothetical protein
REB004370	NOCOG	hypothetical protein
REB004371	NOCOG	hypothetical protein
REB004373	NOCOG	hypothetical protein
REB004374	NOCOG	hypothetical protein
REB004376	NOCOG	hypothetical protein
REB004377	NOCOG	hypothetical protein
REB004378	NOCOG	hypothetical protein
REB004379	COG2303 (E)	gluconate 2-dehydrogenase (FAD) (EC 1.1.99.3) catalytic chain precursor
REB004380	NOCOG	conserved hypothetical protein
REB004381	NOCOG	hypothetical protein
REB004391	NOCOG	hypothetical protein
REB004393	NOCOG	conserved hypothetical protein
REB004413	NOCOG	hypothetical protein
REB004456	COG4957 (K)	putative transcriptional regulator
REB004463	NOCOG	hypothetical protein
REB004467	NOCOG	hypothetical protein
REB004470	NOCOG	hypothetical protein
REB004474	NOCOG	hypothetical protein
REB004478	NOCOG	hypothetical protein
REB004482	NOCOG	hypothetical protein
REB004491	NOCOG	hypothetical protein
REB004494	NOCOG	hypothetical protein
REB004498	NOCOG	hypothetical protein
REB004502	NOCOG	hypothetical protein
REB004505	NOCOG	hypothetical protein
REB004507	NOCOG	hypothetical protein
REB004508	NOCOG	hypothetical membrane protein
REB004510	NOCOG	hypothetical protein
REB004517	NOCOG	hypothetical protein
REB004536	NOCOG	conserved hypothetical protein
REB004549	NOCOG	hypothetical protein

REB004550	NOCOG	hypothetical protein
REB004551	NOCOG	hypothetical protein
REB004555	NOCOG	hypothetical protein
REB004588	NOCOG	hypothetical protein
REB004591	NOCOG	hypothetical protein
REB004595	NOCOG	hypothetical protein
REB004599	NOCOG	hypothetical protein
REB004637	NOCOG	hypothetical protein
REB004645	NOCOG	hypothetical protein
REB004648	NOCOG	conserved hypothetical protein
REB004650	NOCOG	hypothetical protein
REB004651	NOCOG	hypothetical protein
REB004658	NOCOG	hypothetical protein
REB004665	NOCOG	hypothetical protein HK022p34
REB004672	NOCOG	hypothetical protein
REB004690	NOCOG	hypothetical protein
REB004710	NOCOG	hypothetical protein
REB004711	NOCOG	hypothetical protein
REB004720	NOCOG	hypothetical protein
REB004722	NOCOG	hypothetical protein
REB004723	NOCOG	unnamed protein product
REB004725	COG2183 (K)	Hypothetical Protein
REB004726	NOCOG	O (DNA replication;299)
REB004730	NOCOG	hypothetical protein sb54
REB004731	NOCOG	hypothetical protein
REB004732	NOCOG	Orf-129
REB004733	COG0358 (L)	Hypothetical Protein
REB004734	NOCOG	terminase small subunit
REB004737	NOCOG	hypothetical protein
REB004740	NOCOG	hypothetical protein
REB004743	NOCOG	hypothetical protein
REB004744	NOCOG	conserved hypothetical protein
REB004745	NOCOG	hypothetical protein
REB004748	NOCOG	hypothetical protein
REB004749	NOCOG	hypothetical protein
REB004752	COG3525 (G)	Chitobiase precursor (EC 3.2.1.52)
REB004756	NOCOG	hypothetical protein
REB004762	NOCOG	hypothetical protein
REB004763	NOCOG	hypothetical protein
REB004764	NOCOG	hypothetical protein
REB004771	NOCOG	hypothetical protein
REB004772	NOCOG	hypothetical protein
REB004774	NOCOG	hypothetical protein
REB004779	NOCOG	hypothetical protein
REB004782	NOCOG	hypothetical protein
REB004783	NOCOG	hypothetical protein
REB004784	NOCOG	hypothetical protein
REB004785	NOCOG	hypothetical protein
REB004787	NOCOG	hypothetical protein
REB004790	NOCOG	hypothetical protein
REB004792	NOCOG	hypothetical protein
REB004793	COG0739 (M)	hypothetical protein
REB004794	NOCOG	hypothetical protein
REB004795	NOCOG	hypothetical protein
REB004796	NOCOG	hypothetical protein
REB004797	NOCOG	hypothetical protein
REB004799	NOCOG	hypothetical protein
REB004801	NOCOG	hypothetical protein
REB004803	NOCOG	hypothetical protein
REB004804	NOCOG	hypothetical protein
REB004805	NOCOG	hypothetical protein
REB004807	NOCOG	hypothetical protein
REB004808	NOCOG	hypothetical protein
REB004809	NOCOG	hypothetical protein
REB004810	COG2981 (E)	hypothetical protein
REB004811	COG3773 (M)	hypothetical protein
REB004812	NOCOG	hypothetical protein
REB004815	NOCOG	hypothetical protein
REB004816	NOCOG	hypothetical protein
REB004817	NOCOG	hypothetical protein
REB004818	NOCOG	hypothetical protein
REB004819	NOCOG	conserved hypothetical protein
REB004821	NOCOG	hypothetical protein
REB004822	NOCOG	hypothetical protein
REB004824	NOCOG	hypothetical protein
REB004825	NOCOG	hypothetical protein
REB004832	NOCOG	hypothetical protein
REB004833	COG2872 (R)	Predicted metal-dependent hydrolases related to alanyl-tRNA synthetase HxxxH domain
REB004834	COG0697 (GER)	permease
REB004836	NOCOG	hypothetical protein
REB004840	NOCOG	hypothetical protein
REB004841	NOCOG	hypothetical protein
REB004842	NOCOG	hypothetical protein
REB004843	NOCOG	hypothetical protein
REB004844	NOCOG	hypothetical protein
REB004851	NOCOG	hypothetical protein
REB004857	NOCOG	hypothetical protein
REB004859	NOCOG	hypothetical protein
REB004863	NOCOG	hypothetical protein
REB004867	NOCOG	hypothetical protein
REB004870	NOCOG	hypothetical protein
REB004871	NOCOG	hypothetical protein

REB004875	NOCOG	hypothetical protein
REB004877	NOCOG	hypothetical protein
REB004878	NOCOG	hypothetical protein
REB004879	NOCOG	hypothetical protein
REB004880	NOCOG	hypothetical protein
REB004881	NOCOG	hypothetical protein
REB004882	NOCOG	hypothetical protein
REB004883	NOCOG	hypothetical protein
REB004884	NOCOG	hypothetical protein
REB004885	NOCOG	hypothetical protein
REB004888	NOCOG	hypothetical protein
REB004890	NOCOG	hypothetical protein
REB004895	COG0583 (K)	LysR-type regulatory protein
REB004896	NOCOG	hypothetical protein
REB004897	NOCOG	hypothetical protein
REB004910	NOCOG	hypothetical protein
REB004911	NOCOG	hypothetical protein
REB004912	NOCOG	hypothetical protein
REB004914	NOCOG	hypothetical protein
REB004915	NOCOG	hypothetical protein
REB004916	NOCOG	hypothetical protein
REB004918	NOCOG	hypothetical protein
REB004920	NOCOG	hypothetical protein
REB004926	NOCOG	hypothetical protein
REB004927	NOCOG	hypothetical protein
REB004928	NOCOG	hypothetical protein
REB004930	NOCOG	hypothetical protein
REB004931	NOCOG	hypothetical protein
REB004932	NOCOG	hypothetical protein
REB004933	NOCOG	hypothetical protein
REB004934	NOCOG	hypothetical protein
REB004935	NOCOG	hypothetical protein
REB004937	NOCOG	hypothetical protein
REB004939	NOCOG	hypothetical protein
REB004940	NOCOG	hypothetical protein
REB004942	NOCOG	hypothetical protein
REB004944	NOCOG	hypothetical protein
REB004947	NOCOG	hypothetical protein
REB004951	NOCOG	hypothetical protein
REB004953	NOCOG	hypothetical protein
REB004955	NOCOG	hypothetical protein
REB004959	COG3547 (L)	putative IS402 family transposase
REB004962	NOCOG	hypothetical protein
REB004963	NOCOG	hypothetical protein
REB004969	NOCOG	hypothetical protein
REB004970	NOCOG	hypothetical protein
REB004973	NOCOG	hypothetical protein
REB004974	NOCOG	hypothetical protein
REB004976	NOCOG	hypothetical protein
REB004978	NOCOG	terminase small subunit
REB004980	COG3547 (L)	putative IS402 family transposase
REB004983	NOCOG	hypothetical protein
REB004984	NOCOG	hypothetical protein
REB004988	NOCOG	hypothetical protein

A.15. *E. blattae*-spezifische Proteine (Fremdgene)

Liste der *E. blattae*-spezifischen Proteine, die von SIGI als Fremdgene eingestuft wurden.

Genname, COG, (COG-Gruppe), Annotation

REB000025	COG3256 (P)	Nitric-oxide reductase subunit B (EC 1.7.99.7)
REB000026	COG1428 (F)	hypothethical deoxyguanosine/deoxyadenosine kinase
REB000115	NOCOG	hypothetical protein
REB000116	NOCOG	hypothetical protein
REB000121	NOCOG	conserved hypothetical protein
REB000240	COG0270 (L)	cytosine methylase
REB000241	NOCOG	Hypothetical Protein
REB000245	NOCOG	hypothetical protein sb49
REB000808	COG0499 (H)	Adenosylhomocysteinase (EC 3.3.1.1)
REB000957	COG3547 (L)	Transposase
REB001422	COG2931 (Q)	secreted hemolysin-type calcium-binding bacteriocin, putative
REB001650	NOCOG	hypothetical protein
REB001880	COG3306 (M)	putative beta1,4-galactosyltransferase WaaX
REB002127	NOCOG	hypothetical phosphatase

```
REB002132 COG0438 (M)   putative glycosyltransferase
REB002807 NOCOG         unknown protein encoded by cryptic prophage CP-933M
REB002823 NOCOG         Putative toxin of gyrase inhibiting toxin-antitoxin system
REB002870 NOCOG         tail spike protein
REB003118 NOCOG         hypothetical protein
REB003125 NOCOG         gp7
REB003126 NOCOG         conserved hypothetical protein (phage related)
REB003128 NOCOG         conserved hypothetical protein (phage related)
REB003129 NOCOG         conserved hypothetical protein (phage related)
REB003130 NOCOG         conserved hypothetical protein (phage related)
REB003134 NOCOG         putative Chromosome segregation ATPase
REB003215 COG3550 (R)   conserved hypothetical protein
REB003291 NOCOG         host nuclease inhibitor protein
REB003292 NOCOG         conserved hypothetical protein
REB003299 NOCOG         Mu-like phage gp26
REB003304 COG4388 (R)   putative prophage MuSo2, protein Gp32
REB003305 NOCOG         putative prophage MuSo2, major head subunit
REB003306 NOCOG         Mu-like prophage FluMu protein gp36
REB003311 NOCOG         Mu-like prophage FluMu tail tube protein
REB003312 NOCOG         Mu-like prophage FluMu protein gp41
REB003313 NOCOG         probable bacteriophage tail protein
REB003515 NOCOG         hypothetical protein
REB003516 NOCOG         Glycerol dehydratase 21kDa subunit (partial)
REB003519 NOCOG         Cim protein (Kil protein)
REB003520 NOCOG         hypothetical protein gp9
REB003521 NOCOG         hypothetical protein
REB003523 NOCOG         hypothetical protein
REB003525 NOCOG         hypothetical protein
REB003526 NOCOG         hypothetical protein
REB003585 NOCOG         hypothetical protein
REB003586 NOCOG         hypothetical protein
REB003587 NOCOG         hypothetical protein
REB003588 COG0454 (KR)  possible acetyltransferase
REB003589 NOCOG         hypothetical protein
REB003591 NOCOG         hypothetical protein
REB003592 NOCOG         hypothetical protein
REB003593 NOCOG         hypothetical protein
REB003598 NOCOG         hypothetical histidine kinase
REB003599 NOCOG         conserved hypothetical protein
REB003603 NOCOG         hypothetical protein
REB003605 NOCOG         hypothetical protein
REB003607 NOCOG         hypothetical protein
REB003608 NOCOG         hypothetical protein
REB003612 NOCOG         hypothetical protein
REB003614 NOCOG         hypothetical protein
REB003615 NOCOG         hypothetical protein
REB003616 NOCOG         hypothetical protein
REB003617 NOCOG         hypothetical protein
REB003627 NOCOG         hypothetical protein
REB003634 NOCOG         conserved hypothetical protein
REB003638 NOCOG         hypothetical protein
REB003640 NOCOG         putative Phage-like element PBSX protein xkdO
REB003714 NOCOG         hypothetical protein
REB003716 NOCOG         hypothetical protein
REB003717 NOCOG         hypothetical protein
REB003718 NOCOG         conserved hypothetical protein
REB003733 NOCOG         hypothetical protein
REB003734 NOCOG         hypothetical protein
REB003737 NOCOG         hypothetical protein
REB003739 NOCOG         hypothetical protein
REB003741 COG0457 (R)   hypothetical protein
REB003743 NOCOG         hypothetical protein
REB003745 NOCOG         conserved hypothetical protein
REB003748 NOCOG         hypothetical protein
REB003749 NOCOG         conserved hypothetical protein
REB003750 NOCOG         hypothetical protein
REB003753 NOCOG         hypothetical protein
REB003755 NOCOG         hypothetical protein
REB003757 NOCOG         conserved hypothetical protein
REB003761 NOCOG         hypothetical protein
REB003794 NOCOG         hypothetical protein
REB003795 NOCOG         hypothetical protein
REB003808 NOCOG         hypothetical protein
REB003877 COG3549 (R)   conserved hypothetical protein
REB003883 NOCOG         hypothetical protein
REB003979 NOCOG         hypothetical protein
REB003983 NOCOG         hypothetical protein
REB004005 NOCOG         hypothetical protein
REB004006 NOCOG         hypothetical protein
REB004016 NOCOG         conserved hypothetical protein
REB004017 NOCOG         conserved hypothetical protein
REB004018 NOCOG         hypothetical protein
REB004019 NOCOG         hypothetical protein
REB004020 NOCOG         hypothetical protein
REB004022 NOCOG         conserved hypothetical protein
REB004023 NOCOG         hypothetical protein
REB004024 NOCOG         hypothetical protein
REB004112 COG3256 (P)   nitric oxide reductase, subunit B
REB004118 NOCOG         hypothetical protein
REB004214 COG3448 (T)   putative membrane protein
REB004233 COG5302 (R)   putative antitoxin of gyrase inhibiting toxin-antitoxin system
REB004235 NOCOG         resolvase
```

```
REB004278 COG3547 (L)   putative transposase
REB004303 NOCOG         hypothetical protein
REB004376 NOCOG         hypothetical protein
REB004377 NOCOG         hypothetical protein
REB004463 NOCOG         hypothetical protein
REB004588 NOCOG         hypothetical protein
REB004591 NOCOG         hypothetical protein
REB004637 NOCOG         hypothetical protein
REB004645 NOCOG         hypothetical protein
REB004648 NOCOG         conserved hypothetical protein
REB004665 NOCOG         hypothetical protein HK022p34
REB004690 NOCOG         hypothetical protein
REB004723 NOCOG         unnamed protein product
REB004725 COG2183 (K)   Hypothetical Protein
REB004726 NOCOG         O (DNA replication;299)
REB004730 NOCOG         hypothetical protein sb54
REB004732 NOCOG         Orf-129
REB004733 COG0358 (L)   Hypothetical Protein
REB004734 NOCOG         terminase small subunit
REB004740 NOCOG         hypothetical protein
REB004783 NOCOG         hypothetical protein
REB004784 NOCOG         hypothetical protein
REB004810 COG2981 (E)   hypothetical protein
REB004817 NOCOG         hypothetical protein
REB004818 NOCOG         hypothetical protein
REB004822 NOCOG         hypothetical protein
REB004824 NOCOG         hypothetical protein
REB004832 NOCOG         hypothetical protein
REB004833 COG2872 (R)   Predicted metal-dependent hydrolases related to alanyl-tRNA synthetase HxxxH domain
REB004834 COG0697 (GER) permease
REB004870 NOCOG         hypothetical protein
REB004871 NOCOG         hypothetical protein
REB004875 NOCOG         hypothetical protein
REB004880 NOCOG         hypothetical protein
REB004881 NOCOG         hypothetical protein
REB004882 NOCOG         hypothetical protein
REB004910 NOCOG         hypothetical protein
REB004911 NOCOG         hypothetical protein
REB004937 NOCOG         hypothetical protein
REB004939 NOCOG         hypothetical protein
REB004951 NOCOG         hypothetical protein
REB004953 NOCOG         hypothetical protein
REB004955 NOCOG         hypothetical protein
REB004959 COG3547 (L)   putative IS402 family transposase
REB004962 NOCOG         hypothetical protein
REB004963 NOCOG         hypothetical protein
REB004969 NOCOG         hypothetical protein
REB004978 NOCOG         terminase small subunit
REB004980 COG3547 (L)   putative IS402 family transposase
REB004983 NOCOG         hypothetical protein
REB004988 NOCOG         hypothetical protein
```

A.16. *E. blattae*-spezifische Proteine (Vergleich Enterobakterien)

E. blattae-spezifischen Proteine, die sich in allen anderes Enteros wiederfinden. Genname, COG, (COG-Gruppe), Annotation

```
REB000003 COG0484 (O)    Chaperone protein dnaJ
REB000018 COG0324 (J)    tRNA delta(2)-isopentenylpyrophosphate transferase (EC 2.5.1.8)
REB000019 COG0323 (L)    DNA mismatch repair protein mutL
REB000020 COG0860 (M)    N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)
REB000023 COG0541 (U)    SIGNAL RECOGNITION PARTICLE, SUBUNIT FFH/SRP54
REB000054 COG0272 (L)    DNA ligase (EC 6.5.1.2)
REB000078 COG0441 (J)    Threonyl-tRNA synthetase (EC 6.1.1.3)
REB000087 COG0336 (J)    tRNA (Guanine-N(1)-)-methyltransferase (EC 2.1.1.31)
REB000103 COG0526 (OC)   Thiol:disulfide interchange protein dsbA
REB000110 COG0008 (J)    Glutaminyl-tRNA synthetase (EC 6.1.1.18)
REB000152 COG1477 (H)    Thiamine biosynthesis lipoprotein apbE precursor
REB000181 COG0125 (F)    Thymidylate kinase (EC 2.7.4.9)
REB000182 COG0470 (L)    DNA polymerase III, delta' subunit (EC 2.7.7.7)
REB000183 COG0084 (L)    hypothetical hydrolase
```

```
REB000184 COG1263 (G)    PTS system, glucose-specific IIBC component (EC 2.7.1.69)
REB000190 COG3066 (L)    methyl-directed mismatch repair protein
REB000197 COG0019 (E)    Diaminopimelate decarboxylase (EC 4.1.1.20)
REB000208 COG1217 (T)    GTP-binding protein typA/BipA
REB000263 COG1194 (L)    A/G-specific adenine DNA glycosylase (EC 3.2.2.-)
REB000264 COG0220 (R)    putative S-adenosylmethionine-dependent methyltransferase
REB000267 COG0212 (H)    putative 5-formyltetrahydrofolate cyclo-ligase
REB000281 COG0486 (R)    tRNA modification GTPase trmE
REB000282 COG0706 (U)    hypothetical inner membrane protein
REB000284 COG1968 (V)    bacitracin resistance protein (putative undecaprenol kinase)
REB000285 COG0617 (J)    tRNA nucleotidyltransferase (EC 2.7.7.25)
REB000290 COG0108 (H)    3,4-dihydroxy-2-butanone 4-phosphate synthase
REB000295 COG0739 (M)    putative membrane protein
REB000297 COG0469 (G)    Pyruvate kinase II (EC 2.7.1.40)
REB000304 COG0847 (L)    Exodeoxyribonuclease X (EC 3.1.11.-)
REB000326 COG1466 (L)    DNA polymerase III, delta subunit (EC 2.7.7.7)
REB000328 COG0495 (J)    Leucyl-tRNA synthetase (EC 6.1.1.4)
REB000336 COG0319 (R)    conserved hypothetical protein
REB000338 COG0621 (J)    MIAB PROTEIN
REB000340 COG0756 (F)    Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)
REB000342 COG0461 (F)    Orotate phosphoribosyltransferase (EC 2.4.2.10)
REB000346 COG0194 (F)    Guanylate kinase (EC 2.7.4.8)
REB000362 COG0104 (F)    Adenylosuccinate synthetase (EC 6.3.4.4)
REB000364 COG0330 (O)    HflC protein (EC 3.4.-.-)
REB000365 COG0330 (O)    FtsH modulator
REB000386 COG0586 (S)    Putative DedA family, membrane protein
REB000392 COG0169 (E)    Shikimate 5-dehydrogenase (EC 1.1.1.25)
REB000393 COG0009 (J)    SUA5 protein
REB000397 COG0242 (J)    Peptide deformylase (EC 3.5.1.88)
REB000398 COG0223 (J)    Methionyl-tRNA formyltransferase (EC 2.1.2.9)
REB000406 COG0202 (K)    DNA-directed RNA polymerase alpha chain (EC 2.7.7.6)
REB000409 COG0201 (U)    Preprotein translocase secY subunit
REB000416 COG0466 (O)    ATP-dependent protease La (EC 3.4.21.53)
REB000417 COG0760 (O)    Peptidyl-prolyl cis-trans isomerase D (EC 5.2.1.8)
REB000424 COG1132 (V)    ATP-binding component of a transport system
REB000425 COG1132 (V)    Multidrug resistance-like ATP-binding protein
REB000473 COG0193 (J)    Peptidyl-tRNA hydrolase (EC 3.1.1.29)
REB000474 COG0012 (J)    Probable GTP-binding protein ychF
REB000520 COG0040 (E)    ATP phosphoribosyltransferase (EC 2.4.2.17)
REB000521 COG0141 (E)    Histidinol dehydrogenase (EC 1.1.1.23)
REB000522 COG0079 (E)    Histidinol-phosphate aminotransferase (EC 2.6.1.9)
REB000523 COG0131 (E)    imidazole glycerol phosphate dehydratase and histidinol-phosphate phosphatase
REB000524 COG0118 (E)   imidazole glycerol phosphate synthase holoenzyme
REB000525 COG0106 (E)   1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]
 imidazole-4-carboxamide isomerase (Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase)
REB000526 COG0107 (E)    Imidazole glycerol phosphate synthase subunit hisF (EC 4.1.3.-)
REB000527 COG0139 (E)    phosphoribosyl-amp cyclohydrolase; phosphoribosyl-ATP pyrophosphatas
REB000534 COG0362 (G)    6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44)
REB000543 COG0282 (C)    Acetate kinase (EC 2.7.2.1)
REB000544 COG0857 (R)    Phosphate acetyltransferase (EC 2.3.1.8)
REB000566 COG2918 (H)    Glutamate--cysteine ligase (EC 6.3.2.2)
REB000603 COG0172 (J)    Seryl-tRNA synthetase (EC 6.1.1.11)
REB000616 COG0128 (E)    3-phosphoshikimate 1-carboxyvinyltransferase (EC 2.5.1.19)
REB000617 COG0283 (F)    Cytidylate kinase (EC 2.7.4.14)
REB000620 COG0492 (O)    Thioredoxin reductase (EC 1.6.4.5)
REB000624 COG0361 (J)    Bacterial Protein Translation Initiation Factor 1 (IF-1)
REB000663 COG0691 (O)    SsrA (tmRNA)-binding protein
REB000669 COG0061 (G)    Probable inorganic polyphosphate/ATP-NAD kinase (EC 2.7.1.23)
REB000690 COG0354 (R)    conserved hypothetical protein
REB000705 COG2912 (S)    putative transcriptional regulator
REB000706 COG2890 (J)    putative methyltransferase
REB000707 COG0216 (J)    Bacterial Peptide Chain Release Factor 1 (RF-1)
REB000711 COG0462 (FE)   Ribose-phosphate pyrophosphokinase (EC 2.7.6.1)
REB000717 COG0066 (E)    3-isopropylmalate dehydratase small subunit
REB000718 COG0065 (E)    3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)
REB000719 COG0473 (CE)   3-isopropylmalate dehydrogenase (EC 1.1.1.85)
REB000720 COG0119 (E)    2-isopropylmalate synthase (EC 4.1.3.12)
REB000721 COG0028 (EH)   Acetolactate synthase isozyme III large subunit (EC 4.1.3.18)
REB000722 COG0440 (E)    acetolactate synthase III, small subunit
REB000725 COG0275 (M)    putative S-adenosyl methionine adenyltransferase
REB000727 COG0768 (M)    septum formation; penicillin-binding protein 3; peptidoglycan synthetase
REB000728 COG0769 (M)    UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13)
REB000729 COG0770 (M)    UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--
                          D-alanyl-D- alanyl ligase (EC 6.3.2.15)
REB000732 COG0771 (M)    UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9)
REB000734 COG0707 (M)    UDP-N-acetylglucosamine--N-acetylmuramyl-(Pentapeptide)
                         pyrophosphoryl-undecaprenol N-acetylglucosamine transferase (EC 2.4.1.-)
REB000735 COG0773 (M)    L-alanine adding enzyme, UDP-N-acetyl-muramate:alanine ligase
REB000736 COG1181 (M)    D-alanine--D-alanine ligase (EC 6.3.2.4)
REB000738 COG0849 (D)    Cell division protein ftsA
REB000753 COG0635 (H)    putative oxidoreductase
REB000757 COG0325 (R)    conserved hypothetical protein
REB000759 COG0816 (L)    hypothetical endonuclease
REB000760 COG1678 (K)    conserved hypothetical protein
REB000761 COG0189 (HJ)   Glutathione synthetase (EC 6.3.2.3)
REB000763 COG2356 (L)    DNA-specific endonuclease I
REB000773 COG1054 (R)    conserved hypothetical protein
REB000783 COG0728 (R)    putative virulence factor
REB000784 COG1530 (J)    Ribonuclease E (EC 3.1.4.-)
REB000785 COG0564 (J)    Ribosomal large subunit pseudouridine synthase C (EC 4.2.1.70)
REB000803 COG1253 (R)    putative transport protein
REB000823 COG2894 (D)    Septum site-determining protein minD
```

```
REB000824 COG0850 (D)    Cell division inhibitor MinC
REB000825 COG0138 (F)    phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase
REB000841 COG0086 (K)    DNA-directed RNA polymerase beta' chain (EC 2.7.7.6)
REB000842 COG0085 (K)    DNA-directed RNA polymerase beta chain (EC 2.7.7.6)
REB000848 COG0250 (K)    Transcription antitermination protein nusG
REB000854 COG2269 (J)    Putative lysyl-tRNA synthetase (EC 6.1.1.6)
REB000863 COG1509 (E)    conserved hypothetical protein
REB000865 COG0459 (O)    60 kDa chaperonin
REB000866 COG0234 (O)    10 kDa chaperonin GROES
REB000878 COG0539 (J)    30S ribosomal subunit protein S1
REB000895 COG0017 (J)    Asparaginyl-tRNA synthetase (EC 6.1.1.22)
REB000900 COG0594 (J)    Ribonuclease P protein component (EC 3.1.26.5)
REB000902 COG0593 (L)    Chromosomal replication initiator protein dnaA
REB000903 COG0592 (L)    DNA polymerase III, beta chain (EC 2.7.7.7)
REB000905 COG0187 (L)    DNA gyrase subunit B (EC 5.99.1.3)
REB000962 COG0445 (D)    Glucose inhibited division protein A
REB000965 COG0356 (C)    ATP synthase A chain (EC 3.6.3.14)
REB000967 COG0711 (C)    ATP synthase B chain (EC 3.6.3.14)
REB000968 COG0712 (C)    ATP synthase delta chain (EC 3.6.3.14)
REB000969 COG0056 (C)    ATP synthase alpha chain (EC 3.6.3.14)
REB000970 COG0224 (C)    F1-F0-type proton-ATPase subunit gamma
REB000971 COG0055 (C)    ATP synthase B chain (EC 3.6.3.14)
REB000972 COG0355 (C)    ATP synthase epsilon chain (EC 3.6.3.14)
REB000973 COG1207 (M)    GlmU protein; UDP-N-acetylglucosamine pyrophosphorylase;
                         Glucosamine-1-phosphate N-acetyltransferase
REB000974 COG0449 (M)    L-glutamine:D-fructose-6-phosphate aminotransferase
REB001012 COG0568 (K)    RNA polymerase sigma-70 factor
REB001013 COG0358 (L)    DNA primase (EC 2.7.7.-)
REB001014 COG0828 (J)    30S ribosomal subunit protein S21
REB001045 COG0501 (O)    probable protease htpX (Heat shock protein htpX)
REB001074 COG0685 (E)    5,10-methylenetetrahydrofolate reductase (EC 1.7.99.5)
REB001080 COG0165 (E)    Argininosuccinate lyase (EC 4.3.2.1)
REB001100 COG0796 (M)    Glutamate racemase (EC 5.1.1.3)
REB001103 COG2925 (L)    Exodeoxyribonuclease I (EC 3.1.11.1)
REB001150 COG0327 (S)    conserved hypothetical protein
REB001159 COG0567 (C)    2-oxoglutarate dehydrogenase E1 component (EC 1.2.4.2)
REB001160 COG0508 (C)    Dihydrolipoamide succinyltransferase component (E2) of 2-oxoglutarate
                         dehydrogenase complex (EC 2.3.1.61)
REB001176 COG2885 (M)    Peptidoglycan-associated lipoprotein
REB001190 COG1158 (K)    Transcription termination factor rho
REB001191 COG0526 (OC)   Thioredoxin 1
REB001194 COG0210 (L)    ATP-dependent DNA helicase rep (EC 3.6.1.-)
REB001196 COG0059 (EH)   Ketol-acid reductoisomerase (EC 1.1.1.86)
REB001199 COG0129 (EG)   Dihydroxy-acid dehydratase (EC 4.2.1.9)
REB001206 COG0261 (J)    50S ribosomal protein L21
REB001209 COG0536 (R)    GTP-binding protein
REB001215 COG0293 (J)    cell division protein methyltransferase FtsJ
REB001216 COG0465 (O)    Cell division protein ftsH (EC 3.4.24.-)
REB001218 COG1109 (G)    phosphoglucosamine mutase
REB001226 COG0137 (E)    Argininosuccinate synthase (EC 6.3.4.5)
REB001228 COG0195 (K)    transcription pausing; L factor
REB001229 COG0532 (J)    Bacterial Protein Translation Initiation Factor 2 (IF-2)
REB001231 COG0130 (J)    tRNA pseudouridine synthase B (EC 4.2.1.70)
REB001233 COG1185 (J)    Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8)
REB001235 COG0513 (LKJ)  ATP-dependent RNA helicase (dead-box protein)
REB001247 COG0525 (J)    Valyl-tRNA synthetase (EC 6.1.1.9)
REB001249 COG0260 (E)    Cytosol aminopeptidase (EC 3.4.11.1)
REB001266 COG1286 (R)    Colicin V production protein
REB001268 COG0285 (H)    FolC bifunctional protein; Folylpolyglutamate synthase; Dihydrofolate synthase
REB001271 COG0101 (J)    tRNA pseudouridine synthase A (EC 4.2.1.70)
REB001276 COG0304 (IQ)   3-oxoacyl-[acyl-carrier-protein] synthase I (EC 2.3.1.41)
REB001282 COG0082 (E)    Chorismate synthase (EC 4.6.1.4)
REB001284 COG2840 (S)    conserved hypothetical protein
REB001295 COG1965 (P)    CyaY protein
REB001297 COG0253 (E)    Diaminopimelate epimerase (EC 5.1.1.7)
REB001309 COG0561 (R)    hypothetical hydrolase
REB001312 COG0620 (E)    5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14)
REB001348 COG1862 (U)    uncharacterized secreted proteins
REB001355 COG1985 (H)    bifunctional pyrimidine deaminase/reductase
REB001356 COG0054 (H)    riboflavin synthase beta chain
REB001357 COG0781 (K)    N utilization substance protein B
REB001368 COG0109 (O)    cytochrome o ubiquinol oxidase C subunit
REB001370 COG1845 (C)    Cytochrome O ubiquinol oxidase subunit III (EC 1.10.3.-)
REB001373 COG0060 (J)    Isoleucyl-tRNA synthetase (EC 6.1.1.5)
REB001374 COG0597 (MU)   Lipoprotein signal peptidase (EC 3.4.23.36)
REB001376 COG0761 (IM)   Penicillin tolerance protein
REB001381 COG0289 (E)    Dihydrodipicolinate reductase (EC 1.3.1.26)
REB001382 COG0505 (EF)   Carbamoyl-phosphate synthase small chain (EC 6.3.5.5)
REB001383 COG0458 (EF)   Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)
REB001400 COG0262 (H)    Dihydrofolate reductase (EC 1.5.1.3)
REB001401 COG0639 (T)    Bis(5'-nucleosyl)-tetraphosphatase (symmetrical) (EC 3.6.1.41)
REB001403 COG0030 (J)    Dimethyladenosine transferase (EC 2.1.1.-)
REB001434 COG0503 (F)    Xanthine-guanine phosphoribosyltransferase (EC 2.4.2.22)
REB001448 COG0176 (G)    Transaldolase B (EC 2.2.1.2)
REB001452 COG0498 (E)    Threonine synthase (EC 4.2.99.2)
REB001453 COG0083 (E)    Homoserine kinase (EC 2.7.1.39)
REB001454 COG0527 (E)    Aspartokinase (EC 2.7.2.4) / Homoserine dehydrogenase (EC 1.1.1.3)
REB001471 COG0813 (F)    Purine nucleoside phosphorylase (EC 2.4.2.1)
REB001472 COG1015 (G)    Phosphopentomutase (EC 5.4.2.7)
REB001480 COG4108 (J)    Peptide chain release factor 3
REB001483 COG2813 (J)    putative methyltransferase
REB001486 COG0221 (C)    Inorganic pyrophosphatase (EC 3.6.1.1)
```

```
REB001495 COG0312 (R)    putative PmbA protein
REB001520 COG0251 (J)    conserved hypothetical protein
REB001526 COG0078 (E)    Ornithine carbamoyltransferase chain F (EC 2.1.3.3)
REB001574 COG2706 (G)    putative 3-carboxymuconate cyclase
REB001577 COG0161 (H)    Adenosylmethionine-8-amino-7-oxononanoate aminotransferase (EC 2.6.1.62)
REB001578 COG0502 (H)    Biotin synthase (EC 2.8.1.6)
REB001579 COG0156 (H)    8-amino-7-oxononanoate synthase (EC 2.3.1.47)
REB001580 COG0500 (QR)   Biotin synthesis protein bioC
REB001581 COG0132 (H)    Dethiobiotin synthetase (EC 6.3.3.3)
REB001612 COG0807 (H)    GTP cyclohydrolase II (EC 3.5.4.25)
REB001617 COG0616 (OU)   putative periplasmic serine protease
REB001626 COG0147 (EH)   Anthranilate synthase component I (EC 4.1.3.27)
REB001628 COG0134 (E)    N-(5-phosphoribosyl)anthranilate isomerase/indole-3-glycerolphosphate synthetase
REB001629 COG0133 (E)    Tryptophan synthase beta chain (EC 4.2.1.20)
REB001630 COG0159 (E)    Tryptophan synthase alpha chain (EC 4.2.1.20)
REB001633 COG0628 (R)    hypothetical membrane protein
REB001636 COG0722 (E)    3-deoxy-D-arabinoheptulosonate-7-phosphate synthase
REB001646 COG0776 (L)    Integration host factor alpha-subunit
REB001647 COG0072 (J)    Phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20)
REB001648 COG0016 (J)    Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20)
REB001649 COG0292 (J)    50S ribosomal protein L20
REB001716 COG0021 (G)    Transketolase 1 (EC 2.2.1.1)
REB001719 COG0126 (G)    Phosphoglycerate kinase (EC 2.7.2.3)
REB001721 COG0191 (G)    Fructose-bisphosphate aldolase (EC 4.1.2.13)
REB001722 COG0668 (M)    putative membrane protein
REB001726 COG0120 (G)    Ribose 5-phosphate isomerase (EC 5.3.1.6)
REB001786 COG4776 (K)    Exoribonuclease II (EC 3.1.13.1)
REB001796 COG0284 (F)    Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23)
REB001801 COG0717 (F)    Deoxycytidine triphosphate deaminase (EC 3.5.4.13)
REB001834 COG0143 (J)    Methionyl-tRNA synthetase (EC 6.1.1.10) / PROTEIN SECRETION CHAPERONIN CSAA
REB001840 COG0254 (J)    LSU ribosomal protein L31P
REB001844 COG5405 (O)    heat shock protein hslVU, proteasome-related peptidase subunit
REB001845 COG1220 (O)    heat shock protein hslVU, ATPase subunit, homologous to chaperones
REB001852 COG1018 (C)    Ferredoxin--NADP reductase (EC 1.18.1.2)
REB001854 COG0149 (G)    Triosephosphate isomerase (EC 5.3.1.1)
REB001860 COG0205 (G)    6-phosphofructokinase (EC 2.7.1.11)
REB001866 COG1045 (E)    Serine acetyltransferase (EC 2.3.1.30)
REB001870 COG0607 (P)    putative sulfurtransferase
REB001891 COG0267 (J)    LSU ribosomal protein L33P
REB001902 COG0692 (L)    Uracil-DNA glycosylase (EC 3.2.2.-)
REB001911 COG0481 (M)    Elongation factor Tu GTP binding domain
REB001912 COG0681 (U)    Signal peptidase I (EC 3.4.21.89)
REB001913 COG0571 (K)    Ribonuclease III (EC 3.1.26.3)
REB001914 COG1159 (R)    GTP-binding protein era
REB001917 COG0736 (I)    Holo-[acyl-carrier protein] synthase (EC 2.7.8.7)
REB001930 COG0112 (E)    Serine hydroxymethyltransferase (EC 2.1.2.1)
REB001938 COG0483 (G)    putative Inositol phosphatase
REB001942 COG0188 (L)    DNA gyrase subunit A (EC 5.99.1.3)
REB001944 COG0209 (F)    Ribonucleoside-diphosphate reductase alpha chain (EC 1.17.4.1)
REB001945 COG0208 (F)    Ribonucleoside-diphosphate reductase beta chain (EC 1.17.4.1)
REB001946 COG0633 (C)    putative ferredoxin
REB001967 COG1007 (C)    NADH-quinone oxidoreductase chain N (EC 1.6.5.3)
REB001970 COG1009 (CP)   NADH dehydrogenase I chain L (EC 1.6.5.3)
REB001972 COG0713 (C)    NADH-quinone oxidoreductase chain K (EC 1.6.5.3)
REB001973 COG0839 (C)    NADH-quinone oxidoreductase chain J (EC 1.6.5.3)
REB001974 COG1143 (C)    NADH-quinone oxidoreductase chain I (EC 1.6.5.3)
REB001975 COG1005 (C)    NADH-quinone oxidoreductase chain H (EC 1.6.5.3)
REB001976 COG1034 (C)    NADH-QUINONE OXIDOREDUCTASE CHAIN G (EC 1.6.5.3)
REB001977 COG1894 (C)    NADH-QUINONE OXIDOREDUCTASE CHAIN F (EC 1.6.5.3)
REB001978 COG1905 (C)    NADH-QUINONE OXIDOREDUCTASE CHAIN E (EC 1.6.5.3)
REB001981 COG0377 (C)    NADH-quinone oxidoreductase chain B (EC 1.6.5.3)
REB001982 COG0838 (C)    NADH-quinone oxidoreductase chain A (EC 1.6.5.3)
REB002033 COG2812 (L)    DNA polymerase III subunit gamma/tau (EC 2.7.7.7)
REB002034 COG0718 (S)    conserved hypothetical protein
REB002037 COG0326 (O)    Chaperone protein htpG
REB002090 COG0313 (R)    putative methyltransferase
REB002112 COG0861 (P)    putative tellurium resistance protein terC
REB002135 COG0766 (M)    UDP-N-acetylglucosamine 1-carboxyvinyltransferase (EC 2.5.1.7)
REB002136 COG5007 (K)    conserved hypothetical protein
REB002162 COG0103 (J)    SSU ribosomal protein S9P
REB002163 COG0102 (J)    LSU ribosomal protein L13P
REB002175 COG0312 (R)    TldD protein
REB002239 COG0648 (L)    Endonuclease IV (EC 3.1.21.2)
REB002242 COG4668 (G)    PTS system, fructose-specific IIA/FPr component (EC 2.7.1.69)
REB002256 COG0741 (M)    Membrane-bound lytic murein transglycosylase E (EC 3.2.1.-)
REB002318 COG0162 (J)    Tyrosyl-tRNA synthetase (EC 6.1.1.1)
REB002322 COG0177 (L)    Endonuclease III (EC 4.2.99.18)
REB002323 COG4660 (C)    putative electron transport complex protein rnfE
REB002324 COG4659 (C)    putative electron transport complex protein rnfG
REB002325 COG4658 (C)    Electron transport complex protein rnfD
REB002326 COG4656 (C)    hypothetical membrane-bound Fe-S protein
REB002327 COG2878 (C)    putative ferredoxin 2
REB002328 COG4657 (C)    putative NADH-quinone oxidoreductase, subunit 5
REB002380 COG0629 (L)    Single-strand binding protein
REB002388 COG1197 (LK)   Transcription-repair coupling factor
REB002402 COG0015 (F)    Adenylosuccinate lyase (EC 4.3.2.2)
REB002403 COG2915 (R)    conserved hypothetical protein
REB002404 COG0482 (J)    hypothetical ATPase
REB002418 COG0057 (G)    Glyceraldehyde 3-phosphate dehydrogenase A (EC 1.2.1.12)
REB002437 COG0018 (J)    Arginyl-tRNA synthetase (EC 6.1.1.19)
REB002445 COG0173 (J)    Aspartyl-tRNA synthetase (EC 6.1.1.12)
REB002451 COG1108 (P)    High-affinity zinc uptake system membrane protein znuB
```

```
REB002452 COG1121 (P)    High-affinity zinc uptake system ATP-binding protein znuC
REB002469 COG0166 (G)    Glucose-6-phosphate isomerase (EC 5.3.1.9)
REB002489 COG0305 (L)    Replicative DNA helicase (EC 3.6.1.-)
REB002593 COG0320 (H)    Lipoic acid synthetase
REB002597 COG1278 (K)    Cold shock-like protein cspE
REB002660 COG0190 (H)    FolD bifunctional protein
REB002662 COG0215 (J)    Cysteinyl-tRNA synthetase (EC 6.1.1.16)
REB002666 COG1104 (E)    putative Cysteine sulfinate desulfinase/cysteine desulfurase
REB002667 COG0822 (C)    putative NifU protein
REB002668 COG0316 (S)    conserved hypothetical protein
REB002669 COG1076 (O)    Chaperone protein hscB
REB002670 COG0443 (O)    Chaperone protein hscA
REB002671 COG0633 (C)    Ferredoxin, 2Fe-2S
REB002677 COG0820 (R)    putative Fe-S-cluster redox protein
REB002681 COG0124 (J)    Histidyl-tRNA synthetase (EC 6.1.1.21)
REB002682 COG2976 (S)    conserved hypothetical protein
REB002684 COG1160 (R)    putative GTPase
REB002706 COG0329 (EM)   Dihydrodipicolinate synthase (EC 4.2.1.52)
REB002713 COG0624 (E)    putative N-succinyl-diaminopimelate deacylase
REB002732 COG2190 (G)    PTS system, glucose-specific IIA component (EC 2.7.1.69)
REB002735 COG1925 (G)    PTS system protein HPr
REB002743 COG0249 (L)    DNA mismatch repair protein mutS
REB002767 COG0148 (G)    Enolase (EC 4.2.1.11)
REB002768 COG0504 (F)    CTP synthase (EC 6.3.4.2)
REB002786 COG0258 (L)    Exodeoxyribonuclease IX (EC 3.1.11.-)
REB002796 COG0507 (L)    Exodeoxyribonuclease V alpha chain (EC 3.1.11.5)
REB002797 COG1074 (L)    Exodeoxyribonuclease V beta chain (EC 3.1.11.5)
REB002799 COG1330 (L)    Exodeoxyribonuclease V gamma chain (EC 3.1.11.5)
REB002803 COG0207 (F)    Thymidylate synthase (EC 2.1.1.45)
REB002804 COG0682 (M)    Prolipoprotein diacylglyceryl transferase (EC 2.4.99.-)
REB002888 COG0442 (J)    proline tRNA synthetase
REB002894 COG0037 (D)    cell cycle protein MesJ
REB002901 COG0587 (L)    DNA polymerase III alpha subunit (EC 2.7.7.7)
REB002905 COG0764 (I)    (3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase (EC 4.2.1.-)
REB002912 COG0020 (I)    Undecaprenyl pyrophosphate synthetase (EC 2.5.1.31)
REB002914 COG0233 (J)    Ribosome Recycling Factor (RRF)
REB002915 COG0528 (F)    Uridylate kinase (EC 2.7.4.-)
REB002916 COG0264 (J)    elongation factor Ts
REB002917 COG0052 (J)    30S ribosomal protein S2
REB002918 COG0024 (J)    Methionine aminopeptidase (EC 3.4.11.18)
REB002920 COG2171 (E)    2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase (EC 2.3.1.117)
REB002923 COG0265 (O)    protease DO precursor, heat shock protein HtrA (EC 3.4.21.-)
REB002925 COG0775 (F)    MTA/SAH nucleosidase; 5'-methylthioadenosine nucleosidase
REB002928 COG0316 (S)    conserved hypothetical protein
REB002934 COG0744 (M)    Penicillin-binding protein 1B
REB002938 COG1734 (T)    DnaK suppressor protein
REB002939 COG0008 (J)    putative tRNA synthetase
REB002958 COG1249 (C)    Dihydrolipoamide dehydrogenase (EC 1.8.1.4)
REB002959 COG0508 (C)    Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12)
REB002960 COG2609 (C)    Pyruvate dehydrogenase E1 component (EC 1.2.4.1)
REB002977 COG0653 (U)    Preprotein translocase SecA subunit
REB003012 COG0589 (T)    Universal stress protein A
REB003014 COG0306 (P)    Low-affinity inorganic phosphate transporter 1
REB003018 COG0425 (O)    putative transcriptional regulator
REB003023 COG0742 (L)    METHYLTRANSFERASE (EC 2.1.1.-)
REB003024 COG0552 (U)    Signal recognition particle GTPase
REB003027 COG0568 (K)    RNA polymerase sigma-32 factor
REB003043 COG0136 (E)    Aspartate-semialdehyde dehydrogenase (EC 1.2.1.11)
REB003056 COG0316 (S)    conserved hypothetical protein
REB003058 COG0596 (R)    putative biotin biosynthesis protein BioH
REB003080 COG0703 (E)    Shikimate kinase (EC 2.7.1.71)
REB003081 COG0337 (E)    3-dehydroquinate synthase (EC 4.6.1.3)
REB003085 COG0036 (G)    Ribulose-phosphate 3-epimerase (EC 5.1.3.1)
REB003087 COG0180 (J)    Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
REB003093 COG0477 (GEPR)putative transport protein
REB003109 COG0545 (O)    Peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8)
REB003111 COG1553 (P)    putative protein involved in sulfur reduction
REB003112 COG2923 (P)    conserved hypothetical protein
REB003247 NOCOG          hypothetical membrane protein
REB003248 COG2917 (D)    putative intracellular septation protein
REB003249 COG1607 (I)    Acyl-CoA hydrolase (EC 3.1.2.20)
REB003253 COG1502 (I)     Cardiolipin synthetase (EC 2.7.8.-)
REB003261 COG2095 (U)     integral membrane protein
REB003407 COG0605 (P)    Superoxide dismutase [Mn] (EC 1.15.1.1)
REB003440 COG0752 (J)    Glycyl-tRNA synthetase alpha chain (EC 6.1.1.14)
REB003441 COG0751 (J)    Glycyl-tRNA synthetase beta chain (EC 6.1.1.14)
REB003465 COG0537 (FGR)  hypothetical hydrolase
REB003466 COG0236 (IQ)   Acyl carrier protein
REB003475 COG0590 (FJ)   putative deaminase
REB003506 COG1219 (O)    ATP-dependent clp protease ATP-binding subunit clpX
REB003547 COG0359 (J)    LSU ribosomal protein L9P
REB003548 COG0238 (J)    SSU ribosomal protein S18P
REB003550 COG0360 (J)    SSU ribosomal protein S6P
REB003564 COG2168 (P)    conserved hypothetical protein
REB003643 COG0563 (F)    Adenylate kinase (EC 2.7.4.3)
REB003647 COG0290 (J)    Bacterial Protein Translation Initiation Factor 3 (IF-3)
REB003686 COG0218 (R)    Probable GTP-binding protein engB
REB003774 COG0450 (O)    Alkyl hydroperoxide reductase C22 protein (EC 1.6.4.-)
REB003842 COG0588 (G)    Phosphoglycerate mutase 1 (EC 5.4.2.1)
REB003867 COG0533 (O)    Probable O-sialoglycoprotein endopeptidase (EC 3.4.24.57)
REB003888 COG0013 (J)    Alanyl-tRNA synthetase (EC 6.1.1.7)
REB003926 COG0443 (O)    Chaperone protein dnaK
```

```
REB003941 COG0740 (OU)   ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92)
REB004025 COG0812 (M)    UDP-N-acetylenolpyruvoylglucosamine reductase (EC 1.1.1.158)
REB004029 COG1949 (A)    Oligoribonuclease (EC 3.1.-.-)
REB004043 COG0480 (J)    Elongation factor G
REB004044 COG0050 (J)    Protein Translation Elongation Factor Tu (EF-TU)
REB004059 COG0847 (L)    Ribonuclease T (EC 3.1.13.-)
REB004079 COG0331 (I)    Malonyl-CoA-[acyl-carrier-protein] transacylase (EC 2.3.1.39)
REB004080 COG1028 (IQR)  3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)
REB004081 COG0564 (J)    Ribosomal large subunit pseudouridine synthase D (EC 4.2.1.70)
REB004170 COG0328 (L)    RNase HI
REB004173 COG0491 (R)    Hydroxyacylglutathione hydrolase (EC 3.1.2.6)
REB004192 COG1825 (J)    50S ribosomal protein L25
REB004218 COG0231 (J)    Elongation factor P
REB004228 COG0048 (J)    30S ribosomal protein S12
REB004229 COG0049 (J)    30S ribosomal subunit protein S7
REB004247 COG0278 (O)    conserved hypothetical protein
REB004272 COG0333 (J)    LSU ribosomal protein L32P
REB004275 COG1190 (J)    Lysyl-tRNA synthetase (EC 6.1.1.6)
REB004291 COG0211 (J)    LSU ribosomal protein L27P
REB004293 COG0782 (K)    transcription elongation factor GreA (partial)
REB004296 COG0858 (J)    Ribosome-binding factor A
REB004297 COG0184 (J)    SSU ribosomal protein S15P
REB004338 COG0268 (J)    30S ribosomal protein S20
REB004342 COG0206 (D)    Cell division protein ftsZ
REB004365 COG0636 (C)    membrane-bound ATP synthase, F0 sector, subunit c
REB004400 COG1551 (T)    carbon storage regulator; controls glycogen
                         synthesis, gluconeogenesis, cell size and surface properties
REB004493 COG1214 (O)     Inactive homologe of metal-dependent protease
REB004547 COG0228 (J)     30S ribosomal protein S16
REB004682 COG4535 (P)     haemolysin-related protein
REB004753 COG0716 (C)     Flavodoxin
REB004767 COG1932 (HE)    Phosphoserine aminotransferase (EC 2.6.1.52)
REB004778 COG0335 (J)     50S ribosomal subunit protein L19
REB004780 COG0806 (J)     16S rRNA processing protein rimM
REB004813 COG1008 (C)     NADH-quinone oxidoreductase chain M (EC 1.6.5.3)
REB004814 COG0649 (C)     NADH-quinone oxidoreductase chain D (EC 1.6.5.3)
REB004889 COG0230 (J)     50S ribosomal protein L34
REB004902 COG0472 (M)     Phospho-N-acetylmuramoyl-pentapeptide-transferase (EC 2.7.8.13)
REB004903 COG0772 (D)     Cell division protein ftsW
REB004904 COG0516 (F)     GMP reductase (EC 1.6.6.8)
REB004987 COG0088 (J)     50S ribosomal protein L4
```

Tabelle A.1.: COG Gruppen

Gruppe	Funktionen
	INFORMATIONSSPEICHERUNG UND -VERARBEITUNG
J	Translation, ribosomale Struktur und Biogenese
A	RNA-Verarbeitung und -Modifikation
K	Transkription
L	Replikation, Rekombination und Reparatur
B	Chromatin Struktur und Dynamik
	ZELLULÄRE PROZESSE UND SIGNALVERARBEITUNG
D	Zelluläre Kontrolle, Zellteilung Chromosomenaufteilung
Y	Kernstrukturen
V	Verteidigungsmechanismen
T	Signaltransduktionsmechanismen
M	Zellwand-/Membran-/Hüllenbiogenese
N	Zellbeweglichkeit
Z	Cytoskelett
W	Extrazelluläre Strukturen
U	Intrazellulärer Verkehr, Sekretion und vesikulärer Transport
O	Posttranslationelle Modifikation, Proteinumsatz, Chaperone
	METABOLISMUS
C	Energieproduktion und -umwandlung
G	Kohlenhydratmetabolismus
E	Aminosäuretransport und -metabolismus
F	Nukleotidtransport und -metabolismus
H	Coenzymtransport und -metabolismus
I	Lipidtransport und -metabolismus
P	Transport und Metabolismus anorganischer Ionen
Q	Sekundärmetabolit Biosynthese, Transport und Katabolismus
	DÜRFTIG CHARAKTERISIERT
R	Generelle Funktionsvorhersage
S	Funktion unbekannt

a

Danksagung

Mein besonderer Dank gilt Prof. Dr. G. Gottschalk für die Überlassung des Themas und sein anhaltendes großes Interesse an der Arbeit und deren Fortgang. Er hat es mir ermöglicht, selbständig wissenschaftlich zu arbeiten und dabei noch über den Tellerrand der eigenen Forschung zu blicken. Die Teilnahme an nationalen und internationalen Tagungen waren nur ein Teil seiner großen Unterstützung. Vielen Dank!

Herrn Dr. Rainer Merkl danke ich für seine große Unterstützung vor und während der Doktorarbeit, für die fruchtbaren wissenschaftlichen Diskussionen und dafür, daß er immer ein offenes Ohr für die kleinen und großen Probleme während meiner Zeit als Doktorand hatte. Außerdem danke ich Rainer dafür, daß er mit viel Einsatz die Beschaffung neuer Hardware vorangetrieben hat, was es mir ermöglichte neben der Doktorarbeit andere interessante Forschungsbereiche zu erschließen. Der Dank für die konstruktive Durchsicht dieser Arbeit kann gar nicht groß genug ausfallen.

Frau Dr. Anke Henne gilt mein Dank für die erfolgreiche Einführung in das weite Feld der Genom-Sequenzierung. Dafür, daß sie gegenüber neuen „Methoden“ zum schließen von Lücken immer sehr aufgeschlossen war, auch wenn die Aussicht auf Erfolg oft gering erschien. Vielen Dank für die Zeit, die wir bei dem einen oder anderen Kaffee bei den Robotern zugebracht haben („Hilfe, Hilfe, Hilfe...“). Letztendlich sei Anke noch für ihre hervorrangende Arbeit bei der Durchsicht dieses Textes gedankt.

PD Dr. Rolf Daniel danke ich für die Übernahme der Korreferenz und seine gute Zusammenarbeit bei der Sequenzierung von *„blattae“*. Die eingebrachte personelle Unterstützung war sehr hilfreich.

Vielen Dank schulde ich unseren TAs („die Mädels“), auf deren Unterstützung und ausgezeichnete Arbeit in Laborangelegenheiten ich mich jeder Zeit verlassen konnte. Danke Iwona, Silke D., Silke S., Mechthild und Monika!

Danken möchte ich auch allen ehemaligen und aktuellen Mitgliedern des Göttinger Genomlabors für eine gute Arbeitsatmosphäre: Axel Strittmatter, Frank Hoster, Bata (Sebastian) Bäumer (alter ERGOnaut), Jarek Sobkowiak („Olli hat angerufen!“), Yvonne Gläser („Dienstreiseantrag gestellt?“), Kirsten Hahne, Heiko Liesegang („Heiko, wie siehts mit dem Update aus?“), Rosa Martinez-Arias, Holger Brüggemann, Florian Fricke, Elzbieta Brzuszkiewicz und Carsten Raasch.

Dem Werkstatt-Team gilt mein Dank für die gute technische Unterstützung, besonders Olaf, der mich oft in die Geheimnisse der Reparatur oder Entsorgung von Geräten eingeweiht hat.

Meiner kleinen Schwester Anneke (Assessorin des Lehramts an Gymnasien) danke ich dafür, daß (!) sie mich aus den Fängen der neuen deutschen Rechtschreibung befreit hat - knappes Ding!

Meinen Eltern danke ich für ihre Unterstützung während meines Studiums und insbesondere während meiner Doktorarbeit. Sie haben mir sehr viel ermöglicht.

Lebenslauf

- Geboren am 6.7.1974 in Göttingen, deutscher Staatsbürger
- Eltern: Regina Wiezer und Dr. Eckhard Wiezer
- 1980 - 1984 Grundschule Ehrentrup / Lage
- 1984 - 1987 Realschule der Stadt Lage
- 1987 - 1990 Gymnasium der Stadt Lage
- 1990 - 1991 Felix-Klein-Gymnasium Göttingen
- 1991 - 1994 Johanneum Lüneburg
- 14.6.1994 Erlangen der Allgmeinen Hochschulreife am Johanneum Lüneburg
- August 1994 - Oktober 1995 Zivildienst beim Paritätischen Wohlfahrtsverband Lüneburg
- Oktober 1995 Immatrikulation an der Georg-August-Universität Göttingen im Studiengang Biologie (Diplom)
- 16.10.1997 Vordiplom in den Fächern Mikrobiologie, Zoologie, Chemie und Physikalische Chemie
- 5.5.2000 Diplomprüfung in den Fächern Mikrobiologie, Humangenetik und Biochemie
- 22.5.2000 - 27.4.2001 Experimentelle Diplomarbeit unter Prof. Dr. W. Liebl mit dem Thema: „Untersuchung möglicher Kompetenzgene in *Thermus thermophilus*“
- Juni 2001 Beginn der Arbeit zur vorliegenden Dissertation