Komparative Genomanalyse zur Stammoptimierung produktionsnaher *Bacillus*-Stämme

Dissertation

zur Erlangung des

mathematisch-naturwissenschaftlichen

Doktorgrades „Dr. rerum naturalium“ an der

Georg-August-Universität Göttingen

vorgelegt von

Antje Wollherr

aus Peine

Göttingen 2010

Referent: …… Prof. Dr. Wolfgang Liebl
Korreferent: …….....Prof. Dr. Burkhard Morgenstern
Tag der mündlichen Abschlussprüfung: ……...26.10.2010

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

A	Adenin
AA	Aminosäure (*amino acid)*
BBH	bidirektionaler bester Hit
BH	bidirektionaler Hit
BLAST	*basic local alignment search tool*
bp	Basenpaare
bzw.	beziehungsweise
C	Cytosin
COG	*cluster of orthologous genes*
CSF	*competence sporulation factor*
DB	Datenbank
DNA	Desoxyribonukleinsäure
EBI	*European Bioinformatics Institute*
EMBL-Datei	*European Molecular Biology Laboratory* - Dateiformat
ER-Modell	*Entity-Relationship*-Modell
et al.	und andere
etc.	*et cetera*
G	Guanin
ggf.	gegebenenfalls
kb	Kilo-Basenpaare
mRNA	*messenger*-Ribonukleinsäure
Nr.	Nummer
NW	Needleman-Wunsch
ORF	*open reading frame*
RNA	Ribonukleinsäure
sog.	sogenannte
T	Thymin
tRNA	*transfer*-Ribonukleinsäure
u. a.	unter anderem
z. B.	zum Beispiel

1 Einleitung

Bioinformatik ist eine junge, interdisziplinäre Wissenschaft, die informatische Methoden auf biologische Probleme anwendet. Diese Arbeit wird sich auf ein Teilgebiet der Mikrobiologie als bioinformatisches Anwendungsgebiet fokussieren. Mit der Sequenzierung des Genoms, also der Bestimmung der vollständigen Erbinformation eines Organismus, ergeben sich vielfältige bioinformatische Aufgaben, um die Datenmengen zu speichern, verwalten und zu analysieren.

1995 wurde der Grippeerreger *Haemophilus influenzae* als erstes bakterielles Genom entschlüsselt (Fleischmann *et al.*, 1995). 2002 konnte das menschliche Genom entschlüsselt werden (Lander *et al.*, 2001; Venter, 2001). Durch die Weiterentwicklungen von der klassischen Sanger-Sequenzierung (Sanger *et al.*, 1992) zu den Next-Generation-Sequencing-Technologien (Mardis, 2008) ist es in immer kürzerer Zeit möglich, immer mehr Genome zu sequenzieren und damit Datenmengen im *terrabyte*-Bereich (Richter & Sexton, 2009) zu produzieren. Bis heute sind mehr als 1000 Organismen sequenziert worden (Liolios *et al.*, 2010).

Die Sequenzierung liefert jedoch zunächst nur die Genomsequenz. Weitere experimentelle, aber auch bioinformatische Analyseschritte, wie das *ORF finding* und die Annotation sind nötig um Kenntnisse über biologisch relevante Eigenschaften, die von der DNA-Sequenz kodiert werden (Gene, Promotoren, Terminatoren, RNAs etc.) zu erlangen. Beim der Genvorhersage ist das Ziel aus den offenen Leserahmen (*open reading frames*) diejenigen in der DNA-Sequenz zu finden, die möglicherweise für ein Protein kodieren können. Die funktionale Annotation ist die Zuweisung einer Funktion für die vorhergesagten Proteine über Homologievergleiche (Bork *et al.*, 1998) zu bereits bekannten, experimentell gut untersuchten Proteinen. Selbst in nah verwandten Organismen können aber nicht alle Proteine auf diese Weise annotiert werden, da es Gene gibt, die für bisher unbekannte Proteine (sog. hypothetische Proteine) kodieren. *ORF finding* ist eine bioinformatische Vorhersagemethode, die keine 100 % sichere Aussage zur tatsächlichen biologischen Funktion des Gens / Proteins in der Zelle treffen kann. Allerdings liefert sie gut begründete Annahmen, die im Labor verifiziert werden können.

Aus den vorhandenen Proteinen und regulatorischen Elementen ergeben sich alle Fähigkeiten des Organismus wie zum Beispiel die Energiegewinnung durch die

Verwertung bestimmter Nährstoffe (z. B. Zucker, Aminosäuren), die Anpassung an Nährstoffmangelbedingungen oder extreme Standorte als Lebensraum, aber auch um genetische Informationen weiterzugeben (Madigan & Martinko, 2006).

Gerade in den unbekannten Proteinen, die je nach Spezies 30 % (Bork, 2000) und mehr (Fraser *et al.*, 2000) ausmachen können, steckt ein großes Potential (Galperin & Koonin, 2004), das hilft, die metabolischen Stoffwechselwege zu verstehen. Mit dieser Kenntnis wird es möglich, gezielt Stoffwechselwege anzusteuern und den Organismus für eine Fähigkeit zu optimieren.

1.1 Motivation

Bacillus licheniformis ist ein nicht-pathogenes, gram-positives Bodenbakterium, für dessen Typstamm DSM13 die Genomsequenz seit 2004 verfügbar ist (Veith *et al.*, 2004). Aufgrund seiner Fähigkeit bis zu 25 Gramm Protein (Amylasen, Proteasen) pro Liter zu sekretieren hat es eine hohe wirtschaftliche Bedeutung (Schallmey *et al.*, 2004). Industriell wird *B. licheniformis* u. a. in der Waschmittelindustrie eingesetzt. Im Vergleich zum Modellorganismus *B. subtilis* 168 (Kunst *et al.*, 1997), der für den Laboreinsatz optimiert ist, hat *B. licheniformis* DSM13 den Nachteil, schwer genetisch manipulierbar zu sein.

Die Genomsequenzen von zahlreichen *Bacillus*-Stämmen liegen vor (Chen *et al.*, 2007; Gioia *et al.*, 2007; Kunst *et al.*, 1997), so dass vergleichende Analysen (Binnewies *et al.*, 2006) Aufschluss darüber geben können, warum *B. licheniformis* DSM13 im Vergleich zu den bekannten kompetenten Vertretern der *Bacillus*-Gruppe schwer transformierbar ist. Genetische Zugänglichkeit ist eine Grundvoraussetzung, um den Organismus im Labor verändern zu können. Die gezielte genetische Manipulation erlaubt die Erstellung von Mutanten, die für einen bestimmten Zweck (höhere Produktraten, verlängerter Produktionsprozess, bessere morphologische Eigenschaften, etc.) optimiert sind.

Erste experimentelle Ansätze zur Erhöhung der Transformationsraten von *B. licheniformis* DSM13 wurden durch die Deletion zweier Typ1-Restriktionssysteme erreicht (Waschkau *et al.*, 2008). Im Vergleich zu *B. subtilis* 168 sind die Transformationsraten aber dennoch gering.

Um die vergleichenden Analysen durchzuführen, sind bioinformatische Werkzeuge (*tools*) nötig, die eine einfache Auswertung und sinnvolle Visualisierung der Datenmenge erlauben. Für die Zuordnung der bekannten Proteine sind Analysen mit bereits bekannten und experimentell bestätigten Proteinen notwendig. Klassische Algorithmen wie die standardmäßig eingesetzte BLAST-Analyse (Altschul *et al.*, 1990) reichen dafür häufig nicht aus bzw. sind schwer zu interpretieren.

Die vorliegende Arbeit ist in ein Kooperationsprojekt zwischen Industrie und Universität eingebettet. Neben dieser bioinformatisch orientierten Doktorarbeit, beschäftigen sich zwei biologische Doktorarbeiten mit der Methodenentwicklung zur gezielten genetischen Veränderung und der Erstellung chromosomweiter Mutanten von *B. licheniformis* DSM13. Im Laufe des Projektes werden zahlreiche biologische Experimente wie die Expressionsanalyse von *B. licheniformis* DSM13 und die Transposonmutagenese durchgeführt. Außerdem entstehen zahlreiche Mutanten, die charakterisiert und allen Projektbeteiligten in Form einer Datenbank zur Verfügung gestellt werden sollen. Die Entwicklung eines geeigneten Datenbankschemas und dessen Implementierung ist notwendig, um die unterschiedlichen Datentypen strukturiert speichern und abfragen zu können.

1.2 Problemstellung und Lösungsansatz

Das Ziel der vorliegenden Arbeit lässt sich in drei Teilbereiche gliedern. Zunächst sollen auf Basis der vorhanden Genomsequenzen von *B. subtilis* 168 (Kunst *et al.*, 1997), *B. amyloliquefaciens* FZB42 (Chen *et al.*, 2007), *B. pumilus* SAFR-032 (Gioia *et al.*, 2007) und *B. anthracis str.* Sterne (EMBL *accession* Nummer: AE017225) komparative Analysen durchgeführt werden, deren Ziel die Analyse des Kompetenzsystems von *B. licheniformis* DSM13 (Veith *et al.*, 2004) ist. Diese Fragestellung lässt sich mit Hilfe der vorhandenen Genomsequenzen in Form von EMBL-Dateien vom *European Bioinformatics Institute* (EBI, http://www.ebi.ac.uk/) lösen, die nicht nur die chromosomale Lokalisation der Proteine, sondern auch deren funktionellen Annotationen enthalten.

Auf Basis der gesammelten Erfahrungen bei diesem nicht-automatisierten Ansatz, soll ein benutzerfreundliches Programm entwickelt werden mit dessen Hilfe diese komparative Analyse automatisiert werden kann. Neben der einfachen Handhab-

barkeit ist auch die sinnvolle Darstellung der Ergebnisse ein zu berücksichtigendes Kriterium. Da BLAST-Vergleiche (Altschul *et al.*, 1990) nur lokale Ähnlichkeiten berechnen, sollen zur Validierung auch globale Ähnlichkeiten (*similarities)* mit einbezogen werden. Eine BLAST-Analyse gefolgt vom globalen Needleman-Wunsch-Algorithmus (Needleman & Wunsch, 1970) erlaubt dies. Die Einzelproteininformationen können über vergleichende Analysen von benachbarten, konservierten Proteinen zu einer *cluster*-basierten Gesamtgenomsicht erweitert werden. Mit Hilfe des entwickelten *tools* sind auch Bereiche in anderen verwandten Stämmen zu identifizieren, die den Genpool von *B. licheniformis* DSM13 erweitern und im Idealfall die natürliche Kompetenz wiederherstellen oder die Sekretionsleistung erhöhen können.

Das *tool* soll für umfassende komparative Analysen eingesetzt werden, um Bereiche in *B. licheniformis* DSM13 zu identifizieren, die zum Leben des Organismus essentiell sind. Darüber hinaus sollen Bereiche identifiziert werden, die auch im Sinne einer hohen Sekretionsleistung nicht-essentiell sind. Die Deletion dieser Bereiche erlaubt es möglicherweise, die Sekretionsraten nachhaltig zu erhöhen (Morimoto *et al.*, 2008).

Eine Analyse der im Projekt entstehenden Daten erfordert den systematischen Aufbau einer Datenbank vom Datenmodellierungs-Modell (ER-Modell) über das Datenbankschema zu einem funktionalen Datenbanksystem.

1.3 Aufbau der Arbeit

Die Arbeit ist neben diesem Einleitungskapitel in einen Grundlagenabschnitt und drei Hauptkapitel gegliedert, die jeweils einen eigenen Material- und Methodenteil sowie Ergebnisteil haben. Die abschließende Diskussion und Zusammenfassung gibt dann einen Gesamtblick über die Arbeit auch im Kontext der aktuellen Forschung.

2 Grundlagen

Das Grundlagenkapitel liefert das notwendige Basiswissen aus Biologie und Informatik, um die weitere Arbeit verstehen zu können.

Der Informationsfluss innerhalb eines Organismus basiert auf der DNA, als Träger der Erbinformationen. Vereinfacht ausgedrückt, wird DNA in mRNA transkribiert, die dann in ein Protein translatiert wird. Zunächst werden in Kapitel 2.1 die einzelnen biologischen Bausteine der DNA und Prozesse zur Proteinbiosynthese näher beschrieben. Kapitel 2.1 basiert, soweit nicht anders angegeben, auf dem Lehrbuch „Molekulare Genetik“ (Knippers, 2006). Anschließend (Kapitel 2.2) wird die Organismengruppe der *Bacilli* vorgestellt, die den Anwendungsschwerpunkt der nachfolgenden Analysen bildet. Kapitel 2.2 basiert, sofern nicht anders angegeben, auf „*Bacillus subtilis* and it‘s closest relatives: from genes to cells“ (Sonenshine *et al.*, 2001). Eine Eigenschaft einiger Vertreter der *Bacillus*-Spezies ist die Fähigkeit, freie DNA aus der Umgebung aufzunehmen. Kapitel 2.3 beschreibt die zugrundeliegende Regulationskaskade.

In Kapitel 2.4 wird die Verbindung zwischen biologischer Sequenz und bioinformatischen Vergleichsmethoden hergestellt, so dass in Kapitel 2.5 einige bekannte Ansätze zur Identifizierung von Orthologen vorgestellt werden. Kapitel 2.4 basiert, sofern nicht anders angegeben auf: „Bioinformatics: Sequence and Genome Analysis“ (Mount, 2004).

2.1 Vom Gen zum Protein

Gene sind Bereiche auf der DNA, die die Information zur Herstellung eines Proteins oder einer funktionellen DNA enthalten. Prokaryotische Gene sind im Vergleich zu eukaryotischen Genen einfach aufgebaut. Jedes Gen beinhaltet eine protein-kodierende Region, die *open reading frame* (ORF) genannt wird. Der ORF beginnt meistens mit der Basensequenz ATG und endet mit einem Stopp-Codon. ORF’s sind unterschiedlich lang und haben verschiedene Sequenzen. Allerdings ist nicht jeder ORF ein Gen, das in ein Protein translatiert werden kann.

Im 5’-Bereich des ORFs befindet sich der Promoter mit der TATA-Box, die die Konsensussequenz TATA trägt. Der Promoter ist die Bindungsstelle für die RNA-Polymerase, die essentiell für die Transkription ist.

Bei der Transkription lagert sich die RNA-Polymerase an den Promoter an und entspiralisiert die doppelsträngige DNA. Dadurch können am codogenen Strang komplementäre Nukleotide eingebaut werden, die die mRNA bilden. Erreicht die RNA-Polymerase den Terminator, wird die Transkription beendet.

In pro- und eukaryotischen Organismen gibt es verschiedene RNAs, die unterschiedliche Funktionen erfüllen. Die *messenger*-RNA (mRNA) ist das Produkt der Transkription eines Gens und dient als Vorlage zur Translation in ein Protein. *Messenger*-RNAs können aus hundert bis zu einigen tausend Nukleotiden bestehen und entsprechen etwa 10 % der Gesamt-RNA von *Escherichia coli*.

Ribosomen sind Makromoleküle, an denen die Proteinbiosynthese, also die Translation einer mRNA in ein Protein stattfindet. Die mRNA wird durch das Ribosom geschleust und jeweils tripletweise abgelesen. Daraus ergibt sich der genetische *Code*. Ein Nukleotid-Triplet (Codon) entspricht einer Aminosäure. Es gibt 64 Möglichkeiten Triplets aus den vier Basen Adenin (A), Cytosin (C), Guanin (G) und Thymin (D) zu bilden, aber normalerweise nur 20 Aminosäuren. Das bedeutet, dass mehrere Codons für die gleiche Aminosäure kodieren.

Bei der Translation wird die Aminosäure von einer *transfer*-RNA (tRNA) zum Synthesort transportiert. tRNAs aus *B. licheniformis* DSM13 bestehen aus 72-93 Nukleotiden (Veith *et al.*, 2004) im Vergleich dazu haben tRNAs in *E. coli* 74-94 Basen. Eine Kleeblatt-Struktur ist charakteristisch für tRNAs. Wird beim Ablesen der mRNA das Stopp-Codon erreicht, ist die Proteinsynthese beendet.

Proteine sind die Grundbausteine jeder Zelle, die vielfältige Funktionen haben, wie z. B. der Zelle Struktur zu geben, chemische Prozesse zu katalysieren oder Stoffe in die Zelle herein oder heraus zu transportieren.

Prokaryotische Genome bestehen größtenteils aus kodierenden Sequenzen. Charakteristisch für Bakteriengenome sind Promotoren, deren Transkripte oft mehrere Gene umfassen. Diese Genbereiche werden polycistronisch genannt. Wird ein Transkript aus einem Gen gebildet, spricht man von monocistronischer mRNA.

2.2 Die *Bacilli*

Bacilli sind gram-positive Bakterien, die in der Lage sind Sporen zu bilden. Sie kommen ubiquitär vor und können aus Boden, Luft und Wasser isoliert werden. Gemeinsam ist ihnen die Fähigkeit, polymere Substrate wie Stärke oder Proteine mittels sekretierter Enzyme zu verwerten. Es gibt human-pathogene Vertreter, wie z. B. *B. anthracis* oder Insektenpathogene wie *B. thuringiensis* (Schnepf *et al.*, 1998)8). Die *Bacillus subtilis* - Subgruppe besteht allerdings nur aus nicht pathogenen *Bacilli*, wie u. a. *B. subtilis*, *B. amyloliquefaciens* oder den industriell relevanten *B. licheniformis*-Stämmen. Anzumerken ist, dass die Stämme *B. licheniformis* DSM13 und *B. licheniformis* ATTC14580 isogenisch sind und lediglich aus unterschiedlichen Stammzellsammlungen stammen.

Bacillus-Stämme zeichnen sich generell durch einen niedrigen GC-Gehalt von 35-47 % aus.

Als Kohlenstoffquelle nutzen *Bacilli* vor allem Glukose. Sauerstoff dient der Energiegewinnung durch Zellatmung. Unter Nährstoffmangelbedingungen gibt es ein ausgeklügeltes System, das vielfältig auf den Stress reagieren kann. So besteht für *B. subtilis* die Möglichkeit, sich zu nährstoffreicheren Orten zu bewegen, denn *B. subtilis* ist begeißelt. Außerdem kann *B. subtilis* sehr umweltresistente Sporen bilden (Setlow, 2006), die erst unter besseren Nährstoffbedingungen wieder zu vegetativen Zellen werden. Die Aufnahme von DNA über natürliche Kompetenz ist ein weiterer Mechanismus, um auf schlechte Umweltbedingungen zu reagieren, da die damit verbundene mögliche Ausbildung neuer Fähigkeiten zur Nischenadaption führen kann.

Die Eigenschaft der natürlichen Kompetenz hat *B. subtilis* (Spizizen, 1958) zu einem „Arbeitstier“ der Molekularbiologie gemacht.

2.3 Natürliche Kompetenz in Bakterien

Die Fähigkeit eines Bakteriums freie DNA aus der Umgebung aufzunehmen, wird als natürliche Kompetenz (Griffith, 1928) bezeichnet. Die aufgenommene DNA führt in der Zelle zu Rekombinationsereignissen, durch die die DNA ganz oder teilweise ins bakterielle Chromosom integriert wird. Natürliche Kompetenz er-

möglicht einem Bakterium somit die Reparatur und auch Erweiterung des eigenen Genpools.

Erste bakteriologische Untersuchungen zur natürlichen Kompetenz wurden von Griffiths 1928 durchgeführt. Griffith arbeitete mit *Streptococcus pneumoniae*. Diese gram-positiven Bakterien sind tier- und humanpathogen (Musher, 1992). Sie besitzen eine Kapsel, die es dem Immunsystem des Wirtsorganismus unmöglich macht, die Bakterien zu töten. Stämme mit Kapsel sind folglich virulent und werden als S-Stämme bezeichnet. Den avirulenten R-Stämmen fehlt die Kapsel, die sie vor dem Wirtsimmunsystem schützt. Griffith zeigte, dass die getrennten Injektionen von hitzegetöteten S-Zellen und von lebenden R-Zellen in unterschiedliche Mäuse nicht tödlich verlaufen. Die Injektion beider Bakterientypen zusammen töteten jedoch die Mäuse. Aus den toten Mäusen konnten lebende S-Stämme isoliert werden. Mit diesem Versuch wurde gezeigt, dass ein genetischer Austausch stattgefunden hat, der es den lebenden R-Stämmen erlaubte, die genetische Information zur Ausbildung der Kapsel aufzunehmen. Die R-Stämme transformierten zu S-Stämmen.

In weiteren Experimenten konnte die Gruppe um Avery belegen, dass die Transformation auch in vitro durchgeführt werden kann (Avery *et al.*, 1944). Sehr aufwendige Analysen des verwendeten biologischen Materials lieferten den Beweis, dass DNA die fundamentale Einheit bei der Transformation ist (Avery *et al.*, 1944).

Natürliche Kompetenz ist für viele Bakterienstämme sowohl gram-positiver Spezies, wie *Bacillus* (Dubnau, 1991a) und *Streptococcus* (Havarstein *et al.*, 1997), als auch für gram-negative Spezies, wie *Campylobacter* (Nedenskovsorensen *et al.*, 1990), *Helicobacter* (Hofreuter *et al*., 1998) beschrieben. Die molekularbiologische Bedeutung von natürlicher oder induzierbarer Kompetenz ist sehr hoch, weil damit die Handhabung der Stämme zur genetischen Manipulation im Labor erleichtert wird.

Viele Bakterien sind zumeist in einer bestimmten Wachstumsphase natürlich kompetent (Lorenz & Wackernagel, 1994). Bedingt durch äußere Einflüssen, wie z. B. Nährstoffmangel, wird die Fähigkeit zur Transformation aber nur bei einigen

Zellen der Population ausgebildet. Die Aufnahme von DNA findet dann zelldichte-abhängig statt. Allerdings nimmt ein Bakterium nicht wahllos DNA auf, sondern meistens aus nah verwandten Stämmen. DNA aus entfernt verwandten Organismen wird als Fremd-DNA erkannt und von den Nukleasen zerstört.

Die freie DNA aus der Umgebung lagert sich an die Zellwand des Bakteriums an und wird als Einzelstrang durch eine Pore in der Zellmembran in das Bakterium hinein transportiert. Rekombinatorische Ereignisse innerhalb der Zelle führen dann dazu, dass die DNA in das Bakterienchromosom integriert werden kann.

Für *B. subtilis* 168 ist der Vorgang der natürlichen Kompetenz sehr gut untersucht und beschrieben (Dubnau, 1991a; Dubnau, 1991b) (Abbildung 1). Die Aufnahme von DNA ist an das allgemeine Stressantwort-System gekoppelt und abhängig von Wachstumsphase, Zelldichte und Zelltyp, da nicht alle Zellen einer Population kompetent werden.

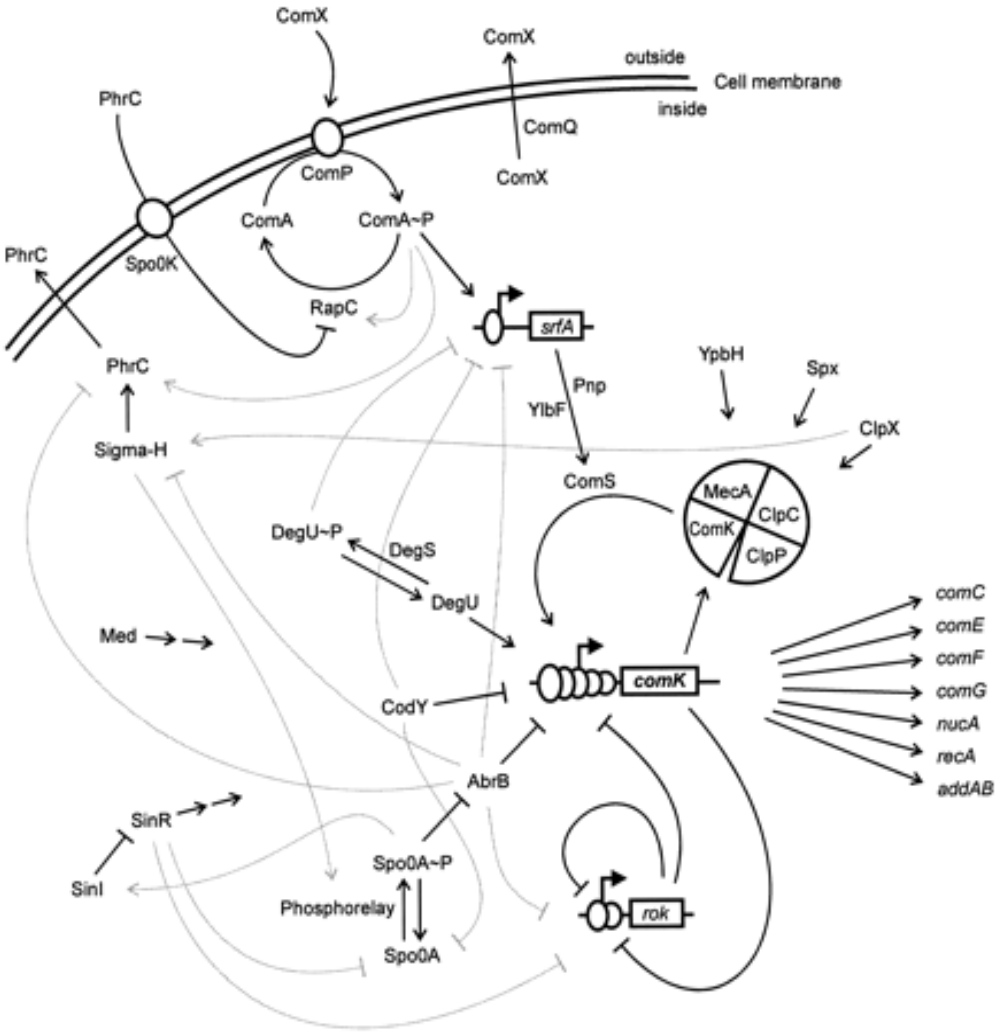


Abbildung 1: Regulationskaskade der natürlichen Kompetenz in *B. subtilis* 168
Dargestellt ist der Mechanismus der natürlichen Kompetenzbildung von den frühen *quorum sensing* Genen des *comQXPA*-Clusters über ComS zum *Mec Switch* mit dem gebundenen Schlüsselregulator ComK. Die späten, strukturellen Kompetenzgene werden von ComK aktiviert.(entnommen aus (Hamoen *et al.*, 2003b))

Mit *quorum sensing* wird ein Zell-Zell-Komunikationsmechanismus zur Reaktion auf die Anwesenheit gleichartiger Populationen in der direkten Umgebung bezeichnet (Bassler & Losick, 2006). In *B. subtilis* 168 ist die Ausbildung der natürlichen Kompetenz an einen *quorum sensing*-Mechanismus gekoppelt. Der Start der Regulationskaskade ist chromosomal im *comQXPA*-Gencluster kodiert (Ansaldi & Dubnau, 2004). ComQ katalysiert die Modifizierung von ComX (Schneider *et al.*, 2002). Das modifizierte ComX wird aus der Zelle in die Umgebung abgegeben und wirkt dort als Pheromon. ComP ist eine Histidin-Kinase, die das Pheromon binden kann und so die Phosphorylierung von ComA bewirkt.

Phosphoryliertes ComA ist für den weiteren Verlauf der Kompetenzregulationskaskade essentiell (Nakano & Zuber, 1991).

Ein zweiter *quorum sensing*-Pfad schließt sich bei der Phosphorylierung von ComA an (Abbildung 2). ComA~P wird von RapC dephosphoryliert. RapC wird von niedrigen extrazellulären Konzentrationen des *competence sporulation factor* (CSF) inhibiert. Pre-CSF ist ein Protein, das in der Zelle durch Transkription und Translation von *phrC* entsteht und dann aus der Zelle ausgeschleust wird. Außerhalb der Zelle wird es geschnitten, so dass CSF entsteht.

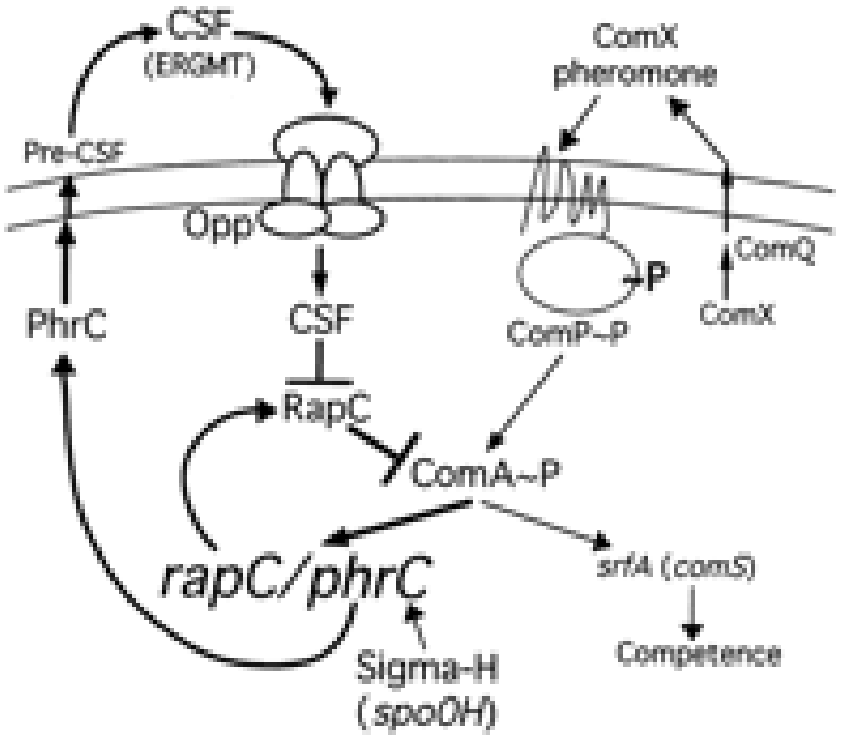


Abbildung 2: Zwei *quorum sensing*-Module, die in die Kompetenzbildung in *B. subtilis* 168 involviert sind
Dargestellt sind beide pheromon-regulierten Pfade, die an der Phosphorylierung von ComA beteiligt sind. Zum einen handelt es sich um das Pheromon ComX, das über das *comQXPA*-Cluster gebildet wird und zum anderen um den *competence sporulation factor*, der von *rapC* / *phrC* gebildet wird. (entnommen aus (Pottathil & Lazazzera, 2003))

Da ComA~P für die Ausbildung der Kompetenz essentiell ist, sind niedrige Konzentrationen von CSF notwendig. Hohe extrazelluläre Konzentrationen von CSF führen jedoch zu einer Inhibierung der Kompetenzausbildung und fördern stattdessen die Sporulation (Pottathil & Lazazzera, 2003).

Phosphoryliertes ComA wirkt als Co-Faktor, damit die RNA-Polymerase an den Promoter für *srfA* binden kann (Nakano & Zuber, 1991). Die Expression von *srfA*

führt gleichzeitig zur Expression von *comS* (Dsouza *et al*., 1995; Hamoen *et al.*, 1995). *comS* ist ein kleines Gen, das innerhalb des Surfactin-Gens lokalisiert ist und zusammen mit *srfA* transkribiert wird.

ComS wirkt als entscheidender Faktor im *Mec Switch* (van Sinderen & Venema, 1994; Vansinderen & Venema, 1994). Um unangebrachte Induktion der Kompetenz in der exponentiellen Phase zu verhindern, bindet der Schlüsselregulator ComK quantitativ in einem ternären Komplex aus MecA, ClpC und ClpP (*Mec Switch*). ComK wird in diesem Komplex von der ClpP-Protease degradiert. Die Bindung von MecA und ClpC im Komplex erhöht die Bindungsaffinität für ComK. ComS bindet kompetetiv zu ComK ebenfalls an MecA, so dass ComK aus dem Komplex freigelassen werden kann, wenn *comS* exprimiert wird.

ComK ist autoreguliert und wirkt auf den eigenen Promoter. Durch die Expression von *comK* wird die Expression der strukturellen Transformationsgene (Hamoen *et al.*, 2003b) initiiert. Die späten Kompetenzgene, die DNA an der Zellwand binden und die eine Pore zur Aufnahme der DNA bilden, sind im *comG*-Operon kodiert (Abbildung 3).

In der Zelle finden rekombinatorische Ereignisse statt, die eine Integration der aufgenommenen DNA-Fragmente in das Chromosom bewirken.

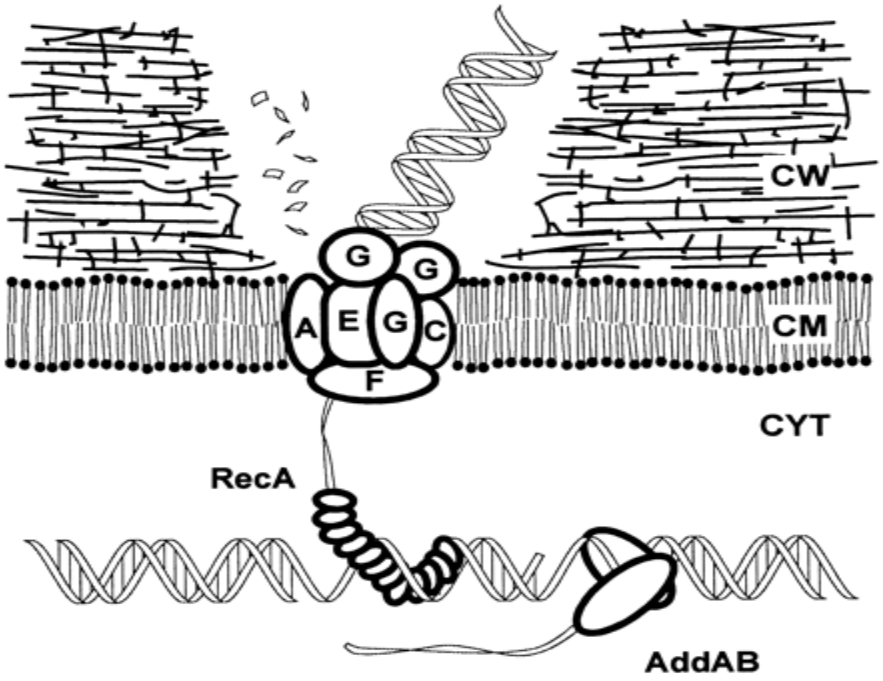


Abbildung 3: Funktionsweise des ComG-Proteinkomplexes (entnommen aus (Hamoen *et al.*, 2003b))
Eine Zelle ist mit der Zellwand (CW), der Zellmembran (CM) und dem Cytoplasma (CYT) dargestellt. Der ComG-Proteinkomplex, der in der Zellmembran lokalisiert ist (ComGA, ComGC, ComGE, Teile von ComGG), weist außen- (Teile von von ComGG)– sowie innenständige (ComGF) Bereiche auf. Doppelsträngige DNA wird über ComG in die Zelle geschleust und dort als einzelsträngige DNA von RecA und AddAB wieder zu doppelsträngiger DNA synthetisiert.

2.4 Sequenzbasierte Vergleichsmethoden

Unter biologischer Sequenz wird die Basenabfolge einer DNA-Sequenz bzw. die Aminosäureabfolge eines Proteins (Kapitel 2.2) verstanden. Um bioinformatisch mit diesen Sequenzen arbeiten zu können, werden die Basen und Aminosäuren ungeachtet ihrer komplexen chemischen *backbone*- und Verknüpfungsstrukturen als Buchstaben in einem Alphabet definiert. Das DNA-Alphabet ist folgendermaßen aufgebaut: D = {A, C, G, T} und repräsentiert die einzelnen Basen während das Aminosäurealphabet aus den 20 kanonischen Aminosäuren sowie ggf. den Selenocysteinen bzw. Pyrolysinen aufgebaut ist.

Mit dieser Abstraktion wird es möglich, bioinformatische Analysen, wie zum Beispiel *alignments* zum direkten Vergleich zweier oder mehrerer Sequenzen oder *pattern matching* zur Suche bestimmter Muster in Sequenzen durchzuführen.

2.4.1 BLAST

Basic local alignment tool (BLAST) (Altschul *et al.*, 1990) ist das Standardwerkzeug in der Biologie, um Sequenzdatenbanken nach ähnlichen Sequenzen zu

durchsuchen. Der Erfolg ist in der einfachen Handhabbarkeit und schnellen, sehr effizienten Suche begründet. Dies bringt den Nachteil mit sich, dass nur lokale Treffer gefunden werden können. Lokal bedeutet, dass nicht die gesamte Sequenz in das *alignment* mit einbezogen werden muss, sondern hochähnliche Sequenzabschnitte ausreichen um einen Treffer zu liefern.

Die Funktionsweise von BLAST basiert auf der Suche nach kurzen, hochkonservierten Sequenzabschnitten, sog. *seeds*, die keine Lücken aufweisen. Für jeden *seed* innerhalb der Datenbank wird versucht, ihn nach links und rechts zu erweitern, so dass die Bewertung des produzierten *alignments* unter einem Schwellenwert bleibt. Die gefunden Treffer werden *high-scoring segment pair* genannt.

BLAST liefert aber nicht nur einen Treffer mit dem zugehörigen *alignment*, sondern auch eine statistisch abgesicherte Bewertung des Treffers in Abhängigkeit von der durchsuchten Datenbank. Dafür werden zwei Werte berechnet: der *bit score* und der *expectation value* (*e-value*). Mit dem *bit score* wird das *alignment* in Bezug auf Ähnlichkeiten und Lücken bewertet. Je höher der *bit score* ist, desto besser der Treffer.

Der *e-value* liefert eine statistische Signifikanz für den Treffer in Abhängigkeit der Datenbank-Größe.

Der *e-value* E wird über folgende Formel berechnet:

$$E = Kmn\, e^{-\lambda S} \qquad (1)$$

Die Parameter K und lambda repräsentieren natürliche Skalare für den Suchraum und das Bewertungssystem. S entspricht dem *bit score*. m steht für die Länge der Suchsequenz und n für die Größe der Datenbank.

Der *raw bit score* S‘ wird über folgende Formel berechnet:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad (2)$$

Diese Bewertungsmöglichkeiten geben Aufschluss darüber, ob der produzierte Treffer durch Zufall entstanden ist oder eine statistische Beziehung reflektiert.

Die Bewertung der *alignments* erfolgt für DNA-Sequenzen typischerweise auf Basis der „*point accepted mutation*" – Matrix (PAM) (Schwartz & Dayhoff, 1978) oder „BLOcks SUbstitution Matrix" (BLOSUM) (Henikoff & Henikoff, 1992). Der Unterschied liegt in der Erstellung der Matrizen. PAM's basieren auf statistischen Beobachtungen über die Häufigkeit von Basenaustauschen nah verwandter Organismen. BLOSUM-Matrizen hingegen basieren auf einer breiteren Datenbasis, der sog. BLOCKS-Datenbank. BLOCKS enthält ~3000 Blöcke von kurzen, hochkonservierten Sequenzen aus 800 Gruppen verwandter Proteine. Damit wird deutlich, dass der Vorteil von PAM in der Spezialisierung auf nah verwandte Organismen liegt und BLOSUM biologisch richtige Ähnlichkeiten eher hervorhebt.

2.4.2 Needleman-Wunsch-Algorithmus

Der Needleman-Wunsch-Algorithmus (Needleman & Wunsch, 1970) bietet die Möglichkeit, globale *alignments* zweier Sequenzen über die gesamte Länge der Sequenzen zu berechnen. Basis ist ein *dynamic programming*-Ansatz, der aus drei Schritten besteht. Zunächst wird eine Tabelle aufgebaut, die aus n+1 Spalten und m+1 Reihen besteht. n entspricht der Länge der ersten Sequenz und m der Länge der zweiten Sequenz. Während der Initialisierung werden die 0. Reihe und die 0. Spalte jeweils auf 0 gesetzt, unter der Voraussetzung, dass keine Lücken zugelassen sind.

Im zweiten Schritt wird die Tabelle M basierend auf folgenden Rekurrenzen aus den vorherigen Zeilen und Spalten ausgefüllt:

$$M_{i,j} = \mathrm{MAXIMUM}[\; M_{i-1,\,j-1} + S_{i,j}\ (1),\; M_{i,j-1} + w\ (2),\; M_{i-1,j} + w\ (3)\;]$$

(1) entspricht einem *match* / *missmatch* in der Diagonalen, $S_{i,j}$ ist die Bewertungsfunktion für *missmatches*

(2) entspricht einer *gap* in Sequenz 1

(3) entspricht einer *gap* in Sequenz 2

Im letzten Schritt wird ein *traceback* beginnend von der Zelle M(n,m) durchgeführt. Dabei werden die Diagonalen bevorzugt, denen ein *match* / *missmatch* zugrunde liegt. Ansonsten werden horizontale oder vertikale Schritte gemacht, bei denen sich der *score* nicht ändert. In diesem Fall werden Lücken in das *alignment* eingefügt. Wird die Zelle M(0,0) erreicht, ist ein optimales *alignment* zwischen den beiden zu vergleichenden Sequenzen gefunden.

Diese Grundidee wurde insofern verfeinert, dass *gap*-Kosten für das Öffnen einer Lücke und für die Erweiterung einer Lücke angeben werden können. Meistens ist es sinnvoll das Öffnen einer Lücke deutlich höher zu bestrafen, als das Verlängern einer Lücke. Für den Needleman-Wunsch-Algorithmus, der im EMBOSS-*package* (Rice *et al.*, 2000) implementiert ist, werden wie beim BLAST (Altschul *et al.*, 1990) auch Bewertungsmatrizen wie EBLOSUM62 für Proteine und EDNAFULL für DNA-Sequenzen verwendet.

Die Bewertung des resultierenden optimalen *alignments* basiert dann auf der *identity* und der *similarity*. Mit der *identity* wird angeben, wie viele identische Treffer es zwischen den beiden Sequenzen gibt. Die *similarity* hingegen gibt Auskunft darüber, wie viele *matches* es gibt. Die *identity* weist immer einen niedrigeren oder gleichen Wert auf im Vergleich zur *similarity*.

2.5 Bekannte Ansätze zur Bestimmung von Orthologen

In diesem Kapitel werden zunächst grundsätzliche Begriffe erklärt, die im Zusammenhang mit komparativen Analysen stehen (Kapitel 2.5.1). Anschließend werden einige Ansätze zur Bestimmung von Orthologen vorgestellt (Kapitel 2.5.2 - 2.5.4).

2.5.1 Homologe im Kontext *pan* und *core* genomes

Homologe sind Proteine, die Ähnlichkeiten zueinander haben, unabhängig von ihrer Evolution und biologischen Funktion. Orthologe und Paraloge werden unter dem Begriff Homologe zusammengefasst.

Orthologe sind Gene bzw. Proteine, die in unterschiedlichen Organismengruppen vorkommen und ähnliche oder die gleiche, biologische Funktionen haben, da sie

durch ein evolutionäres Speziationsereignis entstanden sind (Fitch, 1970). Daher besitzen sie auch ähnliche (homologe) Sequenzen.

Abzugrenzen sind sie von Paralogen, die Genduplikationen innerhalb eines Organismus entsprechen und deshalb ähnliche (homologe) Sequenzen haben. Die Genduplikationen können gleiche Funktionen haben.

Das *core genome* ist die Menge aller Gene, die in allen Stämmen einer Spezies vorkommen (Medini *et al.*, 2005). Möglicherweise bildet das *core genome* die biologische Basis und den charakteristischen Phänotyp einer Spezies (Tettelin *et al.*, 2008). Das *pan genome* schließt das *core genome* sowie alle weiteren Gene ein, die in mindestens einem Stamm der Spezies vorkommen (Medini *et al.*, 2005). Es trägt zur Spezies-Spezifität bei und dient möglicherweise der Nischenadaption (Tettelin *et al.*, 2008).

2.5.2 Bidirektionaler (bester) *hit*

Das Konzept, das bidirektionalen besten *hits* (Overbeek *et al.*, 1999) zugrunde liegt ist, dass Orthologe ähnlicher zueinander sind, als zu anderen Genen bzw. Proteinen der verglichenen Organismen. Um einen bidirektionalen *hit* (BH) zu bestimmen sind zwei aufeinanderfolgende Analyseschritte notwendig. Zunächst wird eine multiple Sequenz mit einer anderen (multiple) Sequenz verglichen. Im zweiten Schritt wird der Vergleich in entgegengesetzter BLAST-Richtung wiederholt. Werden in der weiteren Auswertung nur die jeweils ersten, also besten *hits* verwendet, wird auch von einem bidirektionalen besten *hit* (BBH) gesprochen. Als Vergleichsalgorithmen bieten sich BLAST (Altschul *et al.*, 1990) oder beispielsweise FASTA (Pearson & Lipman, 1988) an.

2.5.3 Das ERGO^{TM}-System

Ein Datenbanksystem zur funktionalen Analyse und Annotation hauptsächlich mikrobieller Genome ist ERGO^{TM} (Overbeek *et al.*, 2003). Intern wird eine Erweiterung des bidirektionalen besten *hits* (BBH) verwendet, die auch Paare benachbarter, konservierter Proteine mit einbezieht (Overbeek *et al.*, 1999). Dadurch können komplexe Gencluster-Analysen (Kapitel 3.2.2) durchgeführt werden. ERGO verwendet standardmäßig nicht den BLAST-Algorithmus (Altschul *et al.*,

1990), sondern den langsameren, aber sensitiveren FASTA-Algorithmus (Pearson & Lipman, 1988).

2.5.4 Weitere Ansätze zur Orthologenbestimmung

Hulsen (Hulsen *et al.*, 2006) gibt einen Überblick zu unterschiedlichen Methoden der Orthologen-Identifizierung sowie eine eingehende Bewertung der einzelnen Ansätze. Untersucht werden BBH (Kapitel 2.5.2) , COG (Tatusov *et al.*, 2003), InParanoid (O'Brien *et al.*, 2005) und OrthoMCL (Li *et al.*, 2003). Der wesentliche Unterschied ist in der Anzahl der Orthologen zu sehen, die in die Analyse eingeschlossen werden. Beim BBH wird genau ein Ortholog betrachtet, wohingegen ein COG (*cluster of orthologous genes*) aus mehreren hundert Homologen bestehen kann. COG verwendet einen triangulären bidirektionalen besten *hit* als Orthologiekriterium, wodurch die deutlich höhere Anzahl erklärt wird. OrthoMCL verwendet einen *Graph-Clustering*-Algorithmus, der ebenfalls mehr als ein Ortholog mit einbeziehen kann. Obwohl InParanoid die besten Resultate nach Hulsen‘s Bewertungsmethode liefert, sind BBHs effektiv, wenn Selektivität (so wenig falsch-positive wie möglich) ein entscheidendes Kriterium ist.

3 Analyse des Kompetenzsystems in *B. licheniformis* DSM13

Der komparative Vergleich der Kompetenzsysteme verschiedener *Bacillus*-Stämme wurde ausgehend von *B. subtilis* 168 durchgeführt. In *B. subtilis* 168 konnten, basierend auf einer Literaturrecherche (Sonenshine *et al.*, 2001) 61 essentielle Gene identifiziert werden, die an der Kompetenzbildung beteiligt sind (Tabelle 1). Für die essentiellen Gene wurden die entsprechenden Orthologen in vier *Bacillus*-Stämmen gesucht.

Tabelle 1: Kompetenzgene in *B. subtilis* 168

Name / *locus tag*	**Produkt/ Funktion**	**Name / *locus tag***	**Produkt/Funktion**
abrB (BSU00370)	Transkriptionaler Übergangszustandsregulator (Hamoen *et al.*, 2003a)	*yqeZ* (BSU24660)	im ComG-Operon (Chung & Dubnau, 1998)
clpC /mecB (BSU00860)	Negativer Regulator der Kompetenz (Msadek *et al.*, 1994)	*comGG* (BSU24670)	DNA-Aufnahme (Chung & Dubnau, 1998)
sigH (BSU00980)	RNA-Polymerase Sigma-Faktor SigH (Weir *et al.*, 1991)	*comGF* (BSU24680)	DNA-Aufnahme (Chung & Dubnau, 1998)
nin / comJ (BSU03420)	Inhibitor der DNA-Degradierungsaktivität von NucA (Provvedi *et al.*, 2001)	*comGE* (BSU24690)	DNA-Aufnahme (Chung & Dubnau, 1998)
nucA /comI (BSU03430)	Membran-assoziierte Nuklease (Provvedi *et al.*, 2001)	*comGD* (BSU24700)	DNA-Aufnahme (Chung & Dubnau, 1998)
srfAA /comL (BSU03480)	Surfactin-Synthetase A (van Sinderen *et al.*, 1990)	*comGC* (BSU24710)	DNA-Aufnahme (Chung & Dubnau, 1998)
srfAB /comL (BSU03490)	Surfactin-Synthetase B (van Sinderen *et al.*, 1990)	*comGB* (BSU24720)	DNA-Aufnahme (Chung & Dubnau, 1998)
comS (BSU03500)	Anti-MecA-Adapterprotein (Hamoen *et al.*, 1995)	*comGA* (BSU24730)	DNA-Aufnahme (Chung & Dubnau, 1998)
srfAC /comL (BSU03510)	Surfactin-Synthetase C (van Sinderen *et al.*, 1990)	*ComEC* (BSU25570)	Extrazelluläre DNA-Bindung (Hahn *et al.*, 1993)
srfA /comL (BSU03520)	Surfactin-Synthetase D (van Sinderen *et al.*, 1990)	*ComEB* (BSU25580)	Extrazelluläre DNA-Bindung (Hahn *et al.*, 1993)
rapC (BSU03770)	Kontrolle der ComA-Aktivität (Lazazzera *et al.*, 1999)	*ComEA* (BSU25590)	Extrazelluläre DNA-Bindung (Hahn *et al.*, 1993)
phrC (BSU03780)	*rapC*-Inhibitor, CSF (Lazazzera *et al.*, 1999)	*comER* (BSU25600)	Regulator vom ComE-Operon (Hahn *et al.*, 1993)
sigB (BSU04730)	RNA-Polymerase Sigma-Faktor SigB (Binnie *et al.*, 1986)	*ComC* (BSU28070)	Prozessierung von ComGC (Chung & Dubnau, 1995)

Name / *locus tag*	**Produkt/ Funktion**	**Name / *locus tag***	**Produkt/Funktion**
comK (BSU10420)	Kompetenz-schlüsselregulator (van Sinderen & Venema, 1994)	*lonB* (BSU28210)	Involviert in die Vorsporenbil-dung (Serrano *et al.*, 2001)
med (BSU11300)	Positiver Regulator von ComK (Ogura *et al.*, 1997)	*clpX* (BSU28220)	ATP-abhängige Clp-Protease (Gerth *et al.*, 1996)
comZ /yjzA (BSU11310)	ComG Operon-Repressor (Ogura & Tanaka, 2000)	*yux /comAB* (BSU31670)	im ComQXPA-Operon (Tran *et al.*, 2000)
oppA (BSU11430)	Oligopeptid-bindendes Protein A (Perego *et al.*, 1991)	*comA* (BSU31680)	Regulation der Kompetenz (Nakano & Zuber, 1991)
oppB (BSU11440)	Oligopeptid-bindendes Protein B (Perego *et al.*, 1991)	*comP* (BSU31690)	Phosphorylierung von ComA (Piazza *et al.*, 1999)
oppC (BSU11450)	Oligopeptid-bindendes Protein C (Perego *et al.*, 1991)	*comX* (BSU31700)	Kompetenzpheromon-Precursor (Schneider *et al.*, 2002)
oppD (BSU11460)	Oligopeptid-bindendes Protein D (Perego *et al.*, 1991)	*comQ* (BSU31710)	schneidet ComX (Hahn *et al.*, 1993)
oppF (BSU11470)	Oligopeptid-bindendes Protein F (Perego *et al.*, 1991)	*degQ* (BSU31720)	Regulation der Exoenzymsynthese (Msadek *et al.*, 1991)
mecA (BSU11520)	Kontrolle der ComK-Degradation (Schlothauer *et al.*, 2003)	*slr / slrR* (BSU34380)	Transkriptionaler Regulator, Paralog von SinR (Kobayashi, 2008)
ylbF (BSU14990)	Kontrolle der ComK-Stabilität (Tortosa *et al.*, 2000)	*clpP* (BSU34540)	ATP-abhängige Clp-Protease (Gerth *et al.*, 1998)
smf/ dpra (BSU16110)	schützt hereinkommende, einzelsträngige DNA (Tadesse & Graumann, 2007)	*comFC* (BSU35450)	Spätes Kompetenzgen (Londonovallejo & Dubnau, 1993)
codY (BSU16170)	Transkriptionaler pleiotropischer Kompetenz-Repressor (Serror & Sonenshein, 1996)	*comFB* (BSU35460)	Spätes Kompetenzgen (Londonovallejo & Dubnau, 1993)
pnpA/ comR (BSU16690)	Nötig für die Kompetenz-entwicklung (Luttinger *et al.*, 1996)	*comFA* (BSU35460)	DNA-Bindeprotein (Londonovallejo & Dubnau, 1993)
cinA (BSU16930)	Induzierung einer Kompe-tenzschädigung (Kaimer & Graumann, 2010)	*yviA/ degV* (BSU35470)	Im ComF-Operon (Msadek *et al.*, 1991)
ymcA (BSU17020)	Regulator der Biofilmbildung (Kearns *et al.*, 2005)	*degU* (BSU35490)	Regulation der Kompetenz (Msadek *et al.*, 1991)
spo0A (BSU24220)	Downregulation von AbrB (Hahn *et al.*, 1995)	*degS* (BSU35500)	Regulation der Kompetenz (Msadek *et al.*, 1991)
sinI (BSU24600)	Antagonist von SinR (Bai *et al.*, 1993)	*ssb/ ssbA* (BSU40900)	Einzelstrang DNA-Bindeprotein (Lindner *et al.*, 2004)
sinR (BSU24610)	Regulator der postexponenti-ellen Genexpression (Bai *et al.*, 1993)		

Die Auswahl der Stämme erfolgte auf Basis beschriebener natürlicher bzw. induzierbarer Kompetenz sowie auf phylogenetischer Verteilung innerhalb der *Bacillus*-Gruppe. *Bacillus subtilis* 168 (Kunst *et al.*, 1997) und *Bacillus amyloliquefaciens* FZB42 (Chen *et al.*, 2007) gehören zu den nicht pathogenen *Bacilli* der *subtilis*-Gruppe und sind natürlich kompetent. *Bacillus licheniformis* DSM13 ist ebenfalls Mitglied der *subtilis*-Gruppe, aber nicht bzw. nur in geringem Maße kompetent (Veith *et al.*, 2004). *Bacillus pumilus*-Stämme wurden als chemisch induzierbar kompetent beschrieben (Droffner & Yamamoto, 1985) und sind phylogenetisch zwischen der *subtilis*-Gruppe sowie der *anthracis*-/ *cereus*-Gruppe einzuordnen. Als Vertreter der *anthracis*-/ *cereus*-Gruppe wurde der avirulente Stamm *Bacillus anthracis* str. Sterne ausgewählt, der chemisch induzierbar kompetent ist (Quinn & Dancer, 1990).

Abbildung 4 zeigt die phylogenetische Verwandtschaft basierend auf der 16S-rRNA der analysierten Stämme.

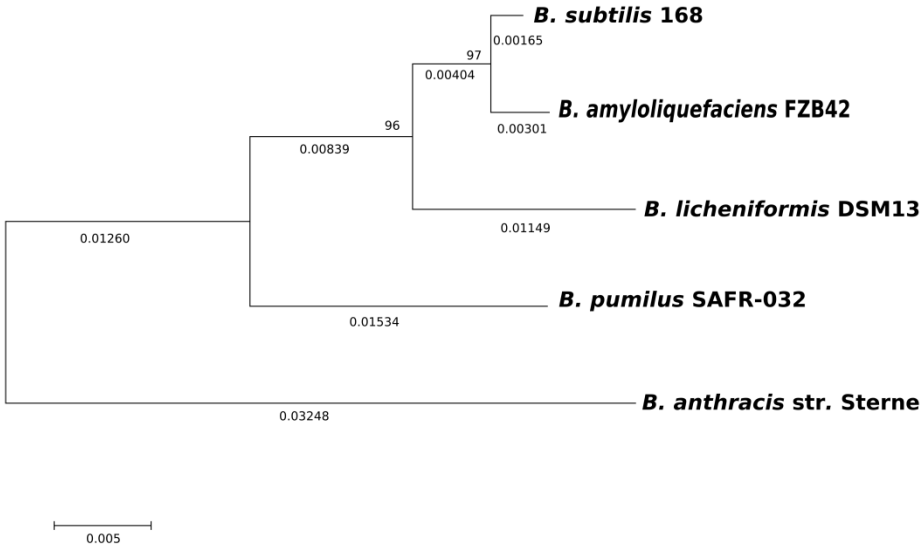


Abbildung 4: 16S-rRNA basierter phylogenetischer Stammbaum: Kompetenzsystemanalyse
Der horizontale Balken repräsentiert 0,005 Substitutionen pro Nukleotidposition

In einem zweiten Analyseschritt wurden das *quorum sensing*-Modul und der *Mec Switch* in der *subtilis*-Gruppe mit Schwerpunkt auf *Bacillus licheniformis*-Stämmen betrachtet. Interessant ist, dass sich die Stämme DSM13 (Veith *et al.*, 2004), ATTC14580 (Rey *et al.*, 2004) und F11 (Waldeck *et al.*, 2006) auf 16S-rRNA-Ebene nicht voneinander unterscheiden lassen. Der 16S-rRNA Stamm-

baum (Abbildung 5) zeigt die phylogenetische Verwandtschaft der analysierten Stämme.

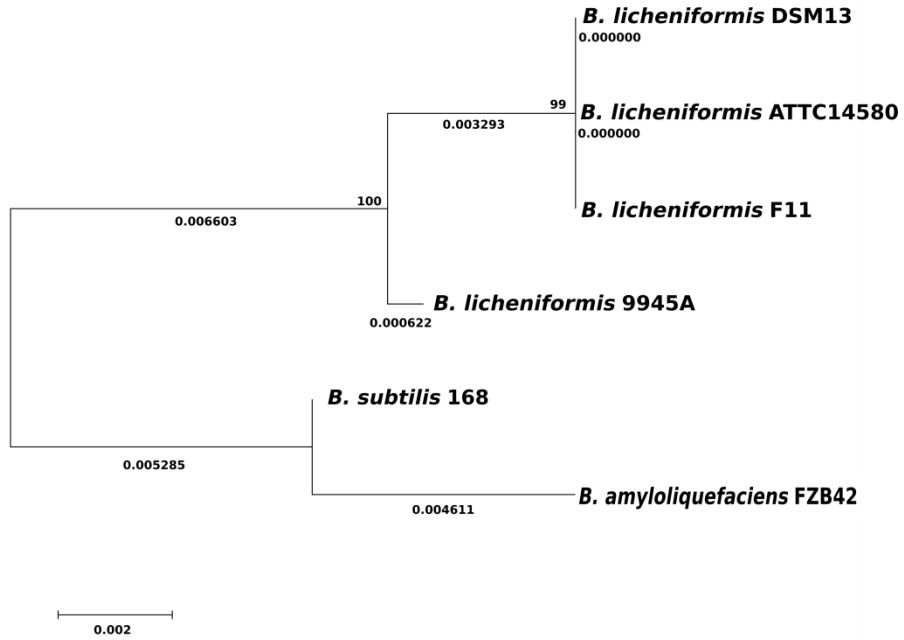


Abbildung 5: 16S-rRNA basierter phylogenetischer Baum: *quorum sensing*-Modul-Analyse
Der horizontale Balken repräsentiert 0.002 Austausche pro Nukleotidsequenz

3.1 Verwendete Daten

Alle Sequenzen und Genumgebungen wurden mit ERGO (Kapitel 2.5.3) identifiziert und den entsprechenden Genomen (Tabelle 2) entnommen. Die ERGO-Einträge der jeweiligen Organismen basieren auf den zugehörigen EMBL-Dateien, die mit den entsprechenden *accession*-Nummern angegeben werden. Folgende Organismen wurden für die Sequenz- und Genumgebungsanalysen verwendet: *Bacillus subtilis* subsp. *subtilis* str. 168, *Bacillus licheniformis* DSM13, *Bacillus amyloliquefaciens* FZB42, *Bacillus pumilus* SAFR-032S und *Bacillus anthracis* str. Sterne.

Für die detaillierte Analyse des Kompetenzregulationsmoduls wurden neben den bisherigen drei Stämmen der *subtilis*-Gruppe folgende weiteren *Bacillus licheniformis*-Stämme verwendet: *Bacillus licheniformis* 9945A, *Bacillus licheniformis* ATCC14580, *Bacillus licheniformis* F11 (Tabelle 2).

Tabelle 2: Übersicht der verwendeten Organismen in der Kompetenzsystemanalyse

Organismus	***accession*-Nr.**	**Kurzinfoformation**	**Referenz**
Bacillus subtilis subsp. *subtilis* str. 168	AL009126	Modellorganismus	(Kunst *et al.*, 1997)
Bacillus licheniformis DSM13	AE017333	industriell relevanter Typstamm	(Veith *et al.*, 2004)
Bacillus amyloliquefaciens FZB42	CP000560	pflanzenassoziierter Stamm	(Chen *et al.*, 2007)
Bacillus pumilus SAFR-032S	AE017225	sehr robuster Sporenbildner	(Gioia *et al.*, 2007)
Bacillus anthracis str. Sterne	CP000813	avirulenter *anthracis* Stamm	unpubliziert
Bacillus licheniformis 9945A	pers. Komm. M. Rachinger, Dissertation 2010	natürlich kompetenter *B. licheniformis*-Stamm	(Thorne & Stull, 1966)
Bacillus licheniformis ATCC14580	CP000002	gleicher Stamm wie DSM13, eine Anzucht Unterschied	(Rey *et al.*, 2004)
Bacillus licheniformis F11	bisher nicht vollständig sequenziert	keine Transposase in *comP*	(Waldeck *et al.*, 2006)

3.2 Komparativer Ansatz

3.2.1 Vergleich der Gen- und Aminosäuresequenzen

Mit Hilfe von ERGO (Overbeek *et al.*, 2003) wurden die Orthologen der 61 Kompetenzgene identifiziert. Sowohl eine Suche in den Annotationen, als auch eine sequenzbasierte Suche wurden durchgeführt. Waren die ORFs nicht vorhanden, wurde eine manuelle Suche mit Hilfe von Artemis (Rutherford *et al.*, 2000) durchgeführt. Mittels des Needleman-Wunsch-Algorithmus (Needleman & Wunsch, 1970) konnten globale Ähnlichkeiten der orthologen Gene und der zugehörigen Aminosäuresequenzen bestimmt werden.

3.2.2 Gencluster-Analyse

Multiple Vergleiche der Kompetenzgen-Umgebungen der fünf genannten (Kapitel 3.1) *Bacillus*-Stämme wurden mit der „*contig regions*“-Funktion aus ERGO (Kapitel 2.5.3) erstellt. Jeweils zwei bis drei Gene in der direkten Umgebung des an der Kompetenzbildung beteiligten Proteins wurden in die Analyse einbezogen.

Diejenigen Genumgebungen, die im Vergleich *B. subtilis* 168 zu *B. licheniformis* DSM13 auffällige Insertions- und / oder Deletionsereignisse zeigen (Tabelle 4), wurden im Detail weiter betrachtet. Alle übrigen konservierten Genumgebungen sind im Anhang (Kapitel 9.2) hinterlegt.

3.2.3 Multipler Vergleich des Kompetenzregulationsmoduls

Der multiple Vergleich des *quorum sensing*-Moduls (Hoffmann *et al.*, 2010) innerhalb der *subtilis*-Gruppe wurde mittels ERGO (Kapitel 2.5.3) durchgeführt. Zusätzlich wurde der GC-*frameplot* aus Artemis (Rutherford *et al.*, 2000) verwendet, um den GC-Gehalt des *comQXPA*-Clusters zu visualisieren.

Für die komparative Analyse von ComS und MecA wurde ein ClustalW-*alignment* (Thompson *et al.*, 1994) erstellt, das mit Hilfe von JalView (Waterhouse *et al.*, 2009) farbig markiert wurde.

3.3 Ergebnisse

3.3.1 Vergleich der Gen- und Aminosäuresequenzen

Die komparative Analyse der Kompetenzgene und –proteine aus Sicht von *B. subtilis* 168 ist in Tabelle 3 zusammengefasst. Für *B. amyloliquefaciens* FZB42 konnten alle 61 Kompetenzproteine identifiziert werden. Die Gensequenzähnlichkeiten liegen im Bereich von 45,4 % bis 95,7 %. Die Proteinsequenzähnlichkeiten decken einen Bereich von 53,5 % bis 99,6 % ab. *B. licheniformis* DSM13 zeigt keine orthologen Gene/Proteine für *rapC*/RapC und *phrC*/PhrC. *comP* ist durch eine Transposase unterbrochen und wurde deswegen in der Analyse nicht berücksichtigt. Die Gensequenzähnlichkeiten reichen von 42,7 % bis 87,9 % und die Proteinsequenzähnlichkeiten von 34,5 % bis 96,6 %. In *B. pumilis* SAFR-032 konnten keine Orthologen für *comS*/ComS, *rapC*/RapC und *phrC*/PhrC identifiziert werden.

Tabelle 3: Vergleich der 61 Kompetenzgene und -proteine von *B. subtilis* 168 mit den Orthologen in den vier Vergleichsorganismen
Links sind die globalen Ähnlichkeitswerte der Gene ausgehend von *B. subtilis* 168 dargestellt; rechts entsprechend die globalen Ähnlichkeitswerte der Proteine; rote Markierungen zeigen fehlende Kompetenzgene /-proteine

Gen	*B. amyloliquefaciens* FZB42	*B. licheniformis* DSM13	*B. pumilus* SAFR-032S	*B. anthracis* str. Sterne	Protein	*B. amyloliquefaciens* FZB42	*B. licheniformis* DSM13	*B. pumilus* SAFR-032S	*B. anthracis* str. Sterne
abrB	89,1 %	85,2 %	82,2 %	77,0 %	**AbrB**	95,9 %	92,7 %	94,8 %	91,7 %
clpC	84,6 %	79,1 %	80,7 %	74,1 %	**ClpC**	99,6 %	98,8 %	98,2 %	93,5 %
sigH	87,1%	80,2 %	79,5 %	66,2 %	**SigH**	98,6 %	96,0 %	96,8 %	85,9 %
nin	71,3 %	50,5 %	47,0 %	47,7 %	**Nin**	88,0 %	65,2 %	33,8 %	44,2 %
nucA	53,8 %	51,2 %	50,7 %	47,2 %	**NucA**	62,3 %	62,2 %	65,0 %	60,4 %
srfAA	72,6 %	64,2 %	58,2 %		**SrfAA**	86,1 %	77,2 %	69,9 %	
srfAB	73,0 %	64,7 %	57,7 %		**SrfAB**	87,0 %	77,1 %	69,7 %	
comS	71,4 %	50,3 %			**ComS**	61,8 %	34,5 %		
srfAC	85,7 %	63,0 %	56,9 %		**SrfAC**	92,9 %	75,1 %	69,7 %	
srfAD	73,3 %	62,9 %	56,2 %		**SrfAD**	86,1 %	75,6 %	71,3 %	
rapC	78,0 %				**RapC**	91,6 %			
phrC	68,2 %				**PhrC**	68,3 %			
sigB	82,9 %	76,6 %	76,0 %	55,0 %	**SigB**	95,8 %	92,5 %	94,0 %	72,7 %
comK	77,8 %	66,4 %	63,9 %	49,4 %	**ComK**	90,2 %	81,9 %	80,3 %	55,3 %
med	74,3 %	68,4 %	64,0 %	42,8 %	**Med**	88,1%	80,2 %	78,3 %	39,3 %
comZ	73,9 %	62,9 %	67,8 %	49,4 %	**ComZ**	76,0 %	85,7 %	87,9 %	53,8 %
oppA	80,7 %	70,6 %	48,2 %	47,5 %	**OppA**	94,1 %	84,0 %	50,6 %	53,6 %
oppB	82,8 %	70,2 %	52,9 %	53,3 %	**OppB**	97,4 %	89,7 %	66,6 %	69,8 %
oppC	84,5 %	67,7 %	53,9 %	54,4 %	**OppC**	98,0 %	88,5 %	64,5 %	65,4 %
oppD	78,7 %	75,1 %	66,4 %	65,6 %	**OppD**	96,1 %	94,1 %	83,6 %	81,3 %
oppF	82,3 %	74,3 %	68,3 %	70,3 %	**OppF**	97,1 %	92,8 %	87,9 %	86,9 %
mecA	85,0 %	73,9 %	72,7 %	61,5 %	**MecA**	94,5 %	90,8 %	89,2 %	75,7 %
ylbF	83,0 %	79,8 %	74,1 %	58,3 %	**YlbF**	96,7 %	92,6 %	88,7 %	75,5 %
smf	68,6 %	59,1 %	59,3 %	40,2 %	**Smf**	82,3 %	72,0 %	70,6 %	37,2 %
codY	81,8 %	79,4 %	80,5 %	73,1 %	**CodY**	99,2 %	98,5 %	98,1 %	93,4 %
pnpA	84,0 %	81,4 %	79,6 %	72,1 %	**PnpA**	98,0 %	97,7 %	96,0 %	91,2 %
cinA	77,3 %	70,6 %	69,0 %	57,5 %	**CinA**	93,8 %	86,8 %	84,4 %	73,3 %
ymcA	83,4 %	77,5 %	76,6 %	67,4 %	**YmcA**	95,8 %	95,1 %	93,1 %	83,2 %
spo0A	85,9 %	80,1 %	76,5 %	66,1 %	**Spo0A**	98,9 %	96,6 %	96,7 %	85,0 %
sinI	78,9 %	60,5 %	59,5 %	50,5 %	**SinI**	82,8 %	67,2 %	67,2 %	42,9 %
sinR	95,7 %	87,9 %	89,1 %	67,4 %	**SinR**	97,4 %	94,7 %	94,6 %	82,9 %
yqzE	58,3 %	51,1 %	56,6 %	50,0 %	**YqzE**	62,0 %	53,8 %	62,0 %	57,0 %
comGG	55,1 %	53,1 %	42,7 %	39,2 %	**ComGG**	74,6 %	61,6 %	46,4 %	34,3 %
comGF	45,4 %	42,7 %	46,4 %	42,5 %	**ComGF**	53,3 %	47,0 %	36,5 %	43,5 %
comGE	58,9 %	49,3 %	48,0 %	42,8 %	**ComGE**	63,8 %	59,1 %	53,3 %	31,2 %

Gen	*B. amyloliquefaciens* FZB42	*B. licheniformis* DSM13	*B. pumilus* SAFR-032S	*B. anthracis* str. Sterne	Protein	*B. amyloliquefaciens* FZB42	*B. licheniformis* DSM13	*B. pumilus* SAFR-032S	*B. anthracis* str. Sterne
comGD	61,3 %	51,0 %	44,5 %	46,0 %	**ComGD**	70,7 %	56,1 %	47,3 %	48,1 %
comGC	69,8 %	64,5 %	68,3 %	49,9 %	**ComGC**	85,4 %	88,8 %	84,7 %	60,0 %
comGB	58,2 %	56,9 %	54,4 %	45,6 %	**ComGB**	75,4 %	71,5 %	60,7 %	53,5 %
comGA	72,0 %	60,8 %	63,5 %	57,4 %	**ComGA**	89,4 %	79,5 %	83,2 %	76,4 %
comEC	60,2 %	54,7 %	54,4 %	48,1 %	**ComEC**	74,1 %	64,7 %	64,3 %	54,9 %
comEB	79,0 %	73,9 %	69,8 %	63,7 %	**ComEB**	95,3 %	93,1 %	90,5 %	77,8 %
comEA	58,5 %	57,2 %	50,9 %	53,3 %	**ComEA**	67,8 %	67,1 %	55,5 %	59,0 %
comER	73,4 %	70,5 %	61,6 %	57,3 %	**comER**	87,6 %	84,2 %	74,9 %	74,9 %
comC	63,3 %	52,3 %	48,9 %	45,8 %	**ComC**	74,9 %	65,1 %	53,6 %	55,5 %
lonB	80,8 %	75,8 %	74,0 %	64,0 %	**LonB**	96,9 %	94,0 %	91,9 %	84,3 %
clpX	82,6 %	79,9 %	80,9 %	76,3 %	**ClpX**	97,1 %	96,7 %	95,7 %	92,6 %
yuxO	72,8 %	68,0 %	61,5 %	58,6 %	**YuxO**	85,2 %	79,8 %	72,3 %	72,1 %
comA	79,6 %	72,2 %	69,1 %	47,2 %	**ComA**	93,0 %	90,7 %	86,0 %	56,2 %
comP	63,2 %		59,5 %		**ComP**	73,9 %		69,5 %	
comX	50,7 %	53,4 %	46,4 %		**ComX**	49,2 %	53,3 %	47,5 %	
comQ	51,6 %	50,7 %	51,2 %		**ComQ**	55,6 %	62,8 %	53,5 %	
degQ	85,4 %	72,8 %	71,6 %		**DegQ**	93,6 %	83,0 %	83,3 %	
slr	74,4 %	64,4 %	61,2 %		**Slr**	90,2 %	84,3 %	75,8 %	
clpP	86,5 %	80,4 %	83,6 %	78,5 %	**ClpP**	98,5 %	95,5 %	98,0 %	94,4 %
comFC	59,9 %	50,8 %	51,7 %	52,0 %	**ComFC**	69,6 %	62,3 %	67,5 %	55,0 %
comFB	72,1 %	61,6 %	59,4 %		**ComFB**	85,7 %	75,0 %	71,8 %	
comFA	61,6 %	61,9 %	57,9 %	57,3 %	**ComFA**	76,6 %	73,3 %	67,0 %	62,4 %
yviA	78,3 %	68,2 %	71,9 %	62,6 %	**YviA**	93,6 %	86,1 %	85,8 %	76,5 %
degU	86,0 %	83,2 %	81,4 %		**DegU**	99,6 %	99,6 %	99,6 %	
degS	81,8 %	73,3 %	72,8 %		**DegS**	98,5 %	91,3 %	91,8 %	
ssb	93,9 %	81,6 %	81,2 %	63,1 %	**Ssb**	98,8 %	94,2 %	91,3 %	81,4 %

Die Gensequenzähnlichkeiten liegen im Bereich von 42,7 % bis 89,1 %. Die Proteinsequenzähnlichkeiten decken einen Bereich von 33,8 % bis 99,6 % ab. *B. anthracis* str. Sterne hat kein Surfactin-Operon und damit auch kein *comS*. Außerdem gibt es keine Orthologen zu *rapC* und *phrC*. Darüber hinaus fehlen *degQ*, *comQ*, *comX* und *comP* sowie *slr*, *comFB*, *degU* und *degS*. Folglich fehlen in *B. anthracis* str. Sterne 15 der Kompetenzgene.

Im Anhang (Kapitel 9.1) befinden sich die kompletten Tabellen der Kompetenzsystemanalyse.

3.3.2 Genclusteranalyse

Eine Zusammenfassung der Genumgebungsanalyse ist in Tabelle 4 dargestellt. Zu jedem Kompetenzprotein bzw. -operon ist angegeben, ob es eine komplette Deletion des Kompetenzproteins gibt. Darüber hinaus ist die Gesamtanzahl der Insertions-/Deletionsereignisse aufgelistet, sowie deren Verteilung auf die einzelnen Organismen. Auffallend ist, dass *B. amyloliquefaciens* FZB42 mit 31 die wenigsten Insertions-/Deletionsereignisse aufweist und *B. anthracis* str. Sterne mit 110 die meisten.

Tabelle 4: Zusammenfassung der Genclusteranalyse
blau markiert sind die Genloci, in deren Umgebung es Insertions- oder Deletionsereignisse im Vergleich von *B. subtilis* 168 und *B. licheniformis* DSM13 gibt; rote Schriftfarbe bedeutet, dass der Genlocus im entsprechenden Stamm deletiert ist

Genlocus	**Deletion des kompletten ORFs**	**Insertions-/Deletionsereignisse in der direkten Genumgebung**	***B. amyloliquefaciens* FZB42**	***B. licheniformis* DSM13**	***B. pumilus* SAFR-032S**	***B. anthracis* str.Sterne**
abrB	0	3	0	0	0	3
clpC /mecB	0	2	0	0	0	2
sigH	0	0	0	0	0	0
nin/ nucA	0	24	4	7	6	7, nur *nin* und *nucA* konserviert
rapC/ phrC	6	36	7	4, (Deletion von *rapC/ phrC*)	15, (Deletion von *rapC/ phrC*)	10, (Deletion von *rapC/ phrC*)
sigB	0	18	3	3	3	9
comK	0	18	6	3	3	6
***opp*-Operon**	0	8	0	1	2	5
mecA	0	8	0	3	2	3
ylbF	0	4	0	0	0	4
smf	0	0	0	0	0	0
codY	0	8	0	0	0	8
pnpA /comR	0	1	0	0	1	0
cinA	0	3	0	1	1	1
ymcA	0	4	0	0	0	4
spo0A	0	16	5	2	5	4
sinI/ sinR	0	5	0	0	0	5
***yqeZ*/*comG*-Operon**	0	15	0	2	2	11
***ComE*-Operon**	0	0	0	0	0	0
ComC	0	3	0	0	0	3
lonB	0	0	0	0	0	0
clpX	0	0	0	0	0	0
***comQXPA*-Cluster**	1	2	0	2	0	Operon deletiert
slr / slrR	0	7	3	2	2	Genumgebung nicht konserviert
clpP	0	undef.	Genumgebung nicht konserviert	Genumgebung nicht konserviert	undef	undef
***comF*-Operon**	0	10	0	0	0	10
***deg*-Operon**	0	4	0	0	0	4
ssb/ ssbA	0	34	3	9	11	11

Im Folgenden (Abbildungen 6-20) wird näher auf die Genumgebungen eingegangen, die im Vergleich *B. subtilis* 168 / *B. licheniformis* DSM13 Insertions- bzw. Deletionsereignisse aufweisen. Abbildung 6 zeigt die Genumgebung von *nucA* und *nin*. Beide Gene sind in den betrachteten Stämmen konserviert.

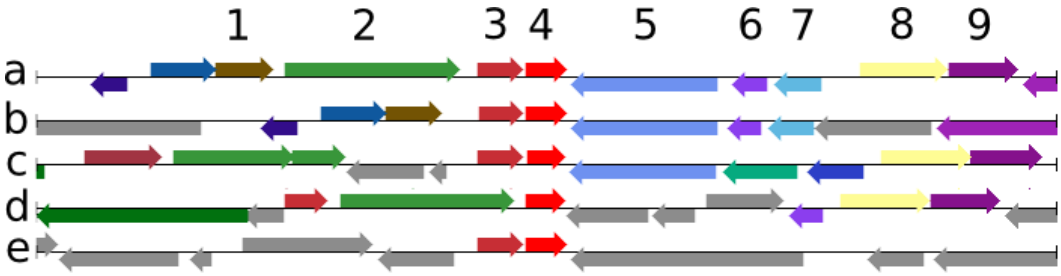


Abbildung 6: Genumgebung von *nucA* und *nin*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *hxlB*, (2) *tlpC* (codiert für ein methyl-akzeptierendes Chemotaxisprotein), (3) *nucA*, (4) *nin*, (5) *yckE*, (6) *yckD*, (7) *yckC*, (8) *yckB*, (9) *yckA*

Es fällt auf, dass in *B. pumilus* SAFR-032 die chromosomalen Lokalisationen von *tlpC* und *nucA* vertauscht sind und dass in *B. anthracis* str. Sterne lediglich *nucA* und *nin*, aber nicht deren Genumgebung konserviert ist. *B. amyloliquefaciens* FZB42 weist Deletionen von *tlpC*, *yckB* und *yckA* auf. Zusätzlich wurde ein Gen zwischen *yckC* und *yciC* inseriert, das möglicherweise für ein Alkoholdehydrogenase-ähnliches Protein kodiert. In *B. licheniformis* DSM13 gibt es drei Deletionen (*hxlB*, *yckD* und *yckC*) und vier Insertionen. Zwei Insertionen liegen zwischen *tlpC* und *nucA*, bei denen es sich um ein hypothetisches Protein und möglicherweise um eine Proteinkinase handelt. Die anderen beiden Insertionen sind zwischen *yckE* und *yckB* und kodieren für eine mögliche NAD(P)H-Dehydrogenase und einen ABC-Transporter. Außerdem ist *tlpC* durch ein Stoppcodon unterbrochen. *B. pumilus* SAFR-032 zeigt Deletionen von *hxlB*, *yckE* und *yckC* sowie Insertionen von drei hypothetischen Genen zwischen *nin* und *yckD*.

RapC und *phrC* sind nur in *B. subtilis* 168 und *B. amyloliquefaciens* FZB42 konserviert und in den anderen betrachteten Stämmen nicht vorhanden (Abbildung 7).

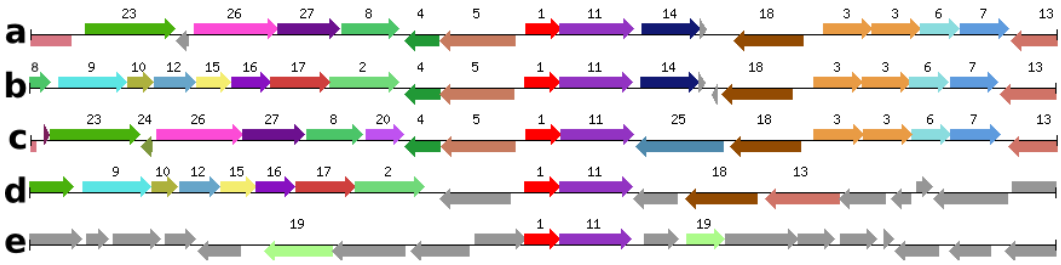


Abbildung 7: Genumgebung von *rapC/phrC*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032, (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (8) *gerKB*, (4) *yclH*, (5) *yclI*, (1) *yclJ*, (11) *yclK*, (14) *rapC* und daneben *phrC*, (18) *yclM*, (3) *yclN*, (3) *yclO*, (6) *yclP*, (7) *yclO*

B. licheniformis DSM13 weist zusätzlich zwei Insertionen auf: (20) hypothetisches Protein und (25) tlpA, das für ein methyl-akzeptierendes Chemotaxis-Protein codiert. Die übrigen Gene sind ebenfalls konserviert. *B. amyloliquefaciens* FZB42 hat die sieben Gene yxeK – yxeQ zwischen gerKB und yclH inseriert, die u. a. für einen möglichen ABC-Transporter kodieren. *B. pumilus* SAFR-032 hat die Gene *yxeK-yxeQ* ebenfalls zwischen *gerKB* und *yclH* inseriert. Außerdem befindet sich eine Transposase vor *yclJ* und ein hypothetisches Protein nach *yclK*. Die Gene *yclN-yclO* sind deletiert. In *B. anthracis* str. Sterne sind lediglich die Gene *yclJ* und *yclK* konserviert.

B. amyloliquefaciens FZB42, *B. licheniformis* DSM13 und *B. pumilus* SAFR-032 weisen in der Genumgebung *sigB* Deletionen von drei hypothetischen Proteinen (*ydcF*, *ydcG* und *ydcH*) auf (Abbildung 8). In *B. anthracis* str. Sterne sind lediglich *rsbV*, *rsbW* und *sigB* konserviert. Hinter *sigB* wurde ein hypothetisches Protein inseriert, dem sich *rsbU* anschließt.

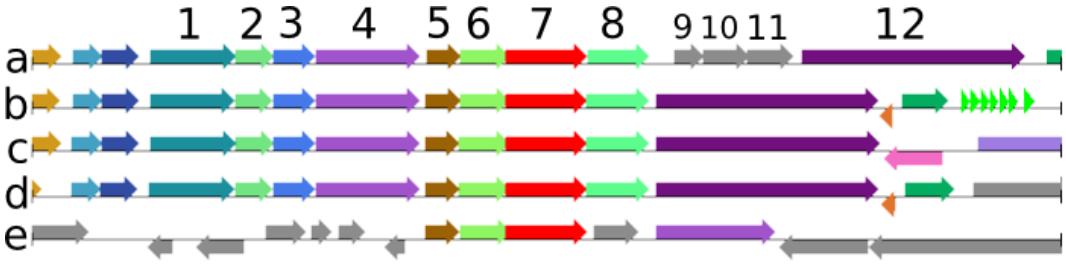


Abbildung 8: Genumgebung von *sigB*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032, (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *rsbRA*, (2) *rsbS*, (3) *rsbT*, (4) *rsbU*, (5) *rsbV*, (6) *rsbW*, (7) *sigB*, (8) *rsbX*, (9) *ydcF*, (10) *ydcG*, (11) *ydcH*, (12) *ydcI*

Abbildung 9 zeigt, dass die Genumgebung von *comK* zahlreiche Insertions- und Deletionsereignisse aufweist. *yhfW* – *comK* sind in allen betrachteten Stämmen außer in *B. anthracis* str. Sterne konserviert. *yhxD* ist in keinem der Organismen konserviert und *yhjA* sowie *yhjB* kommen mit einigen Abweichungen nur noch in *B. amyloliquefaciens* FZB42 vor. Zwischen *comK* und *yhjA* sind vier hypothetische Proteine und zwischen *yhjA* und *yhjB* ist ein hypothetisches Protein inseriert.

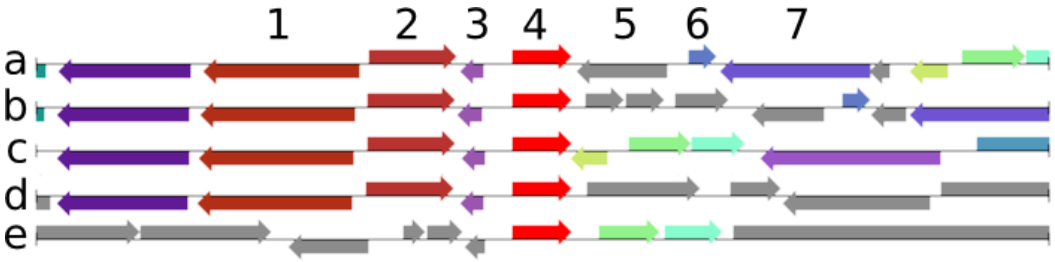


Abbildung 9: Genumgebung von *comK*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032, (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *yhfW*, (2) *yhxC*, (3) *yhzC*, (4) *comK*, (5) *yhxD*, (6) *yhjA*, (7) *yhjB*

Das *opp*-Operon (Abbildung 10) und dessen Umgebung sind in *B. amyloliquefaciens* FZB42 konserviert. *B. licheniformis* DSM13 weist lediglich eine Deletion von *yjbB* auf. *B. pumilus* SAFR-032 hat *yjbB* deletiert und zwischen *oppF* und *yjbC* ein hypothetisches Protein inseriert. *B. anthracis* str. Sterne weist drei Deletionen von *yjbB* bis *yjbD* auf. Auffällig ist, dass statt *oppA* ein anderes für ein oligopeptid-bindendes Protein kodierendes Gen inseriert ist, das keine Ähnlichkeit zu den anderen *opp*-Genen hat.

Abbildung 10: Genumgebung des *opp*-Operons
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne ; Zuordnung der Gene: (1) *oppA*, (2) *oppB*, (3) *oppC*, (4) *oppD*, (5) *oppF*, (6) *yjbB*, (7) *yjbC*, (8) *yjbD*

Der Genbereich um *mecA* (Abbildung 11) ist in *B. amyloliquefaciens* FZB42 konserviert. *B. licheniformis* DSM13 weist eine Deletion von *yjbB*, sowie *coiA* und *pepF* auf. In der Lücke zwischen *yjbD* und *mecA* konnte ein nicht annotiertes Gen mit Ähnlichkeit zu *yjbE* identifiziert werden. Das Gen *yjbB* ist in *B. pumilus*

SAFR-032 deletiert und zwischen *mecA* und *coiA* wurde eine Cardiolipin-Synthetase integriert. Diese Synthetase ist auch in *B. anthracis str.* Sterne integriert. Zusätzlich sind in *B. anthracis str.* Sterne yjbB und yjbC deletiert.

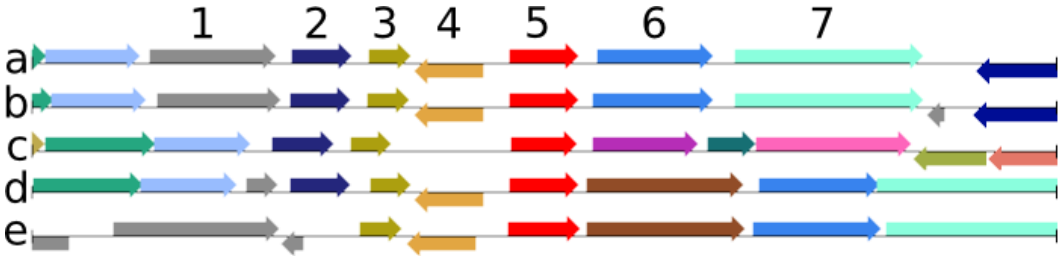


Abbildung 11: Genumgebung von *mecA*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne ; Zuordnung der Gene: (1) *yjbB*, (2) *yjbC*, (3) *yjbD*, (4) *yjbE*, (5) *mecA*, (6) *coiA*, (7) *pepF*

Zunächst fällt auf, dass *B. subtilis* 168 in der Genumgebung von *cinA* (Abbildung 12) ein Stoppcodon in *ymfK* hat und dadurch statt einem, zwei Gene vorhergesagt wurden. In der aktuellen, überarbeiteten *B. subtilis* 168 – Version wurde dieser ORF entfernt. Der gesamte Genbereich um *cinA* ist in allen verglichenen Organismen konserviert. Allerdings weisen *B. licheniformis* DSM13, *B. pumilus* SAFR-032 und *B. anthracis* str. Sterne eine Deletion von *pbpX* auf.

Abbildung 12: Genumgebung von *cinA*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032, (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *ymfK*, (2) *ymfM*, (3) *pgsA*, (4) *cinA*, (5) *recA*, (6) *pbpX*, (7) *ymdA*

Abbildung 13 zeigt, dass die Gene vor und einschließlich *spo0A* in allen Organismen konserviert sind. In *B. amyloliquefaciens* FZB42 sind zwischen *spo0A* und *yqiG* zwei und zwischen *yqiG* und *yqiH* drei hypothetische Proteine eingefügt (nicht mehr in Abbildung 13 zu sehen). Die Gene *yqiG–yqiK* sind vorhanden. *B. licheniformis* DSM13 weist eine Deletion von *yqiG* auf und hat stattdessen ein hypothetisches Protein inseriert. Die Gene *yqiH – yqiK* sind konserviert, wohingegen die Gene *yqiG – yqiI* in *B. pumilus* SAFR-032 deletiert sind. Zwischen *spo0A* und *yqiK* sind zwei hypothetische Gene eingefügt. Die gleiche Genumge-

bung wie in *B. pumilus* SAFR-032 ist in *B. anthracis* str. Sterne konserviert. Allerdings ist statt zwei hypothetischen Proteinen lediglich ein hypothetisches Protein eingefügt.

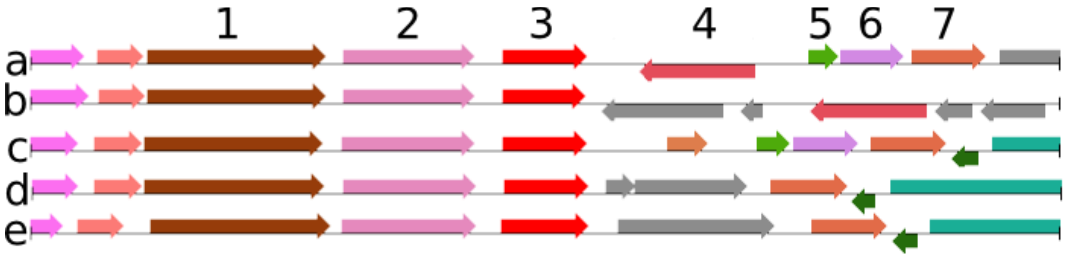


Abbildung 13: Genumgebung von *spo0A*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *recN*, (2) *spoIVB*, (3) *spo0A*, (4) *yqiG*, (5) *yqiH*, (6) *yqiI*, (7) *yqiK*

Das *comG*-Operon (Abbildung 14) ist in allen betrachteten Organismen außer *B. anthracis* str. Sterne konserviert. Auffällig ist, dass in *B. pumilus* SAFR-032 statt *comGG* ein anderes Gen eingefügt ist, das keine Ähnlichkeit zu den übrigen *comG*-Genen aufweist. Im Bereich vor *comGA* sind in *B. licheniformis* DSM13 und *B. pumilus* SAFR-032 die Gene *yqhB* und *yqxL* deletiert. In *B. anthracis* str. Sterne sind nur *comGA – comGD* und *yqzE* konserviert.

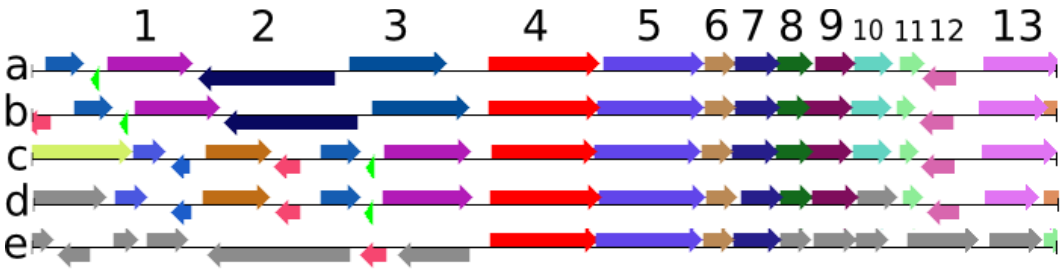


Abbildung 14: Genumgebung des *comG*-Operons
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne ; Zuordnung der Gene: (1) *rsbRD*, (2) *yqhB*, (3) *yqxL*, (4) *comGA*, (5) *comGB*, (6) *comGC*, (7) *comGD*, (8) *comGE*, (9) *comGF*, (10) *comGG*, (11) *yqzE*, (12) *yqzG*, (13) *yqxM*

Das *comQXPA*-Cluster ist in allen Organismen außer *B. anthracis* str. Sterne konserviert. In *B. anthracis* str. Sterne fehlt das gesamte *comQXPA*-Cluster. Abbildung 15 zeigt die Insertion zweier Transposasen in der Mitte von *comP* für *B. licheniformis* DSM13 (Kapitel 3.3.3).

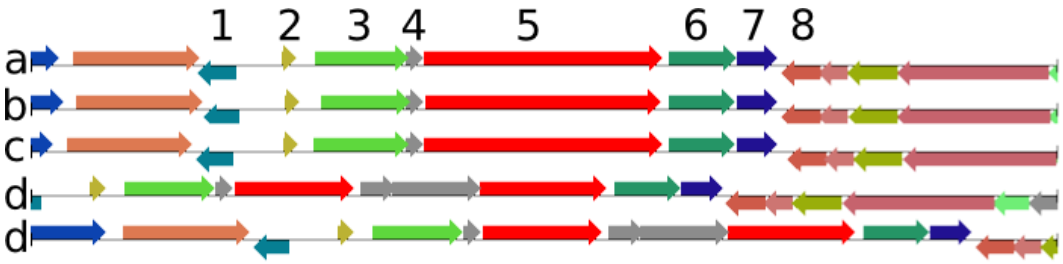


Abbildung 15: Genumgebung des *comQXPA*-Clusters
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. amyloliquefaciens* FZB42, (c) *B. pumilus* SAFR-032, (d) *B. licheniformis* DSM13; Zuordnung der Gene: (1) *yuzC*, (2) *degQ*, (3) *comQ*, (4) *comX*, (5) *comP*, (6) *comA*, (7) *yuxo*, (8) *mrpG*

Abbildung 16 zeigt, dass die Gene *slrR* bis *epsD* in allen Organismen konserviert sind. In *B. anthracis* str. Sterne ist nur *slrR* konserviert, aber nicht dessen genomischer Kontext. *B. amyloliquefaciens* FZB42 weist eine Insertion dreier hypothetischer Proteine zwischen *padC* und *pnbA* auf. In *B. licheniformis* DSM13 und *B. pumilus* SAFR-032 sind *padC* und *pnbA* deletiert.

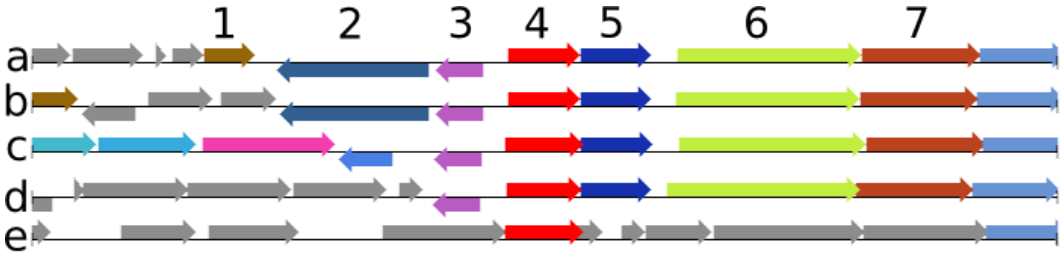


Abbildung 16: Genumgebung von *slrR*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *padC*, (2) *pnbA*, (3) *slrR*, (4) *epsA*, (5) *epsB*, (6) *epsC*, (7) *epsD*

Abbildung 17 zeigt die Genumgebung von *clpP*. *clpP* ist in allen betrachteten Organismen konserviert. Der Genkontext weist erhebliche Deletions- und Insertionsereignisse auf.

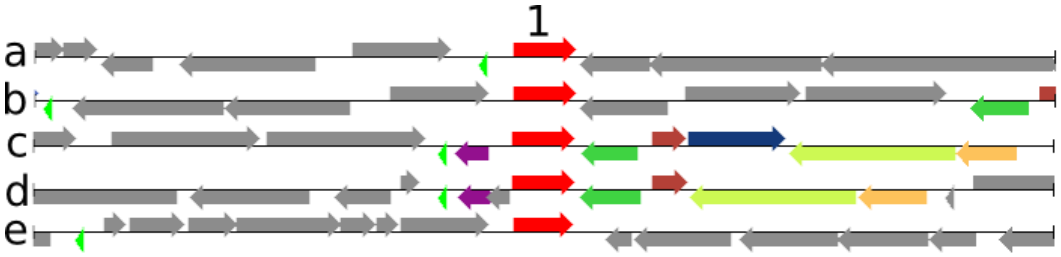


Abbildung 17: Genumgebung von *clpP*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *clpP*

ssb ist in allen betrachteten Organismen konserviert (Abbildung 18). *B. amyloliquefaciens FZB42* hat zwischen *rpsR* und *exoAA* die transkriptionalen Regulatoren *adaA* und *adaB* inseriert. Das Gen *maa* ist nicht vorhanden. In *B. licheniformis* DSM13 sind *yyaE* sowie *exoAA-maa* deletiert. Nach *rpsR* sind die Gene *yybO*, *ybaR*, *yxlE* und *cotF* inseriert. In *B. pumilus* SAFR-032S und *B. anthracis* str. Sterne sind die Gene *yyaE* bis *rspR* konserviert. Der sich anschließende Genbereich ist nicht konserviert.

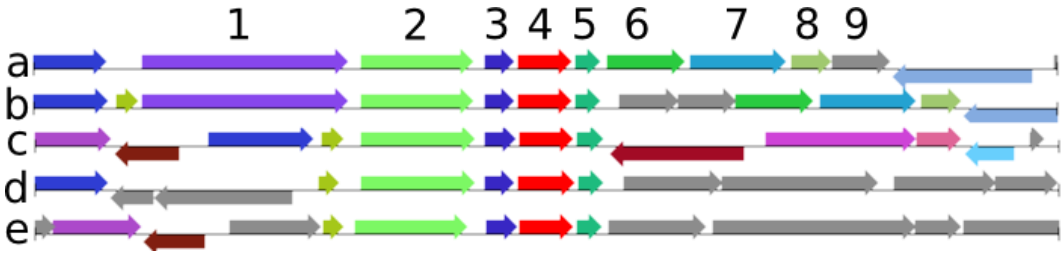


Abbildung 18: Genumgebung von *ssb*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne ; Zuordnung der Gene: (1) *yyaE*, (2) *yyaF*, (3) *rpsF*, (4) *ssb*, (5) *rpsR*, (6) *exoAA*, (7) *ccpB*, (8) *yyaH*, (9) *maa*

3.3.3 Multipler Vergleich des Kompetenzregulationsmoduls

Die komparative Analyse des *comQXPA*-Clusters verschiedener *Bacillus*-Stämme (Kapitel 3.1) zeigt, dass das Cluster in allen Stämmen konserviert ist (Hoffmann *et al.*, 2010) (Abbildung 19).

B. licheniformis DSM13 und *B. licheniformis* ATCC14580 weisen ein Insertionselement in *comP* auf. Die Insertion erfolgte bei DSM13 in gleicher Leserichtung wie die anderen Gene des *comQXPA*-Clusters, während die Insertion in ATTC14580 in entgegengesetzter Richtung erfolgte. Darüber hinaus weisen *B. subtilis* 168, *B. amyloliquefaciens* FZB42, *B. licheniformis* 9945A und *B. licheniformis* F11 eine Senkung des GC-Gehalt ab der Mitte von *comP* in 5‘-Richtung auf, die sich auch über *comX* und *comQ* erstreckt. Genau an der Stelle, an der der GC-Gehalt einbricht, ist die Transposase in DSM13 und ATTC14580 integriert.

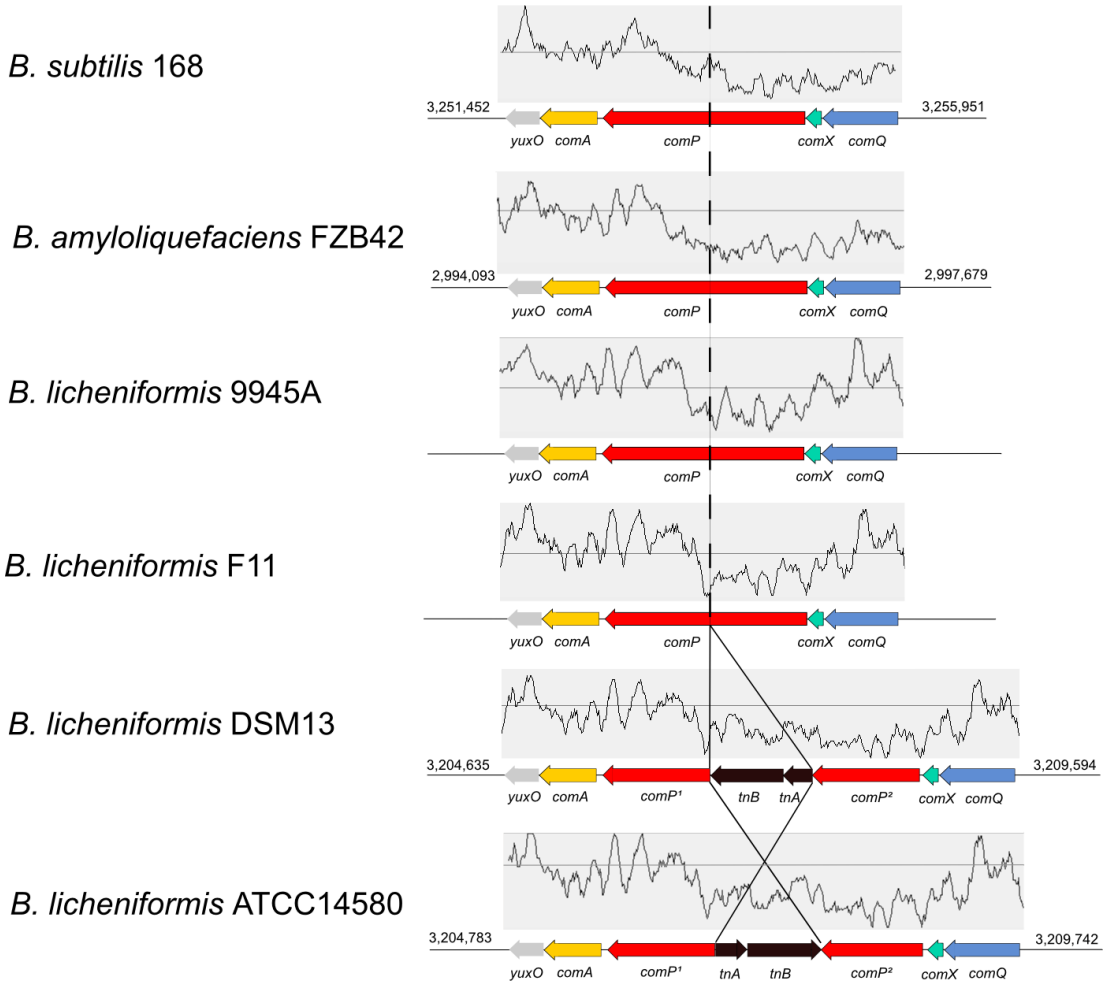


Abbildung 19: Multipler Vergleich des *comQXPA*-Clusters (entnommen aus (Hoffmann *et al.*, 2010))
Das *comQXPA*-Cluster wurde für sechs *Bacillus*-Stämme mit GC *frame plot alignt.* gelb entspricht *comA*, rot *comP*, türkis *comX*, hellblau *comQ* und schwarz den Transposasegenen. Mit der gestrichelten Linie ist der Einbruch des GC-Gehalts markiert, an dessen Position in DSM13 und ATCC14580 die Transposase integriert ist.

Tabelle 5 zeigt die detaillierten Ähnlichkeitswerte der *comQXPA*-Cluster zueinander aus Abbildung 19. Bis auf wenige Ausnahmen liegt die Proteinsequenzähnlichkeit immer über der Nukleotidsequenzähnlichkeit.

Betrachtet man alle Vergleiche einzeln, zeigt sich, dass die höchsten Ähnlichkeitswerte jeweils für die Vergleiche von Yuxo/*yuxo*, ComA/*comA* und DegQ/*degQ* erreicht werden. Die niedrigsten Vergleichswerte finden sich bei ComX/*comX* und ComQ/*comQ*. Für alle Proteine einzeln sind die geringsten Ähnlichkeitswerte jeweils in Vergleichen mit *B. subtilis* 168 bzw. *B. amyloliquefaciens* FZB42 zu finden und die höchsten Werte in Vergleichen von *B. licheniformis*-Stämmen untereinander.

Tabelle 5: Globale Protein- und Nukleotidanzahlen und -sequenzähnlichkeiten der *comQXPA*-Cluster
Im oberen Bereich der Tabelle sind die Längen der Proteine bzw. Nukleotide der einzelnen Stämme angegeben. Im unteren Bereich stehen die Protein- bzw. Nukleotidsequenzähnlichkeiten für die jeweils angegebenen Organismennummern, die sich auf die Beschriftung im oberen Bereich beziehen. Die linke bzw. obere Zahl einer Zelle bezieht sich auf die Proteinsequenz und die rechte bzw. untere Zahl auf die Nukleotidsequenz; rot markiert sind die Felder, in denen das unterbrochene *comP* auftaucht und deswegen keine Vergleiche berechnet wurden, grün markiert sind die jeweils niedrigsten Vergleichswerte für ein Protein; gelb sind die jeweils höchsten Werte für ein Protein

		YuxO/ *yuxO*	**ComA/ *comA***	**ComP/ *comP***	**ComX/ *comX***	**ComQ/ *comQ***	**DegQ/ *degQ***
1 *B. subtilis* 168		127 / 381	215 / 645	770 / 2310	56 / 168	300 / 900	47 / 141
2 *B. amyloliquefaciens* FZB42		128 / 384	215 / 645	767 / 2301	58 / 174	287 / 861	47 / 141
3 *B. licheniformis* DSM13		130 /390	213 / 639		55 / 165	290 / 870	47 / 141
4 *B. licheniformis* ATTC14580		130 /390	213 / 639		55 / 165	290 / 870	47 / 141
5 *B. licheniformis* 9945A		105 /315	213 / 639	772 / 2316	57 / 171	304 / 912	47 /141
6 *B. licheniformis* F11		130 /390	213 / 639	774 / 2322	55 / 165	290 / 870	47 / 141
1	**2**	85.2 % / 72.8 %	93 % / 79.6 %	73.9 % / 63.2 %	49.2 % / 50.7 %	55.6 % / 51.6 %	93.6 % / 85.4 %
1	**3**	8 % / 68 %	91 % / 72.2 %		53 % / 53.4 %	63 % / 50.7 %	83 % / 72.8 %
1	**4**	8 % / 68 %	91% / 72.2 %		53 % / 53.4 %	63 % / 50.7 %	83 % / 72.8 %
1	**5**	65.1 % / 57.1 %	90.7 % / 71.7 %	71.7 % / 60.1 %	48.4 % / 43. %	62.9 % / 50.6 %	76.6 % / 66 %
1	**6**	79.8 % / 67 %	90.7 % / 72%	70.8 % / 58.2 %	53.3 % / 52.9 %	62.8 % / 50.9 %	76.6 % / 66 %
2	**3**	80 % / 63.3 %	89.3 % / 71.8 %		60.3 % / 41 %	57.2 % / 50.4 %	83.3 % / 75. %
2	**4**	80 % / 63.3 %	89.3 % / 71.8 %		60.3 % / 41 %	57.2 % / 50.4 %	83.3 % / 75. %
2	**5**	64.6 % / 55.1 %	89.7 % / 72.3 %	74.8 % / 63.1 %	58.7 % / 46.9 %	60.5 % / 53.4 %	76.6 % / 72.4 %
2	**6**	79.2 % / 64.2 %	89.7 % / 72.3 %	72.1 % / 60.2 %	61.4% / 40.2 %	57.4 % / 50.6 %	76.6 % / 68.2 %
3	**4**	10 % / 10 %	10 % / 10 %		10 % / 10 %	10 % / 10 %	10 % / 10 %
3	**5**	76 % / 76.4 %	10 % / 97.2 %		64.3 % / 57.4 %	93.4 % / 88.7 %	95,7 % / 90.2 %
3	**6**	10 % / 99 %	10 % / 99.5 %		10 % / 99.5 %	10 % / 99.7 %	95.7 %/ 90.2 %
4	**5**	76 % / 76.4 %	10 % / 97.2 %		64.3 % / 57.4 %	93.4 % / 88.7 %	95,7 % / 90.2 %
4	**6**	10 % / 99 %	10 % / 99.5%		10 % / 99.5 %	10 % / 99.7 %	95.7 %/ 90.2 %
5	**6**	76 % / 77 %	10 % / 97.7 %	80.8 % / 74.5 %	64.3 % / 56.7 %	93.4 % / 88.4 %	10 % / 94 %

Auffällig ist, dass das gesamte Cluster in *B. licheniformis* DSM13 und *B. licheniformis* ATTC14580 zu 100 % identisch ist. Zu diesen beiden Stämmen ist das Cluster von *B. licheniformis* F11 am ähnlichsten. Es gibt einige Austausche auf Nukleotidebene, wohingegen die Proteine auch zu 100 % identisch sind. In allen Vergleichen weist *B. licheniformis* 9945A die geringsten Ähnlichkeiten im *comQXPA*-Cluster auf.

Der multiple Vergleich von ComS (Abbildung 20 A) zeigt Unterschiede zwischen kompetenten und nicht kompetenten *Bacillus*-Stämmen. ComS aus *B. licheniformis* 9945A weist eine N-terminale Verlängerung um 16 Aminosäuren im Vergleich zu *B. subtilis* 168 und *B. amyloliquefaciens* FZB42 auf. Darüber hinaus gibt es eine Insertion von vier Aminosäuren an Position 27. Das MecA-Bindemotiv, das für *B. subtilis* mit ILLYPR beschrieben ist (Ogura *et al.*, 1999) unterscheidet sich ebenfalls erheblich (ITRFRP).

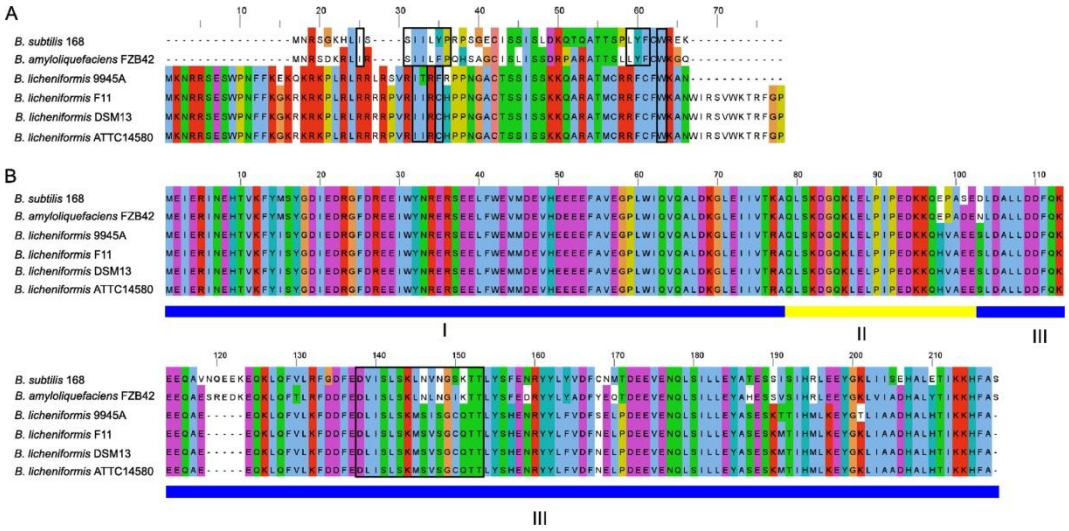


Abbildung 20: ClustalW-*alignments* von ComS(A) und MecA(B) (entnommen aus (Hoffmann *et al.*, 2010))
Gleiche Farben zeigen gleiche Aminosäuren. (A) Schwarze Boxen zeigen für die Kompetenz essentielle Aminosäuren. (B) Die Domänenstruktur von MecA ist hervorgehoben. I zeigt die ComK und ComS – Bindedomäne, II eine *loop*-Region und III gibt die ClpC-Bindedomäne an.

Die nicht auf Kompetenz untersuchte *B. licheniformis*-Stamm F11 sowie die nicht kompetenten DSM13 und ATTC14580 weisen die gleichen ComS-Veränderungen auf und besitzen zusätzlich eine C-terminale Verlängerung von ComS um 12 Aminosäuren.

MecA ist in allen betrachteten Stämmen bis auf wenige Aminosäureaustausche konserviert. Abbildung 20 B zeigt die Domänenstruktur von MecA. Die ComK/

ComS-Bindedomäne ist konserviert. Die *loop*-Region ist in den *B. licheniformis*-Stämmen einheitlich und zeigt im Vergleich zu *B. subtilis* 168 zwei Aminosäureaustausche (Position 98 und 101) sowie zu *B. amyloliquefaciens* FZB42 einen Aminosäureaustausch (Position 98). Die ClpP-Binderegion ist in den *B. licheniformis*-Stämmen konserviert außer in *B. licheniformis* 9945A. Zwei Aminosäureaustausche an Position 168 und 200 markieren den Unterschied. *B. subtilis* 168 weist 16 Aminosäureaustausche im Vergleich zu *B. licheniformis* DSM13 auf. Davon liegen vier in der zweiten ComS-Bindedomäne. *B. amyloliquefaciens* FZB42 hat 15 Aminosäureaustausche in der ClpP-Domäne, drei davon befinden sich in der zweiten ComS-Bindedomäne.

4 Genomweite Identifikation von Orthologen

4.1 Komparativer Ansatz

Zur genomweiten Identifikation von Orthologen und flexiblen Genombereichen wurde das *softwaretool* BiBaG entwickelt. Bakterielle Gesamtgenomvergleiche können damit erstellt werden. BiBaG basiert auf bidirektionalen besten BLAST-*hits* sowie globalen *alignments* auf Proteinebene.

4.1.1 Bestimmung von Orthologen

Zur Bestimmung von Orthologen (Kapitel 2.5.1) wird die Methode des bidirektionalen besten *hits* (Kapitel 2.5.2) verwendet. Abbildung 21 zeigt eine schematische Übersicht des algorithmischen Ablaufes.

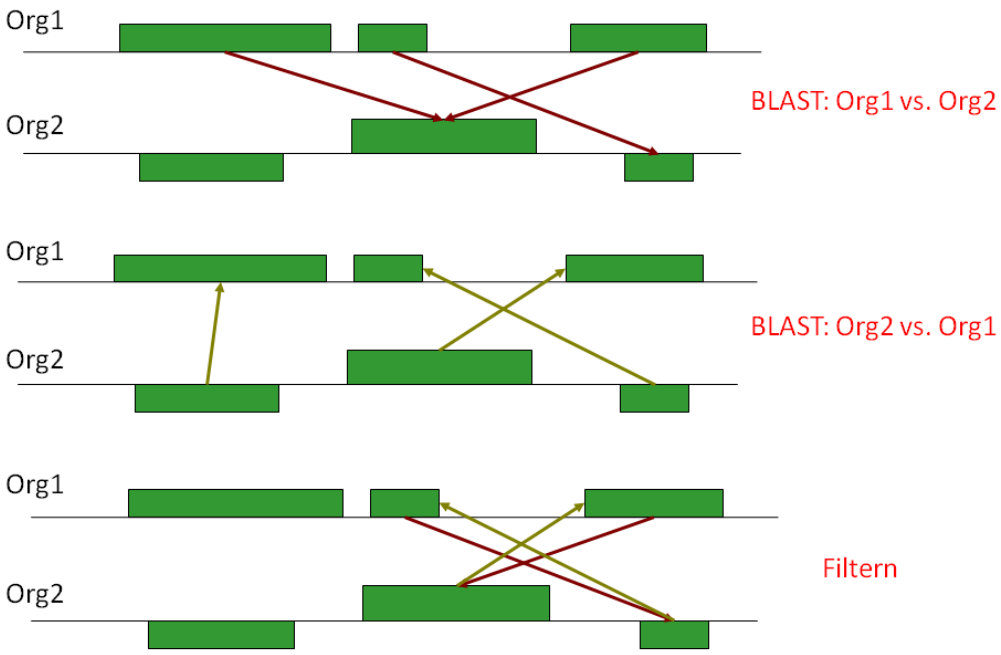


Abbildung 21: Schematische Darstellung des bidirektionalen BLASTs

Betrachtet werden zwei Organismen mit unterschiedlichen Genomen. Die Proteine des Organismus 1 werden mittels BLAST-Analyse mit den Proteinen von Organismus 2 verglichen und der erste Treffer jeden Proteins gespeichert. Anschließend wird die BLAST-Analyse in entgegengesetzter Richtung, also Organismus 2 gegen Organismus 1, wiederholt. Der erste Treffer jeden Proteins wird wieder gespeichert. Im Filterschritt werden dann diejenigen Proteine übernommen, die

gegenseitig in beiden BLAST-Richtungen den jeweils besten Treffer auf sich haben.

In einem zweiten Analyseschritt werden die Proteine mit bidirektionalen besten BLAST *hits* mit Hilfe des Needleman-Wunsch-Algorithmus (Needleman & Wunsch, 1970) verglichen, um globale Sequenzähnlichkeiten zu bestimmen.

Neben den *e-values* und NW-*similarities* sind zur weiteren Beurteilung, ob es sich tatsächlich um Orthologe handelt, die Annotation und der genetische Kontext entscheidend.

4.1.2 Bestimmung von Homologieclustern

Benachbarte Genloci in einem Organismus, die konserviert in einem anderen Organismus vorkommen nennt man Cluster. Liefern die Proteine, deren Gene chromosomal benachbart sind, im Vergleichsorganismus jeweils auch bidirektionale beste *hits* liegt ein Homologiecluster vor.

Die Bestimmung von Homologieclustern (Abbildung 22) basiert auf den bidirektionalen besten *hits* (BBH). Ein einzelnes Protein, das einen bidirektionalen besten *hit* liefert und dessen benachbarte Proteine keine BBH‘s haben, gilt auch als Homologiecluster.

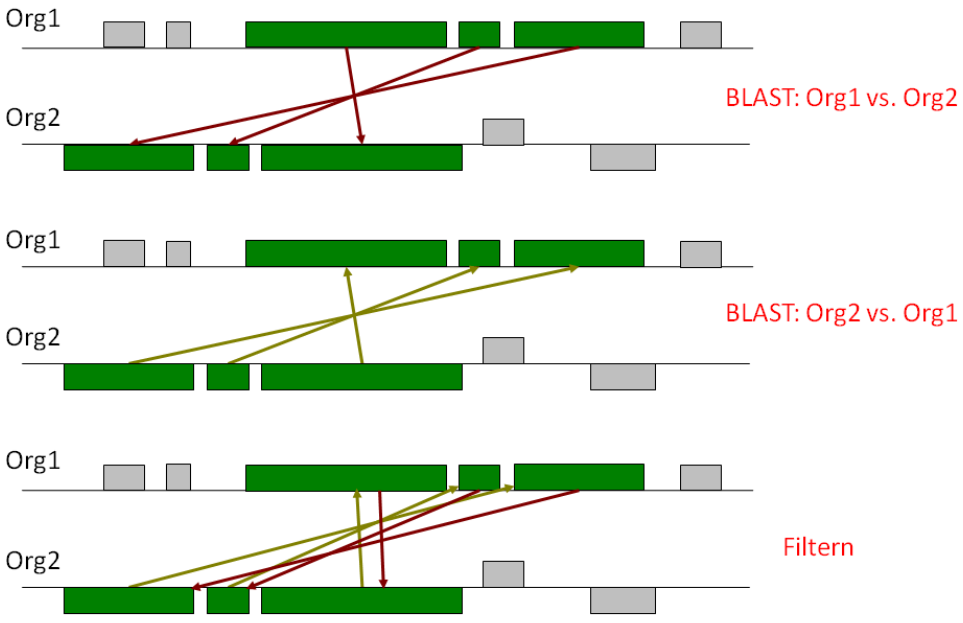


Abbildung 22: Schematische Darstellung des Ablaufs zur Bestimmung von Homologieclustern

4.1.3 Bestimmung von *pan* und *core genome* in drei Organismen

Als *core genome* (Kapitel 2.5.1) werden im bioinformatischen Sinne die Proteine bezeichnet, die in allen Organismen einer Organismengruppe vorkommen. Das *pan genome* setzt sich dieser Definition folgend aus den Proteinen zusammen, die spezifisch für einzelne Organismen der Gruppe sind. Das *pan* und *core genome* lässt sich für drei Organismen durch die Methode der bidirektionale besten *hits* bestimmen (Abbildung 23).

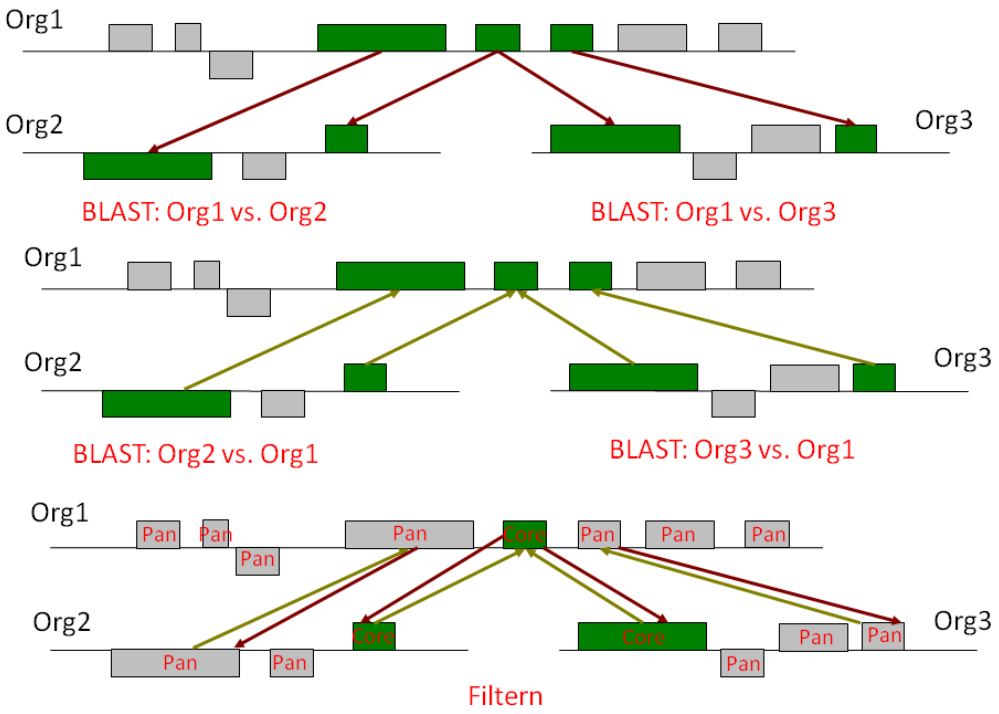


Abbildung 23: Schematische Darstellung der Bestimmung von *pan*/*core genomes* für drei Organismen

Ausgehend von einem Referenzorganismus Org1 werden zwei BLAST-Vergleiche durchgeführt. Dabei wird Org1 zunächst gegen Org2 und dann gegen Org3 geblastet. Im zweiten Schritt werden diese BLAST's umgekehrt durchgeführt. Liefert ein Protein im Referenzorganismus BBHs zu Proteinen beider Vergleichsorganismen, wird dieses Protein des Referenzorganismus als Teil des *core genome* betrachtet. Ein Protein kann folglich entweder in einem, beiden oder keinem Vergleichsorganismus BBHs haben. Anschließend wird die gesamte Analyse für Org2 als Referenzorganismus und anschließend Org3 als Referenzorganismus

wiederholt. Die hier beschriebene Methode wird im Folgenden Triple-BLAST genannt.

Venn-Diagramme sind Kreise mit Überlappungsbereichen von keinem, einem oder beiden der anderen Kreise. Diese Form der Darstellung ist besonders geeignet für einen Überblick über *pan* und *core genome* in drei zu analysierenden Organismen (Abbildung 24).

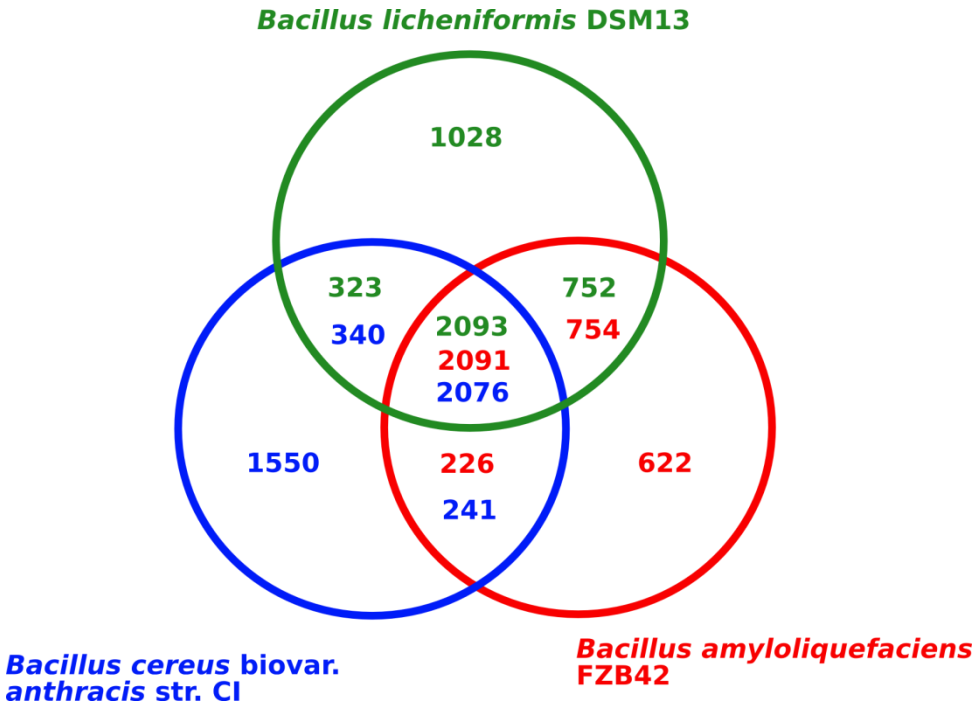


Abbildung 24: Beispiel einer Venn-Diagramm-Darstellung der Triple-BiBaG-Analyse
Jeder organismenspezifische BiBaG ist aus der Sicht des jeweiligen Organismus in einer Farbe dargestellt. Grün entspricht *B. licheniformis* DSM13, rot *B. amyloliquefaciens* FZB42 und blau *B. cereus* biovar. *anthracis* str. CI. Je nachdem welche Organismen Proteine teilen ist die Anzahl der gemeinsamen Proteine angegeben. In der Mitte überlappen alle drei Organismen. Hier sind die Proteinanzahlen der *core genomes* angegeben. In den äußeren Bereichen, in denen es keine Überlappungen gibt befinden sich die stammspezifischen Proteine.

4.1.4 Statistischer Überblick der Orthologenbestimmung für größere Organismenanzahlen

Vergleicht man in einer BBH-Analyse mehrere Organismen miteinander, ist die Information, wie viele bidirektionale beste *hits* es in Subgruppen der Organismen gibt von hoher Bedeutung. Dazu wird für alle möglichen Kombinationen von 2er, 3er, 4er, …-Subgruppen, die Anzahl von BBH's benötigt. Wird beispielsweise eine bidirektionale BLAST-Analyse mit einem Referenzorganismus und vier Ver-

gleichsorganismen durchgeführt, ergeben sich folgende Gruppen g = {12, 13, 14, 123, 124, 134, 1234}. Basierend auf den abhängig vom Referenzorganismus berechneten BBHs lassen sich dann die entsprechenden Anzahlen ermitteln.

4.2 Implementierung

BiBaG wurde in Java 1.6 implementiert. Abbildung 25 zeigt einen Überblick über die erstellten Pakete und Klassen. Es gibt sieben klar voneinander abgegrenzte Pakete, die jeweils spezifische Klassen und Methoden bereitstellen. Das zentrale, steuernde Paket heißt „bibag_logic“. Es beinhaltet die *main*-Methode, die die Eingabedaten einliest und in der angegeben wird, ob eine bidirektionale BLAST-Analyse (BiBaG) (Kapitel 4.1.1 und 4.1.2) oder eine Triple-BLAST-Analyse (TripleBiBaG) (Kapitel 4.1.3) mit genau drei Organismen durchgeführt werden soll. Außerdem enthält das Paket alle Methoden, die den bidirektionalen BLAST bzw. Triple-BLAST auswerten und anschließend das globale *alignment* durchführen sowie die Ausgabedateien bereitstellen.

Für jeden dieser Aufgabenbereiche gibt es ein einzelnes Paket, das die entsprechende Funktionalität beinhaltet. Im Paket „emblIO“ wird das Einlesen der EMBL-Dateien realisiert. „seqComp“ ermöglicht es, den BLAST und das NW-*alignment* durchzuführen. Im „statistics“-Paket werden Methoden bereitgestellt, die übergreifende Aussagen zu den Orthologen einzelner Organismenvergleichsgruppen erlauben (Kapitel 4.1.4).

Im „clustering“-Paket wird die Funktionalität zur organismusübergreifenden Bestimmung von benachbarten, konservierten Genombereichen zur Verfügung gestellt (Kapitel 4.1.2). „excelIO“ dient der Ausgabe der im Programmablauf erstellten Daten als Excel-Tabelle. „graphicsOut“ realisiert die Ausgabe der mit TripleBiBaG (Kapitel 4.1.3) berechneten Werte für die Venn-Diagramme.

Tabelle 6 zeigt die entwickelten Methoden und eine Kurzbeschreibung ihrer Funktionen.

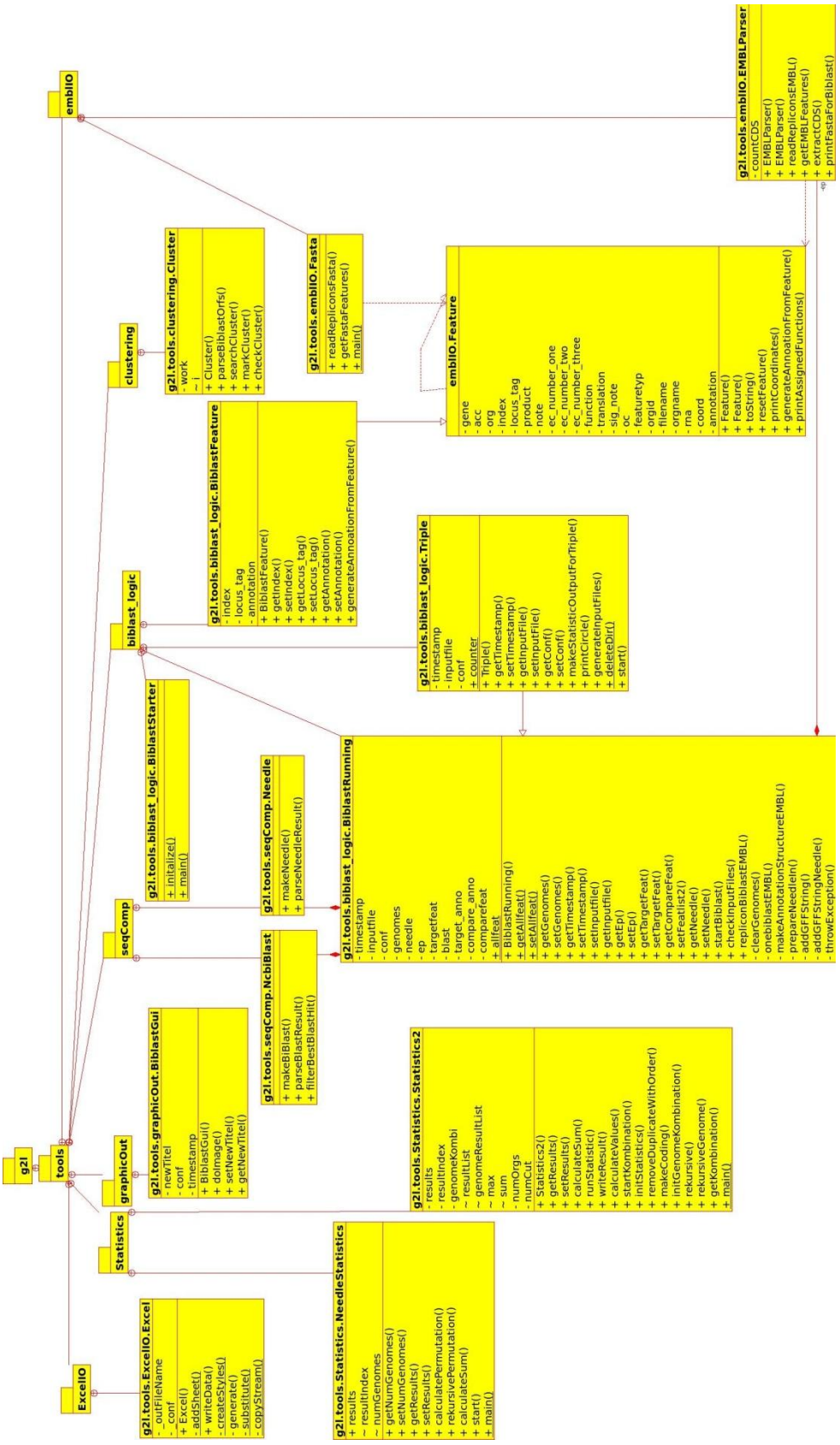


Abbildung 25: UML-Diagramm der BiBaG-Klassen

Tabelle 6: Übersicht der Pakete, Klassen und Methoden von BiBaG
Die Klassen ExcelIO und Combination enthalten im wesentlichen Code, der von anderer Stelle übernommen wurde. Genauere Angaben direkt im Quelltext (Kapitel 9.1, BiBaG.jar).

Paket	Klasse	Methodenname	Kurzbeschreibung
g2l.tools	**Con-fig**	**readConfigFile()**	Einlesen der *bibag.config*-Datei und setzen der nutzerspezifischen Parameter
g2l.tools.bibagLogic	**BiBaG Starter**	**initialize()**	Prüft die notwendige Ordnerstruktur und legt ggf. den *data*-Ordner an
		main()	Startet den BiBaG bzw. TripleBiBaG und gibt die Eingabeparamter weiter
	BiBaGRunning	**startBiBaG()**	Legt *working* und *result*-Ordner mit Zeitstempel an, prüft Eingabedateien
		checkInputFles()	Prüft, ob der Zugriff auf alle Eingabedateien möglich ist
		repliconBiblast EMBL()	Zentrale Methoden, die einzelnen bidirektionalen BLASTs ausführt, Analysen durchführt und die Ergebnisdateien ausgibt
		clearGenomes()	Entfernt einen Organismus aus der internen *genomes*-Liste
		oneBiblastEMBL()	Führt einen bidirektionalen Blast für zwei Organismen durch
		makeAnnotation StructureEMBL()	Legt für eine *feature*-Liste eine *map* mit Annotationen an
		prepareNeedleIn()	Schreibt zwei Fasta-Dateien mit je einer Proteinsequenz
		addGFF String()	Schreibt für ein *feature* eine Zeile in einer GFF-Datei abhängig vom *featuretyp* (*pan* / *core*)
		addGFFString Needle()	Schreibt für ein *feature* eine Zeile in einer GFF-Datei abhängig vom NW -Wert
	TripleRunning	**makeStatistics OutputForTriple()**	Erstellt eine Textdatei mit den Anzahlen der Proteine , die in einem, beiden oder keinem der Organismen vorkommen
		printCircle()	Zeichnet die Venn-Diagramme als Grafikausgabedatei
		generateInput Files()	Legt basierend auf den TripleBiBaG-Eingabedateien, die Eingabedateien für die Vertauschung der Referenzorganismen an
		deleteDir()	Löscht den angegebenen Ordner und seine Unterordner
		start()	Führt den TripleBiBaG aus
g2l.tools.clustering	**Cluster**	**parseBlastResult()**	Liest eine BLAST m8 Tabelle ein
		searchCluster()	Sucht in einer Liste von Proteinen, nach benachbarten, konservierten Proteinen
		markCluster()	Markiert zwei aufeinanderfolgende Proteine einer Liste, wenn sie benachbart sind
		checkCluster()	Prüft, ob zwei Proteine benachbart sind
		addGFFString Cluster()	Schreibt eine Zeile in einer GFF-Datei abhängig davon, ob das *feature* in einem Cluster vorkommt
		calculateGFF Cluster()	Schreibt die Cluster-GFF-Datei für alle Features

Paket	Klasse	Methoden-name	Kurzbeschreibung
2l.tools. emblIO	**EMBL Parser**	**readReplicons EMBL()**	Liest alle Replikondateien zu einem Organismus ein
		getEMBL Features()	Erstellt eine Liste mit allen *features* einer EMBL-Datei
		extractCDS()	Liefert alle Informationen für eine CDS
		printFasta ForBiblast()	schreibt für eine *feature*-Liste eine multiple FASTA-Datei
		getTranslation FromLocus()	liefert die Übersetzung für einen übergebenen *locus tag*
		renumberFeat List()	nummeriert eine *feature*-Liste neu
		getIndex FromLocus()	liefert den internen Index für einen *locus tag*
	Feature	**generateAnnotatio nFromFeature()**	erstellt einen Annotations-String für ein *feature*
		printAssigned Functions()	schreibt eine Zeile in der Ausgabedatei, die alle Annotationen enthält
g2l.tools .excelIO	**Excel***	**writeData()**	schreibt die BiBaG – Ergebnisse in eine Excel-Tabelle
g2l.tools.gra-phicsOut	**Triple BiBaGOut**	**doImage()**	zeichnet das TripleBiBaG-Venn-Diagramm
g2l.tools.seqComp	**NcbiBlast**	**makeBiblast()**	führt den bidirektionalen Blast auf der Konsole aus
		parseBlastResult()	liest eine BLAST m8 – Ausgabetabelle ein
		filterBestBlastHit()	liefert den jeweils besten BLAST-*hit*
	Needle	**makeNeedle()**	führt den NW-Algorithmus des EMBOSS-Pakets auf der Konsole aus
		parseNeedleResult ()	liest die *needle* – Ausgabe ein
g2l.tools. statistics	**Combi-nation***	**startKombinati-on()**	initialisiert die Berechnung aller Kombinationsmöglichkeiten der Subgruppen für eine bestimmte Anzahl von Organismen
	coreSta -tistic	**calculateSum()**	berechnet für ein n : 2^{n-1}
		initGenome Kombination()	berechnet für alle Gruppen, die Anzahl der BBH-basierten Proteine

4.2.1 Format der Eingabedaten

Als Eingabedaten werden EMBL-Dateien verwendet, die den Standards vom EBI genügen müssen. Darüber hinaus muss dem Programm eine Eingabedatei überge-

ben werden, die tab-separiert organismenspezifische *identifier* und die zugehörigen Namen der EMBL-Dateien enthält. Abbildung 26 zeigt eine Beispieleingabedatei. Jeder Organismus erhält eine eigene Zeile und kann aus mehreren Replikons bestehen. Die erste Zeile der Eingabedatei enthält den Referenzorganismus.

```
Bacillus_licheniformis_DSM13    AE017333
Bacillus_subtilis_subsp._subtilis_str._168      AL009126
Bacillus_amyloliquefaciens_FZB42        CP000560
Bacillus_anthracis_str._Ames_Ancestor   AE017334        AE017336        AE017335
Bacillus_thuringiensis_str._Al_Hakam    CP000485        CP000486
```

Abbildung 26: Beispiel einer BiBaG-Eingabedatei
In der ersten Zeile wird der Referenzorganimus, in diesem Fall *B. licheniformis* DSM13, mit dem EMBL-Dateinamen angegeben. Jede weitere Zeile enthält die Vergleichsorganismen, ggf. mit Plasmiden (wie in diesem Beispiel *B. anthracis* str. Ames Ancestor und *B. thuringiensis* str. Al Hakam)

4.2.2 Einstellbare Parameter

Bei der Installation wird die bibag.config-Datei mit Standardparametern angelegt. Die Parameter können in dieser Datei verändert werden, so dass individuelles Programmverhalten ermöglicht wird. Tabelle 7 gibt einen Überblick über die Parameter und Einstellungsmöglichkeiten. Die Parameter können in zwei Gruppen eingeteilt werden. Die ersten Gruppen bilden „maindir“, „formatdb“, „blast“, „blastA“ sowie „needle“, die systemabhängig individuell gesetzt werden müssen. Die zweite Gruppe besteht aus Parametern, die die Art der Ausgabe, abhängig von den Eingabedaten und dem gewünschten Ergebnis individuell beeinflussen und so dem Benutzer zusätzliche Flexibilität erlauben.

Tabelle 7: Überblick der BiBaG-Konfigurationsparameter

Name des Parameters	Funktion des Parameters	Standardeinstellung
maindir	Hauptverzeichnis unter dem alle Eingabe- und Ausgabedateien liegen	/opt/tools/BiBaG/
formatdb	Pfad zum formatdb-Programm des NCBI Blast	/opt/blast/formatdb
blast	Pfad zum blastall-Programm des NCBI Blast	/opt/blast/blastall
blastA	Anzahl der für den Blast verwendeten Prozessorkerne	4 (=maximaler Wert)
needle	Pfad zum needle-Programm des EMBOSS-Packages	/opt/tools/emboss/bin/needle
evalue1 – evalue5	*e-value*-Abstufungen für die Färbung der Exceltabelle	1e-120,1e-100,1e-90, 1e-50, 1e-20
needle20-needle100	Prozentabstufungen der NW-Alignmentlängen für die GFF-Datei	20.0 , 30.0 , 50.0 70.0, 90.0, 100.0
needleTriple1	Erster TripleBiBaG-CutOff für die Venn-Diagramm-Darstellung	25.0
needleTriple2	zweiter Triple-BiBaG-CutOff für die Venn-Diagramm-Darstellung	90.0
orgCut	Anzahl der Organismen der *core*-Statistik	5

4.2.3 Format der Ausgabedaten

Abhängig davon, ob ein BiBaG (Kapitel 4.1.1 und 4.1.2) oder ein TripleBiBaG (Kapitel 4.1.3) gestartet wurde, unterscheiden sich die Ausgabedaten. Für eine BiBaG-Analyse wird eine Excel-Tabelle erzeugt, die alle wesentlichen Informationen enthält. Diese Tabelle besteht aus sechs Datenblättern. Das erste Datenblatt „BiBaG_outevalue“ zeigt farbkodierte *e-values* der BLAST-Analysen. Jeder Vergleichsorganismus erhält eine Spalte. Je kleiner der *e-value* der bidirektionalen *hits* ist, desto dunkler wird die Farbe auf einer sechsstufigen Skala von weiß bis rot (Tabelle 8).

Tabelle 8: Farbcode der *e-values* im ersten, BLAST-basierten Datenblatt der BiBaG-Ergebnisdatei

Farbe	*e-value*
White	> e-20 and 1
LightYellow	<e-20 and >e-50
Gold	<e-50 and >e-90
LightOrange	<e-90 and >e-100
Orange	<e-100 and >e-120
Red	<e-120 and 0

„BiBaG_outneedle“ zeigt für jeden bidirektionalen *hit* die Needleman-Wunsch-Sequenzidentität oder 0, falls es keinen bidirektionalen *hit* gibt. Das dritte Datenblatt „BiBaG_outanno“ enthält die Annotationen der Trefferproteine im Ver-

gleichsorganismus. „BiBaG_outorf“ zeigt die ORF-IDs der bidirektionalen *hits*. Das fünfte Excel-Datenblatt enthält Informationen darüber, ob benachbarte Proteine im Referenzorganismus auch in den Vergleichsorganismen benachbart vorliegen. Sollte kein bidirektionaler *hit* vorhanden sein, wird die Zahl 10000 in die entsprechende Zeile eingetragen. Alle zusammenhängenden Cluster werden mit der gleichen Nummer versehen. Aufeinanderfolgende Cluster werden aufsteigend nummeriert. Das kleinste Cluster besteht aus nur einem Protein. Das sechste Datenblatt heißt „BiBaG_core_statistics“ und enthält die Anzahlen von bidirektionalen *hits* für bestimmte Kombinationen von Vergleichsorganismen (Kapitel 4.1.4).
Zur Visualisierung der Ergebnisse werden drei verschiedene Typen von GFF-Dateien (General Feature Format, http://www.sequenceontology.org/gff3.shtml) bereitgestellt. Die Dateien mit „blast“ im Dateinamen beinhalten die Information, ob es einen bidirektionalen *hit* gegeben hat oder nicht. Die dazu verwendeten *tags* in der Datei heißen *core* und *pan*. „needle“ enthält die Information, wie groß die prozentuale Sequenzidentität ist, basierend auf dem Needleman-Wunsch-*alignment*. Dazu wird ein Stufensystem ähnlich dem *e-value* in der Excel-Tabelle verwendet (Tabelle 9).

Tabelle 9: Zuordnung von GFF-*tags* zu Needleman-Wunsch-Prozentidentitäten

tag	**NW-Identität**
Needle20	0 % - 20 %
Needle30	20 % - 30 %
Needle50	30 % - 50 %
Needle70	50 % - 70 %
Needle90	70 % - 90 %
Needle100	90 % - 100 %

Die dritte GFF-Datei beinhaltet die „cluster“-Informationen. Es werden die *tags* cluster1 und cluster2 verwendet, um direkt benachbarte Proteincluster voneinander unterscheiden zu können. Für jeden Vergleichsorganismus werden also drei GFF-Dateien erstellt. Mit einer entsprechend angepassten „options“-Datei können die Genome und die vergleichenden Analysen (GFF-Dateien) mittels DNA-Plotter (Carver *et al.*, 2009) oder Artemis (Rutherford *et al.*, 2000) visualisiert werden.

Drei weitere Dateien helfen bei der Analyse der Daten bzw. beim Erkennen fehlerhafter Eingabeorganismen. Die Datei „*errorlist*“ enthält alle Organismen, deren EMBL-Dateien fehlerhaft waren und infolgedessen nicht *geparst* werden konnten.

„*checkhit“* beinhaltet alle Proteine, die in einer Richtung einen BLAST-Treffer lieferten, insgesamt aber keinen bidirektionalen *hit* aufwiesen. In der Datei „*nohit“* sind die Proteine vermerkt, die in beiden Richtungen keinen Treffer aufweisen.

Im Gegensatz zum BiBaG werden beim TripleBiBaG mehr Ausgabedateien erzeugt. Es gibt drei Excel-Tabellen für die BiBaG-Analysen aus jeder Sicht der drei zu vergleichenden Organismen. Es werden drei Bilder erzeugt, die Venn-Diagrammdarstellungen (Abbildung 24) der Vergleiche zeigen. Das „blast“-Bild zeigt die jeweiligen Anzahlen der BLAST-Treffer für alle drei Organismen sowie Proteine, die nur in zwei Organismen vorkommen bzw. die organismenspezifischen Proteine, die keinen bidirektionalen *hit* zu einem der anderen beiden Organismen lieferten. „needle25“ zeigt die Proteine, die eine Identität von mehr als 25 % aufweisen, als gemeinsame Treffer und „needle90“ zeigt entsprechend die Treffer, die eine 90 %ige Identität aufweisen. Die Proteine, die diesen Treffern zugrunde liegen, werden als Protein-Fasta-Dateien zur weiteren Analyse mit ausgegeben. Die Dateien „*errorlist*“, „*checkhit*“ und „*nohit*“ werden ebenfalls generiert.

4.3 Test der genomweiten Identifikation von Orthologen

Es wurden drei unabhängige BiBaG-Analysen durchgeführt. Jeder der dafür verwendeten Organismen wurde mittels BiBaG gegen sich selbst verglichen und anschließend analysiert, welche Proteine nicht als Orthologe identifiziert wurden. Grundannahme ist, dass jeder ORF auf sich selbst *mappt*.

4.3.1 Vergleichsdaten

Beim Test der genomweiten Identifikation von Orthologen in Organismenvergleichen mit sich selbst wurden *B. licheniformis* DSM13, *B. subtilis* 168 und *B. licheniformis* 9945A (pers. Komm. Michael Rachinger, Dissertation 2010) verwendet.

4.3.2 Ergebnis

Bei der BiBaG-Analyse von drei Organismen gegen sich selbst zeigt sich, dass es keine Treffer gibt, die nicht auf sich selbst zeigen (falsch-positive). Allerdings gibt es einige Treffer, die keinen BiBaG-*hit* liefern (falsch-negative) (Tabelle 10). Die prozentuale Abweichung der Anzahl nicht *gemappter* Proteine zur Gesamtanzahl der Proteine eines Organismus liegen in allen drei betrachteten Fällen unter 0,5 %. Das entspricht weniger als 5 Proteinen auf 1000 Proteinen, die nicht *gemappt* werden können. Für *B. subtilis* 168 liegt die prozentuale Abweichung sogar nur bei 0,02 %. Das entspricht einem Protein auf 5000 Proteine.

Für *B. licheniformis* DSM13 konnten unter den falsch-negativen Proteinen 16 Transposasen (Bli00522, BLi00523, BLi01678, BLi01679, BLi02279, BLi02280, BLi02483, BLi02484, BLi03103, BLi03104, BLi03340, BLi03341, BLi03355, BLi03356, BLi03590, BLi03591) und zwei Domänentreffer mit 65 AA (BLi01575) sowie 106 AA (BLi01583) identifiziert werden.
Bei *B. licheniformis* 9945A entsprechen die nicht *gemappten* Proteine drei Transposasen (RBLi04298, RBLi04328, RBLi04327).
Für *B. subtilis* 168 sind in Tabelle 9 zwei Werte angegeben, da 68 Proteine als Pseudogene annotiert sind und deswegen keine *coding sequence* enthalten. Bezieht man diese Pseudogene mit ein, so liegt die prozentuale Anzahl *ungemappter* Proteine bei 1,63 %. Werden die Pseudogene nicht berücksichtigt gibt es lediglich ein Protein, das nicht auf sich selbst *gemappt* werden kann. Es handelt sich um ein potentielles Phagenprotein (BSU05099).
Die BiBaG-Ergebnisdateien sind im Anhang (Kapitel 9.1) hinterlegt.

Tabelle 10: Übersicht der BiBaG-Analysen ausgewählter Organismen mit sich selbst

Organismus	**Gesamtanzahl der ORFs**	**Anzahl der *gemappten* ORFs**	**Prozentuale Abweichung**
B. licheniformis DSM13	4196	4178	0,43 %
B. subtilis 168	4245	4176	1,63 % (0,02 %)
B. licheniformis 9945A	4167	4164	0,07 %

4.4 Ergebnisse der Anwendung der genomweiten Identifikation von Orthologen

Im Folgenden werden die Ergebnisse mehrerer BiBaG-Analysen präsentiert. Zunächst wird *B. licheniformis* DSM13 mit ausgewählten Organismen der *Bacillus*-Gruppe verglichen (Kapitel 4.4.1), um gemeinsame und unterschiedliche Genombereiche zu identifizieren. Anschließend wird die Möglichkeit der Annotationsübertragung (4.4.2) vorgestellt. BiBaG wurde außerdem zur Deletions- und Insertionstargetbestimmung (Kapitel 4.4.3, Kapitel 4.4.4) in *B. licheniformis* DSM13 eingesetzt. Tabelle 11 enthält alle Organismennamen und die *accession*-Nummern der verwendeten EMBL-Dateien.

Tabelle 11: Verwendete Bacillus-Stämme in den BiBaG-Analysen

Organismus	***accession*-Nummer**
B. licheniformis DSM13	AE017333
B. subtilis subsp. *subtilis* str. 168	AL009126
B. amyloliquefaciens FZB42	CP000560
B. cereus biovar. *anthracis* str. CI	CP001746 - CP0001749
B. halodurans C-125	BA000004
B. pumilus SAFR-032	CP000813
B. anthracis str. Ames Ancestor	AE017334 - AE017336
B. cereus E33L	CP000001, CP000040 - CP000044
B. clausii KSM-K16	AP006627
B. megaterium DSM319	CP001982
B. thuringiensis str. Al Hakam	CP000485, CP000486
B. anthracis str. A0248	CP001597 - CP001599
B. anthracis str. Ames	AE016879
B. anthracis str. CDC 684	CP001214 - CP001216
B. anthracis str. Sterne	AE017225
B. cereus 03BB102	CP001407, CP001406
B. cereus AH187	CP001177 - CP001181
B. cereus AH820	CP001283 - CP001286
B. cereus ATCC 10987	AE017194, AE017195
B. cereus ATCC 14579	AE016877, AE016878
B. cereus B4264	CP001176
B. cereus G9842	CP001186 - CP001188
B. cereus Q1	CP000227 - CP000229
B. cytotoxicus NVH 391-98	CP000764, CP000765
B. licheniformis ATCC 14580	CP000002
B. licheniformis 9945A	(pers. Komm. M. Rachinger, Dissertation 2010)
B. megaterium QM B1551	CP001983 - CP001990
B. thuringiensis serovar. *konkukian str.* 97-27	AE017355, CP000047
B. tusciae DSM 2912	CP002017
B. weihenstephanensis KBAB4	CP000903
Geobacillus kaustophilus HTA426	BA000043

4.4.1 *B. licheniformis* DSM13 und seine Orthologen in anderen *Bacillus*-Stämmen

Zur komparativen Vergleich von *B. licheniformis* DSM13 im Kontext weiterer Vertreter der *Bacillus*-Spezies wurde eine BiBaG-Analyse durchgeführt. Dabei wurden zunächst Orthologe (Kapitel 4.1.1) und Homologie-Cluster (Kapitel 4.1.2) bestimmt. Anschließend wurden für drei Organismen die *pan* und *core genomes* (Kapitel 4.1.3) identifiziert. Zuletzt wurde für 28 sequenzierte *Bacilli* eine statistische Analyse (Kapitel 4.1.4) der *pan* und *core genomes* durchgeführt.

4.4.1.1 Daten

Für die erste BiBaG-Analyse wurde *B. licheniformis* DSM13 als Referenzorganismus verwendet. Die Vergleichsorganismen sind *B. subtilis* 168, B. *amyloliquefaciens* FZB42, *B. pumilus* SAFR-032, *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 (Tabelle 11). Alle Organismen wurden ohne ihre Plasmide ausgewertet, sofern sie welche haben. Anschließend wurde ein TripleBiBaG mit *B. licheniformis* DSM13, *B. amyloliquefaciens* FZB42 und *B. cereus* biovar *anthracis* str. CI inklusive der Plasmide durchgeführt. Im letzten Schritt wurden eine BiBaG-Analyse mit *B. licheniformis* DSM13 gegen 28 sequenzierte *Bacilli* ebenfalls mit Plasmiden durchgeführt (Tabelle 12, 11).

Die elf Organismen, die in die statische Analyse der *core genomes* (Kapitel 4.1.4) mit einbezogen wurden sind in Tabelle 12 angegeben.

Tabelle 12: 29 Organismen der BiBaG-Analyse
Mit einem Sternchen sind die Organismen markiert, die für die *core genomes* – Analyse verwendet wurden

Organismen			
B. licheniformis DSM13 *	*B. clausii* KSM-K16 *	*B. cereus* AH187	*B. licheniformis* ATCC 14580
B. subtilis subsp. *subtilis* str. 168 *	*B. megaterium* DSM319 *	*B. cereus* AH820	*B. megaterium* QM B1551
B. amyloliquefaciens FZB42 *	*B. thuringiensis* str. Al Hakam *	*B. cereus* ATCC10987	*B. thuringiensis* serovar. *konkukian* str. 97-27
B. cereus biovar. *anthracis* str. CI *	*B. anthracis* str. A0248	*B. cereus* ATCC14579	*B. tusciae* DSM 2912
B. halodurans C-125 *	*B. anthracis* str. Ames	*B. cereus* B4264	*B. weihenstephanensis* KBAB4
B. pumilus SAFR-032 *	*B. anthracis* str. CDC 684	*B. cereus* G9842	
B. anthracis str. Ames Ancestor *	*B. anthracis* str. Sterne	*B. cereus* Q1	
B. cereus E33L *	*B. cereus* 03BB102	*B. cytotoxicus* NVH 391-98	

4.4.1.2 Ergebnisse

Die komparative Analyse von *B. licheniformis* DSM13 mit den fünf Vergleichsorganismen zeigt, dass *B. subtilis* 168 mit 3071 Proteinen die meisten bidirektionalen besten *hits* zu *B. licheniformis* DSM13 hat. *B. amyloliquefaciens* FZB42 teilt 2845, *B. pumilus* SAFR-032S 2751, *B. cereus* biovar. *anthracis* str. CI 2405 und *B. halodurans* C-125 2293 Orthologe mit *B. licheniformis* DSM13.

Abbildung 27 stellt die Visualisierung der globalen NW-*similarities* in der Anordnung abnehmender Orthologenanzahlen von außen nach innen dar.

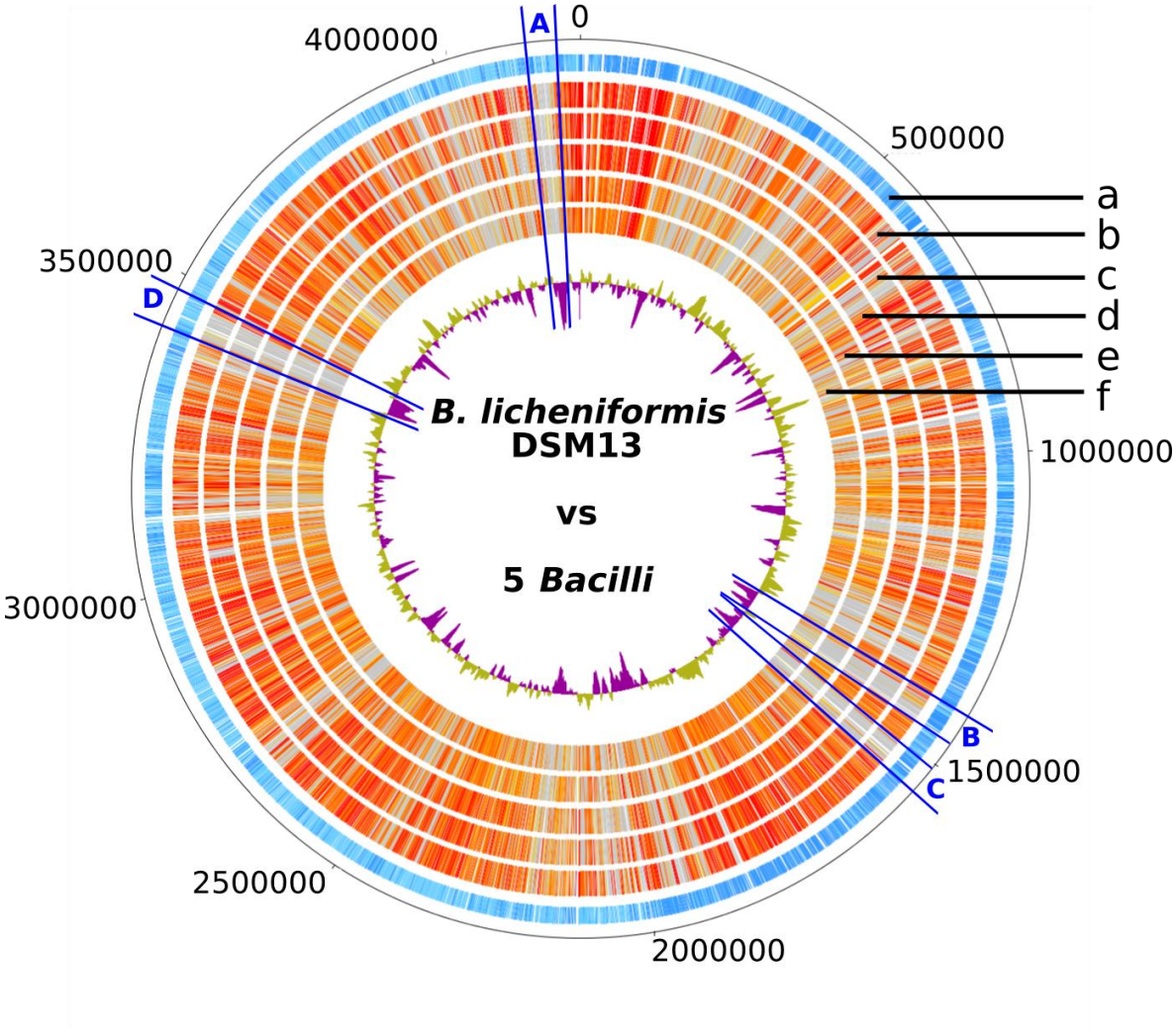


Abbildung 27: Gensonne zur Visualisierung der prozentualen, globalen Sequenzähnlichkeiten der Orthologen
Die Ringe zeigen den Referenzorganismus *B. licheniformis* DSM13 mit seinen *forward* und *reverse* ORFs (a) sowie die Vergleichsorganismen (b-f): (b) *B. subtilis 168*, (c) *B. amyloliquefaciens FZB42*, (d) *B. pumilus* SAFR-032S, (e) *B. cereus* biovar. *anthracis* str. CI, (f) *B. halodurans* C-125. Die Farbkodierung folgt Tabelle 8. Markiert sind vier Bereiche (A-D), die nur im Referenzorganismus vorkommen.

Es sind vier Bereiche markiert, die spezifisch für *B. licheniformis* DSM13 sind. Bereich A umfasst 27 Proteine (= 30.666 bp) und enthält überwiegend hypothetische Proteine sowie ein mögliches Typ-1-Restriktionssystem. Bereich B besteht aus 68 (= 46.788 bp), Bereich C aus 48 (= 40.730 bp) und Bereich D aus 56 (= 46.775 bp) ORFs, die überwiegend für hypothetische Proteine kodieren. Aufgrund der vorhandenen funktionalen Annotation einiger ORFs (Kapitel 9.1, „markierteBereicheDSM13.xls“) sind die Bereiche B, C und D möglicherweise Prophagen. Auffällig ist, dass alle vier Bereiche mit einem vom Durchschnitt abweichenden GC-Gehalt einhergehen.

Es gibt 1515 Proteine, die *B. licheniformis* DSM13 mit allen fünf Vergleichsorganismen teilt. Darunter sind u. a. Proteine zu finden, die wichtig bei der Transkription und Translation sind, wichtige metabolische Funktionen übernehmen oder an der Sporenbildung beteiligt sind (Kapitel 9.1, „DSM13_orthologe_5.xlsx“).

Tabelle 13 zeigt die Verteilung der orthologen Proteine auf prozentuale Sequenzähnlichkeiten. Deutlich ist, dass die niedrigsten Ähnlichkeiten im Bereich von 20-50 % in allen fünf Organismen erreicht werden. Für *B. subtilis* 168, *B. amyloliquefaciens* FZB42 und *B. pumilus* SAFR-032S werden die meisten Orthologen zwischen 70 % und 90 % globaler Sequenzähnlichkeit gefunden. *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 haben die meisten Proteine in der Gruppe von 0 % – 20 %.

Tabelle 13: Verteilung der Orthologenanzahlen auf prozentuale Sequenzidentitätswerte
Für jeden der fünf Organismen ist angegeben, wie viele Orthologe es in den sechs Needleman-Wunsch-Sequenzähnlichkeitsintervallen gibt

NW-Wert / Organismus	***B. subtilis* subsp. *subtilis* str. 168**	***B. amyloliquefaciens* FZB42**	***B. pumilus* SAFR-032**	***B. cereus* biovar. *anthracis* str. CI**	***B. halodurans* C-125**
0 % - 20 %	1126	1355	1447	1797	1912
20 % - 30 %	10	11	5	16	27
30 % - 50 %	83	84	95	242	231
50 % - 70 %	368	361	459	778	780
70 % - 90 %	1666	1556	1535	1173	1100
90 % - 100 %	943	829	655	190	146

Für die gleiche BiBaG-Anaylse mit *B. licheniformis* DSM13 als Referenzorganismus und den fünf Vergleichsorganismen wurde auch eine Homologie-Cluster-Analyse durchgeführt.

Abbildung 28 zeigt die Verteilung der Cluster in den einzelnen Organismen. In den Bereichen mit einem GC-Gehalt, der geringer ist als der durchschnittliche GC-Gehalt gibt es keine oder nur kurze Cluster aus ein bis zwei Proteinen.

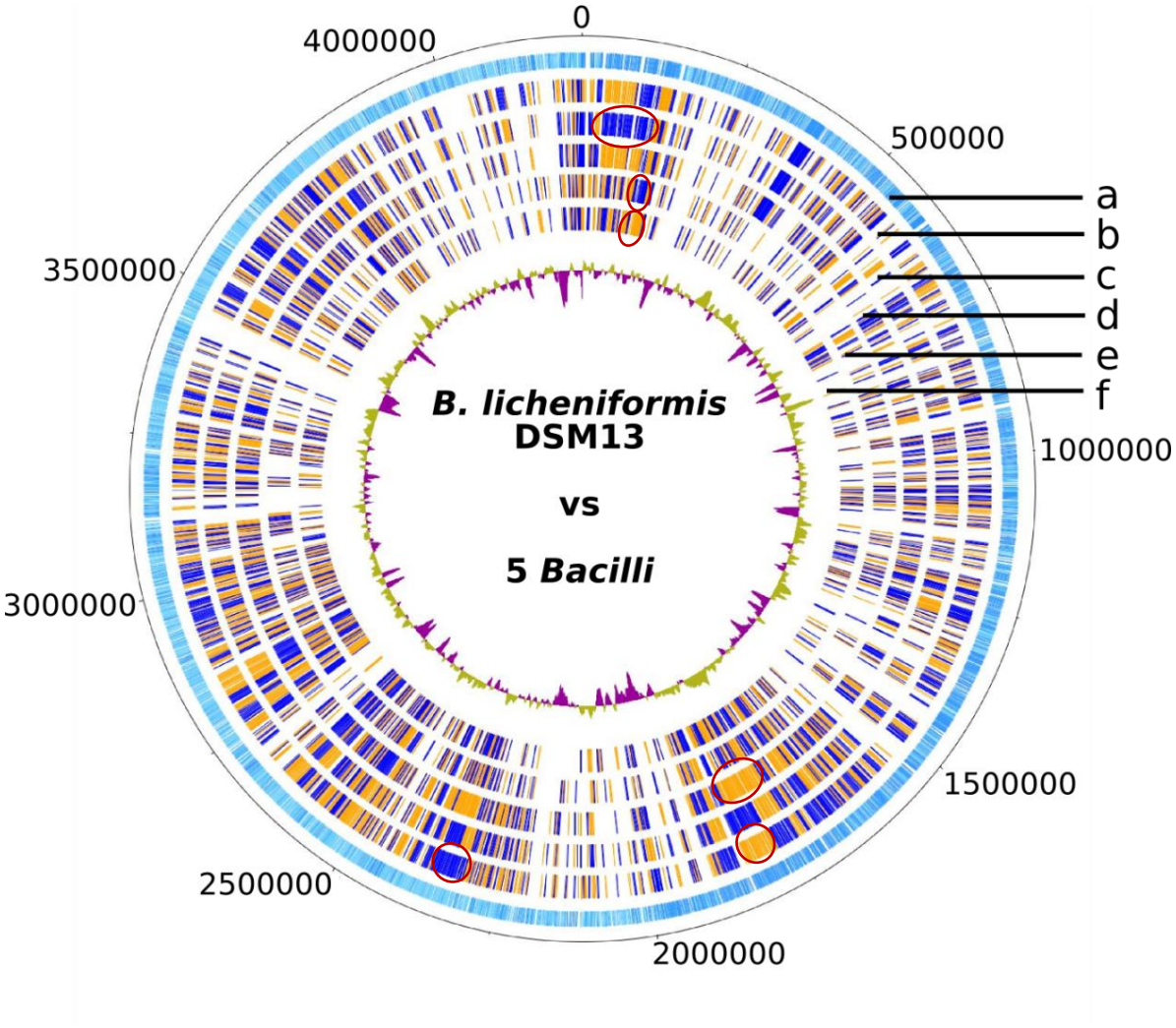


Abbildung 28: Gensonne mit Visualisierung der BiBaG-Clusteranalyse
Die Ringe zeigen den Referenzorganismus *B. licheniformis* DSM13 mit seinen *forward* und *reverse* ORFs (a) sowie die Vergleichsorganismen (b-f): (b) *B. subtilis 168*, (c) *B. amyloliquefaciens FZB42*, (d) *B. pumilus* SAFR-032S, (e) *B. cereus* biovar. *anthracis* str. CI, (f) *B. halodurans* C-125. Homologie-Cluster sind abwechselnd orange und blau gefärbt. Das längste Cluster ist jeweils rot eingekreist.

Tabelle 14 fasst die Ergebnisse der Homologie-Cluster-Analyse zusammen. Die meisten Cluster liegen in *B. cereus* biovar. *anthracis* str. CI vor.

Tabelle 14: Zusammenfassung der Clusterinformationen aus der Gensonne (Abbildung 28)

Cluster-information / Organismus	***B. subtilis* subsp. *subtilis* str. 168**	***B. amylo-liquefaciens* FZB42**	***B. pumilus* SAFR-032**	***B. cereus* biovar. *anthracis* str. CI**	***B. halodu-rans* C-125**
Gesamtanzahl Cluster	967	890	854	1330	1188
Anzahl der Proteine des längsten Clusters	63	88	92	38	44

Die geringste Anzahl an Homologie-Clustern findet sich in *B. pumilus* SAFR-032. Die Anzahlen der Proteine des jeweils längsten Clusters sind in *B. subtilis* 168, *B. amyloliquefaciens* FZB42 und *B. pumilus* SAFR-032 im Vergleich zu *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 um wenigstens 19 Proteine größer. Die längsten Homologiecluster für *B. amyloliquefacienzs* FZB42, *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 befinden sich in der Region um den Replikationsursprung. *B. pumilus* SAFR-032 weist das längste Homologiecluster im Bereich *downstream* des PBSX-Prophagen auf. Für *B. subtilis* 168 konnten zwei Homologiecluster mit der jeweils gleichen Anzahl an Proteinen identifiziert werden. Das eine liegt ebenfalls in der Region *downstream* des PBSX-Prophagen und das Andere etwa entgegengesetzt dem Replikationsursprung. Die genauen Bereiche und ihre Annotationen sind im Anhang hinterlegt (Kapitel 9.1, „markierteCluster.xls“).

Abbildung 29 zeigt die Ergebnisse der TripleBiBaG-Analyse der Organismen *B. licheniformis* DSM13, *B. cereus* biovar. *anthracis* str. CI und *B. amylo-liquefaciens* FZB42. Da die TripleBiBaG-Analyse auf drei BiBaG-Analysen beruht, sind für jeden Organismus vier Zahlen angegeben, die die spezifischen und mit einem oder beiden anderen Organismen geteilten Proteinanzahlen darstellen.

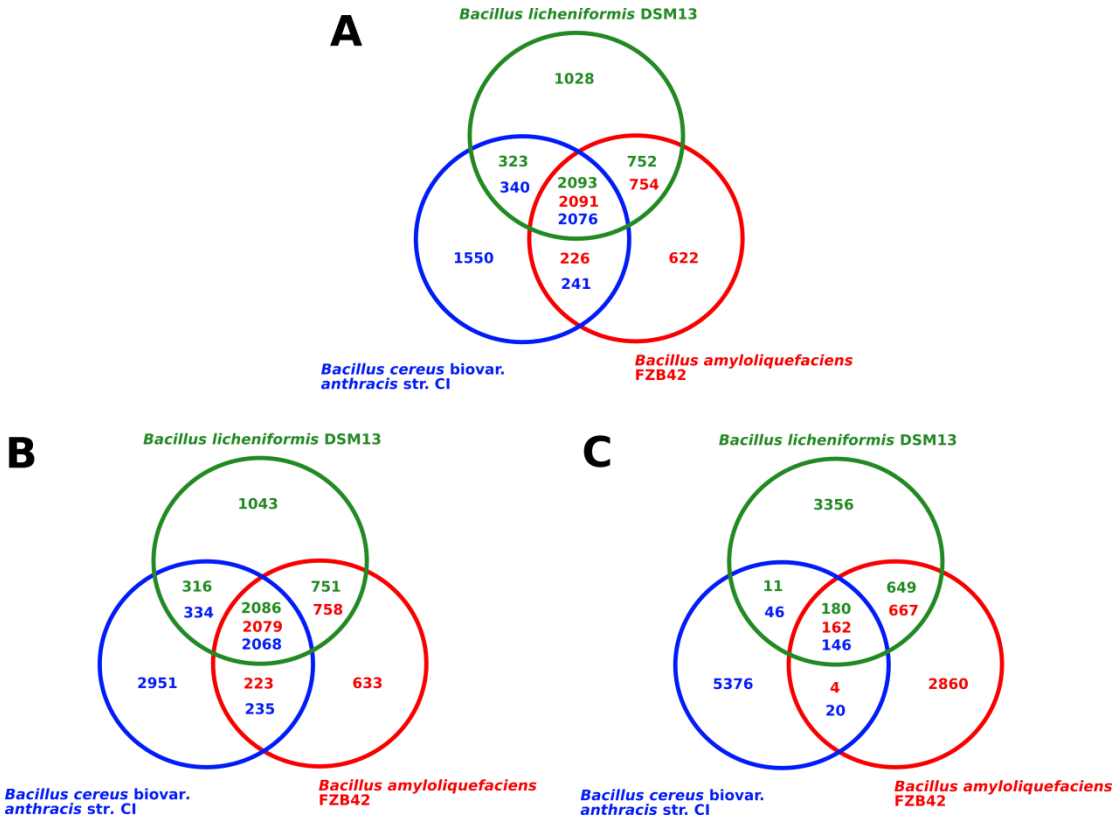


Abbildung 29: Venn-Diagramme der TripleBiBaG-Analyse
Verglichen wurden jeweils *B. licheniformis* DSM13, *B. cereus* biovar. *anthracis* str CI und *B. amyloliquefaciens* FZB42. Es wurden drei unterschiedliche Kriterien zur *core genome* Bestimmung angewandt, die in den drei Grafiken A, B und C dargestellt sind: (A) ein bidirektionaler *hit* liegt vor, (B) NW-*similarity* > 25 %, (C) NW-*similarity* > 90 %

Die Anzahl der *core genome*-Proteine zwischen dem bidirektionalen *hit* (Abbildung 29 A) und einer NW-*similarity* von minimal 25 % (Abbildung 29 B) differiert nur um maximal 12 Proteine. Für das strengere Kriterium von minimal 90 % NW-*similarity* (Abbildung 25 C) gibt es weniger als 200 Proteine, die allen Organismen gleich sind. Gleichzeitig steigt die Anzahl der organismenspezifischen Proteine.

In allen Vergleichen teilen *B. licheniformis* DSM13 und *B. amyloliquefaciens* FZB42 mehr Proteine miteinander, als *B. cereus* biovar. *anthracis* str. CI mit einem der anderen beiden Organismen. In allen drei Vergleichen ist das organismenspezifische Genom von *B. cereus* biovar. *anthracis* str. CI am größten. Im Anhang (Kapitel 9.1, „*core genomes*“) befinden sich neben den zugrunde liegenden Daten (TripleBiBaG-Tabellen, multiple FASTA-Dateien) auch Listen mit den Proteinen der *core genomes* aus allen drei Organismensichten und für alle drei *core genome* Kriterien.

Das *B. licheniformis* DSM13 *core genome*, basierend auf der 90 % NW-*similarity*, besteht aus 180 im wesentlichen lebensnotwendigen Proteinen, z. B. ribosomalen Proteinen, tRNA-Synthetasen, Polymerasen, Zellteilungsinitialisierungsproteinen und Proteine die am zentralen Energiestoffwechsel beteiligt sind. Ein Vergleich dieser 180 Proteine mit den von *B. subtilis* 168 *gemappten* essentiellen Genen (Kapitel 4.4.3) zeigte, dass 83 Proteine in beiden Analysen identisch sind (Kapitel 9.1 „essential_DSM13_abgleich_TripleBiBaG.xls“.

Betrachtet man die 1028 organismenspezifischen Proteine von *B. licheniformis* DSM13 basierend auf den BBH’s (Abbildung 29 A), so stellt man fest, dass neben den Proteinen des PBSX-Prophagen hauptsächlich hypothetische Proteine vorhanden sind (Kapitel 9.1, „blast_unique1“). ComX zählt auch zu den organismenspezifischen Proteinen.

Abbildung 30 basiert auf einer statistischen Analyse (Kapitel 4.1.4) der *core* und *pan genomes* von elf *Bacillus*-Stämmen (Tabelle 12, mit * gekennzeichnet). Die Anzahl der Organismen pro Vergleich ist gegen die Anzahl der Proteine aufgetragen. Ausgangspunkt ist *B. licheniformis* DSM13 mit 4196 Proteinen. Jede Hinzunahme eines weiteren Organismus, führt zu einer erhöhten Proteinanzahl im gesamten Genpool. Für jede mögliche Kombination der 2er, 3er, ... –Gruppen wurde der Mittelwert der *core* und der *pan genomes* aufgetragen. Hierbei ist *pan genome* ohne das *core genome* angegeben. Außerdem sind die maximalen und minimalen Werte innerhalb einer Gruppe dargestellt.

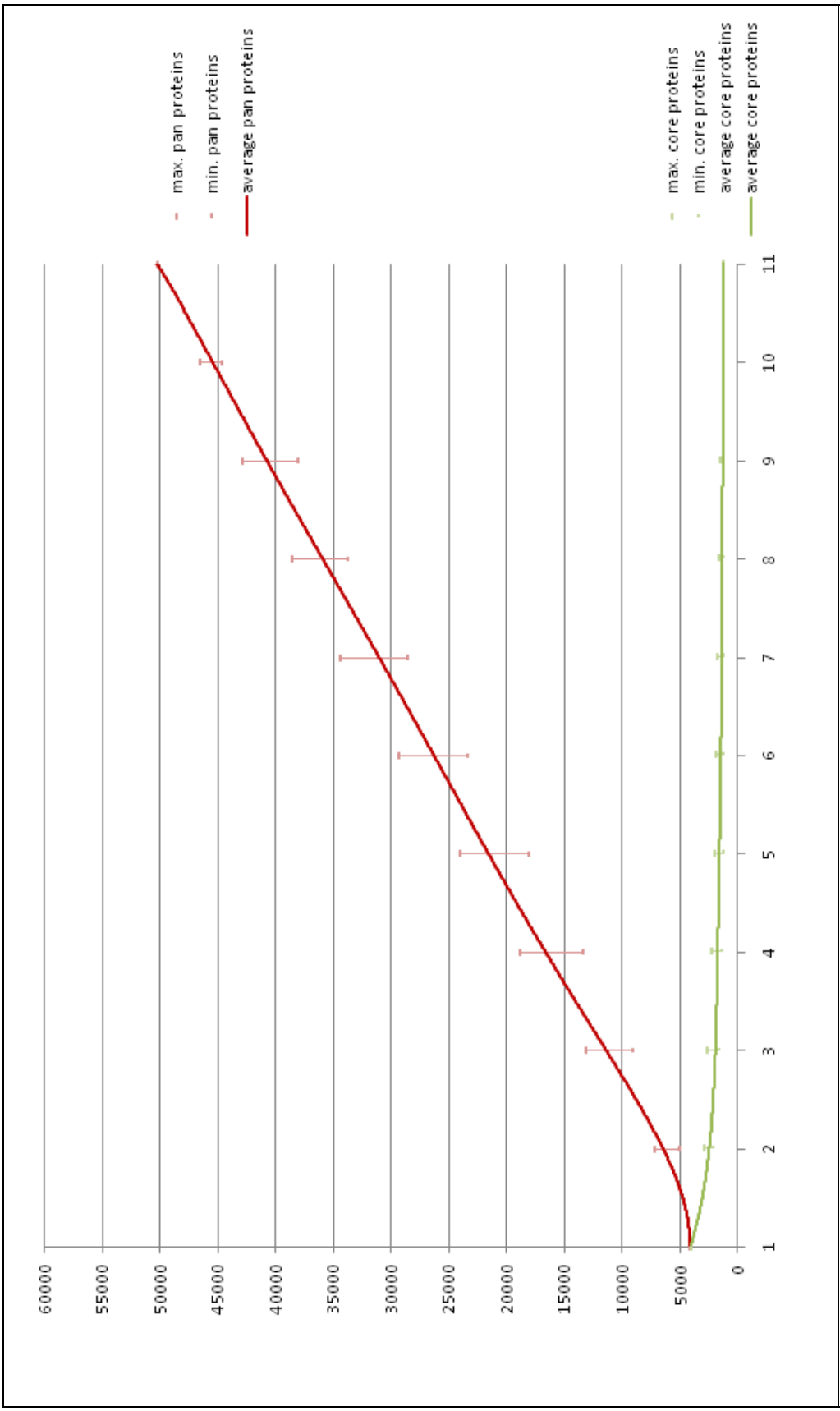


Abbildung 30: Statistischer Überblick der *pan* und *core genome* Statistik
Auf der x-Achse ist die Anzahl der Organismen angegeben und die y-Achse zeigt die Anzahl der Proteine

Das *core genome* nimmt mit jeder Hinzunahme eines weiteren Organismus ab und scheint ein Plateau zu erreichen. Auch die minimalen und maximalen Werte nehmen in ihrem Abstand zueinander ab. Das *pan genome* steigt hingegen sehr schnell an und erreicht auch kein Plateau. Der Abstand der minimalen und maximalen Werte der *pan genomes* nimmt bis zur Hälfte der Organismenanzahlen zu und erreicht dort sein Maximum. Anschließend nimmt der Abstand ab, bis es bei elf Organismen nur noch ein *pan genome* gibt und deshalb Minimal- und Maximalwert gleich sind. Die zugrundeliegenden Berechnungen sind im Anhang hinterlegt (Kapitel 9.1, „biblast_core_sorted.xlsx“, „pan_10_berechnung.xlsx“, „pan_10_graph.xlsx“).

Das *Bacillus-core genome* ausgehend von *B. licheniformis* DSM13 im Vergleich zu den 28 *Bacillus*-Stämmen basierend auf BBHs besteht aus 875 orthologen Proteinen (Kapitel 9.1, „DSM13_orthologe_28.xlsx“). Es konnten 396 stammspezifische Proteine für *B. licheniformis* DSM13 identifiziert werden (Kapitel 9.1, „DSM13_specific.xlsx“). Dabei wurde *B. licheniformis* ATTC14580 nicht berücksichtigt, da es sich um einen zu DSM13 isogenen Stamm handelt. Die stammspezifischen Proteine sind größtenteils als hypothetische Proteine annotiert.

4.4.2 Annotationsübertragung von *B. subtilis* 168 auf *B. licheniformis* 9945A und *B. licheniformis* DSM13

Mit Hilfe von BiBaG wurden die Annotationen von *B. subtilis* 168 (Barbe *et al.*, 2009)auf die Orthologen in *B. licheniformis* 9945A und *B. licheniformis* DSM13 übertragen. Als Orthologe wurden alle Proteine angesehen, die einen bidirektionalen *hit* geliefert haben und zusätzlich mindestens zu 70 % globaler *alignment*-Länge übereinstimmen.

4.4.2.1 Daten

Es wurden die EMBL-Dateien von der *B. subtilis* 168, *B. licheniformis* DSM13 und *B. licheniformis* 9945A verwendet (Tabelle 10).

4.4.2.2 Ergebnisse

Die Annotationsübertragung mittels BiBaG lieferte 2609 von *B. subtilis* 168 auf *B. licheniformis* DSM13 übertragene Annotationen und 2608 übertragene Annotationen auf *B. licheniformis* 9945A (Tabelle 15). Es konnten folglich unter den

angegebenen Bedingungen (4.4.2.1) mehr als die Hälfte der Proteine beider Organismen mit funktionalen Annotationen versehen werden.

Tabelle 15: Übersicht der Anzahl übertragener ORF-Annotationen

Stamm	***B. licheniformis* DSM13**	***B. licheniformis* 9945A**
ORFs	4196	4167
Orthologe	2609	2608
Nicht-orthologe Proteine	1587	1559

Die Analyse der einzelnen übertragenen Annotationen zeigt, dass die 2609 Orthologen aus DSM13 nicht vollständig mit den übertragenen Annotationen aus 9945A übereinstimmen. Es gibt 2514 Orthologe, die beide Organismen mit *B. subtilis* 168 gemeinsam haben (Kapitel 9.1, „common.xls“). 189 unterschiedliche Orthologe konnten identifiziert werden und damit 189 Annotationen, die in beiden Stämmen unterschiedlich sind (Kapitel 9.1, „unique.xls“). Im Anhang befinden sich außerdem zwei Excel-Tabellen mit den Zuordnungen der Annotationen auf die jeweiligen orthologen Proteine (Kapitel 9.1, „DSM13_annotated.xls“, „9945_annotated.xls“).

4.4.3 Deletionstargetbestimmung für *B. licheniformis* DSM13

Die Bestimmung von Deletionstargets ist entscheidend, um *B. licheniformis* DSM13 für den Produktionsprozess zu optimieren. Für *B. subtilis* 168 gibt es bereits Analysen dazu, welche Proteine essentiell sind (Kobayashi *et al.*, 2003). Es wurde schrittweise jedes Gen einzeln deletiert und überprüft, ob die Mutanten in Vollmedium noch wachsen konnten. Mit diesem Ansatz wurden für *B. subtilis* 168 271 essentielle Gene bestimmt (Kapitel 9.1, „essential_subtilis.xls“).

Im zweiten Schritt wurden produktionsirrelevante Bereiche, wie z. B. Phagen, identifiziert und deletiert (Ara *et al.*, 2007). Der dabei entstehende Stamm *B. subtilis* MGB469 weist gegenüber dem Wildtypstamm ein um 469 kb reduziertes Genom auf. *B. subtilis* MGB469 wurde durch die Deletion weiterer Bereiche zu *B. subtilis* MGB876 transformiert (Morimoto *et al.*, 2008). Dessen Genom ist 876 kb kleiner als das Genom von *B. subtilis* 168.

Abbildung 31 zeigt das Genom von *B. subtilis* 168 mit den essentiellen und deletierten Genbereichen. (Kapitel 9.1)

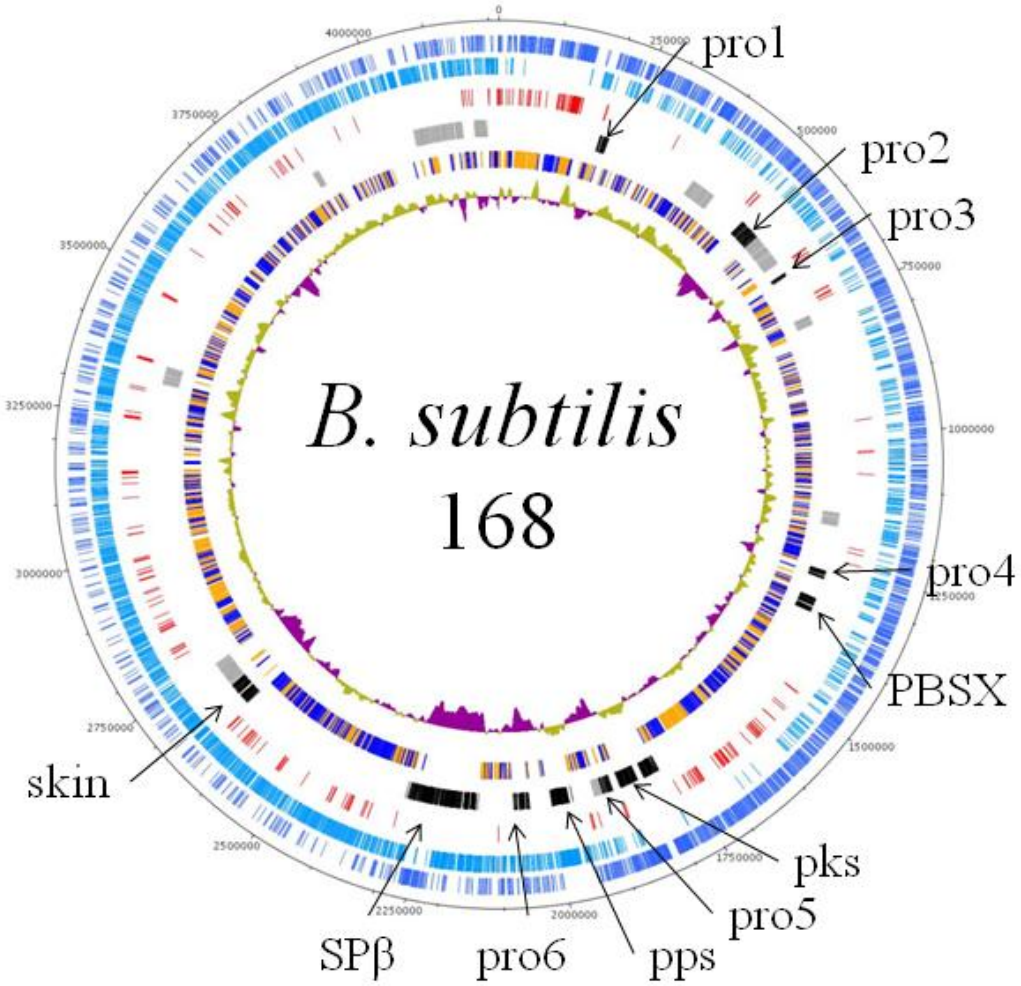


Abbildung 31: Grafische Darstellung essentieller und deletierter Bereiche in *B. subtilis* 168
Die Ringe zeigen *B. subtilis* 168 mit seinen *forward* (blau) und *reverse* (hellblau) ORFs. Rot markiert sind die essentiellen Gene. Schwarz sind die ersten Deletionen, die dann um weitere Deletionsbereiche (grau) erweitert wurden und zu *B. subtilis* MGB876. Der orange / blaue – Ring zeigt die konservierten *cluster* zwischen *B. subtilis* 168 und *B. licheniformis* DSM13. Der innerste Ring stellt den GC-Gehalt von *B. subtilis* 168 dar.

Sowohl die als essentiell identifizierten Gene, als auch die deletierten Bereiche wurden mit Hilfe von BiBaG auf B. *licheniformis* DSM13 *gemappt*. Anschließend wurden zusammenhängende chromosomale Bereiche ausgewählt, sodass keine essentiellen Gene oder Proteine mit bidirektionalen *hits* zu *B. subtilis* 168 im Deletionsbereich liegen.

4.4.3.1 Daten

Für den BiBaG-Vergleich wurden die EMBL-Dateien von *B. subtilis* 168 und *B. licheniformis* DSM13 verwendet (Tabelle 10). Zusätzlich wurden die essentiellen und die deletierten Gene aus den Veröffentlichungen von Kobayashi (Kobayashi *et al.*, 2003) und Morimoto (Morimoto *et al.*, 2008) übernommen.

Tabellen mit den entsprechenden ORFs befinden sich im Anhang (Kapitel 9.1, „essentiell_subtilis.xls“, „deleted_regions_subtilis.xls“).

4.4.3.2 Ergebnisse

Beim BiBaG-*Mapping* (Kapitel 9.1, „biblast.xls“) der 271 essentiellen Proteine aus *B. subtilis* 168 (Kapitel 9.1, „essential_subtilis.xls“) auf *B. licheniformis* DSM13 konnten 268 Orthologe (Kapitel 9.1, „essential_dsm13.xls“) identifiziert werden. Das *mapping* der ersten Deletionen von 456 Genen (Kapitel 9.1, „deleted_regions_subtilis.xls“), die zu MGB469 führten, lieferten 103 potentiell deletierbare ORFs in *B. licheniformis* DSM13 (Kapitel 9.1, „mapped_first _deletions.xls“). Bei der anschließenden Deletion in *B. subtilis*, die zu MGB867 führte, wurden 438 ORFs deletiert (Kapitel 9.1, „deleted_regions_subtilis.xls“). Die Identifikation von Orthologen in *B. licheniformis* DSM13 lieferte weitere 231 potentiell deletiertbare Gene (Kapitel 9.1, „mapped_seconddeletions_including _first.xls“). Tabelle 16 zeigt eine Zusammenfassung der Analyse.

Tabelle 16: Zusammenfassung der BiBaG-*Mapping*-Analyse zur Identifikation von Deletionstargets

	essentielle Gene	**Genanzahl erste Deletionen**	**Genanzahl zweite Deletionen**	**Gesamtgenanzahl der Deletionen**
***B. subtilis* 168**	271	460	438	894
***B. licheniformis* DSM13**	268	103	231	334

Bei der Erweiterung der 334 potentiell deletierbaren Genbereiche um Proteine, die weder essentiell sind, noch bidirektionale *hits* zu anderen Proteinen in *B. subtilis* 168 zeigten, konnten zusätzlich 335 Deletionskandidaten (Kapitel 9.1, „mapped_ deletionresgions_enhanced.xls“) identifiziert werden. Auf Basis der kompletten Deletionstargetliste, die 669 Deletionskandidaten umfasst, wurden 11 zusammenhängende Genombereiche als Deletionstargets ausgewählt (Tabelle 17). Bei Deletion aller Deletionstargets würde sich ein um 253.665 Basenpaare reduziertes Genom ergeben.

Tabelle 17: Deletionstargets für *B. licheniformis* DSM13

	Funktionen	**Start**	**Stop**	**bp**
1	GerK-Operon, hyp. Prot.	447.992	484.939	36.947
2	GerP-Operon, Sporulation	1.173.043	1.191.082	18.039
3	PBSX-Prophage	1.312.595	1.345.262	32.667
4	YesL-YesX-Operon	1.356.552	1.374.478	17.926
5	Phage , hyp. Prot.	1.429.495	1.463.620	34.125
6	hyp. Prot.	1.504.875	1.537.730	32.855
7	Sporulation, multidrug-Resistenz, hyp. Prot.	1.897.791	1.917.461	19.670
8	Sporulation	2.137.787	2.148.371	10.584
9	hyp. Prot.	2.674.360	2.688.203	13.843
10	Antibiotikaresistenz-Proteine	2.722.575	2.733.069	10.494
11	hyp. Prot.	4.151.506	4.178.021	26.515
			Gesamt	**253.665**

Die Gensonne von *B. licheniformis* DSM13 (Abbildung 32) zeigt sowohl die *gemappten* essentiellen Proteine, als auch die *gemappten*, potentiell deletierbaren Bereiche.

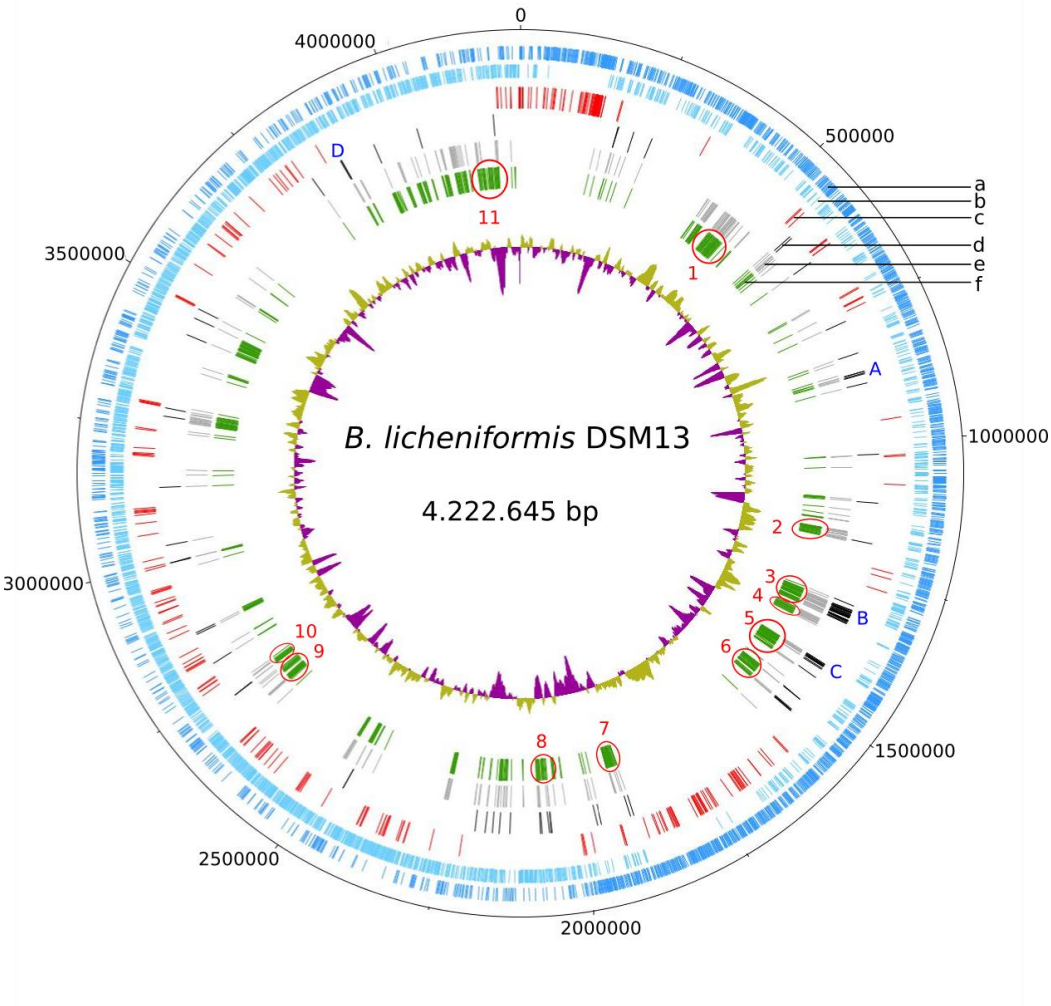


Abbildung 32: *B. licheniformis* DSM13 mit essentiellen und potentiell deletierbaren Gene
Die Ringe zeigen den Referenzorganismus *B. licheniformis* DSM13 mit seinen *forward* (a) und *reverse* ORFs (b). Ring (c) zeigt die essentiellen Gene aus *B. subtilis* 168 markiert, die Orthologe in *B. licheniformis* DSM13 haben. Die Orthologen der ersten Deletionen, die zu B. subtilis MGB469 führten, sind in Ring (d) dargestellt. Ring (d) zeigt die *gemappten* Orthologen, aller Deletionen, die zu *B. subtilis* MGB876 führten. Ring (d) stellt die um hypothetische Proteine erweiterten potentiellen Deletionstargets dar.
(A) zeigt einen Teil des *gemappten* Prophagen pro1. (B) ist der PBSX-Prophage. Bereich (C) ist ein Teil des skin-Elements bzw. ein Prophage und Bereich (D) enthält einen Teil des Prophagen pro5.
Die roten Zahlen 1 – 11 korrespondieren zu den in Tabelle 17 beschriebenen Deletionstargets.

4.4.4 Insertionstargetbestimmung für *B. licheniformis* DSM13

4.4.4.1 Daten

Zur Insertionstargetbestimmung wurden die EMBL-Dateien der drei nah verwandten Stämme *B. subtilis* 168, *B. amyloliquefaciens* FZB42 und *B. licheniformis* 9945A sowie der etwas entfernter verwandte *Geobacillus kaustophilus* HTA426 verwendet (Tabelle 10).

4.4.4.2 Ergebnisse

Die Bestimmung der möglichen Insertionstargets lieferte zwischen 418 und 1452 Proteine (Tabelle 18). Im Anhang befinden sich vier Listen (Kapitel 9.1, „möglicheInsertionsTargets_BAM.xls“, „möglicheInsertionsTargets_BSU.xls“, „möglicheInsertionsTargets_BLI.xls“, „möglicheInsertionsTargets_GBK.xls“) mit den entsprechenden Proteinen und ihren Annotationen. Bei den identifizierten Proteinen handelt es sich überwiegend um hypothetische Proteine. *B. licheniformis* 9945A kann im Gegensatz zu *B. licheniformis* DSM13 auf Urease wachsen. Das Operon, das für die Urease-Proteine kodiert, ist ein Insertionstarget, das bereits in *B. licheniformis* DSM13 integriert wurde (pers. Komm. M. Rachinger, Dissertation 2010).

Tabelle 18: Übersicht der Anzahlen möglicher Insertionsproteine

	***B. subtilis* 168 (BSU)**	***B. amyloliquefaciens* FZB42 (BAM)**	***B. licheniformis* 9945A (BLI)**	***Geobacillus kaustophilus* HTA426 (GBK)**
Anzahl der ORFs, die nicht in DSM13 vorkommen	1174	848	418	1452

5 Integration und Speicherung experimenteller und sequenzbasierter Daten

In diesem Kapitel wird eine Analyse der zu speichernden Daten (Kapitel 5.1) vorgenommen und darauf basierend ein Datenbankschema (Kapitel 5.2) entworfen. Kapitel 5.3 beschreibt, wie das Datenbankschema in eine relationale Datenbank übersetzt wird. Anschließend wird die Implementierung der Anwendungsebene (Kapitel 5.4) erläutert und abschließend ein Anwendungsbeispiel (Kapitel 5.5) die Fähigkeiten der DB demonstriert.

5.1 Daten

Bei den verwendeten Daten handelt es sich zum einen um sequenzbasierte Genomdaten aus öffentlichen Datenbanken (http://www.ebi.ac.uk), zum anderen auch um experimentelle Daten, die im Laufe des Projektes erstellt werden (Expressionsanalysen, Transposonmutagenesen).

Für die Sequenzdaten werden EMBL-Dateien (Abbildung 33) verwendet, die strukturiert das Genom eines Organismus mit seinen wesentlichen genetischen Eigenschaften beschreiben. Die DNA-Sequenz ist dabei ein ebenso wichtiger Bestandteil, wie die sog. *features*. Als *feature* ist alles zu betrachten, was durch die Sequenz kodiert wird. *features* haben somit immer eine Position auf der Sequenz. Dabei kann es sich zum Beispiel um Gene, Promotoren, Terminatoren, regulatorische RNAs o. ä. handeln. Jedes Feature besitzt eine funktionale Annotation. Daneben stehen für die sog. *coding sequences* auch die Translationen in die Aminosäuresequenzen zur Verfügung.

```
FT   gene            99924..100048
FT                   /locus_tag="BLi00100"
FT   rRNA            99924..100048
FT                   /locus_tag="BLi00100"
FT                   /product="5S ribosomal RNA"
FT   gene            100226..100690
FT                   /locus_tag="BLi00101"
FT                   /gene="ctsR"
FT   CDS             100226..100690
FT                   /locus_tag="BLi00101"
FT                   /protein_id="AAU39075.1"
FT                   /gene="ctsR"
FT                   /transl_table=11
FT                   /note="transcriptional repressor of class III stress
FT                   genes; RBL03205"
FT                   /db_xref="InterPro:IPR008463"
FT                   /db_xref="UniProtKB/TrEMBL:Q65PD9"
FT                   /codon_start=1
FT                   /translation="MGKNISDIIEQYLKQILEQNGKEILEIKRSEIADKFQCVPSQINY
FT                   VINTRFTSERGYIVESKRGGGGYIRIIKIKMNDKIDLINNIMNQIYTRLSQAASDDIIL
FT                   RLLENGVITESEAKLMVSVMDRSVLYIDLPERDELRARMMKAMLTSLKFK"
FT                   /product="CtsR"
FT   gene            100705..101259
FT                   /locus_tag="BLi00102"
FT                   /gene="mcsA"
FT   CDS             100705..101259
FT                   /locus_tag="BLi00102"
FT                   /protein_id="AAU39076.1"
FT                   /gene="mcsA"
FT                   /transl_table=11
FT                   /note="modulator of CtsR repression; RBL03206"
FT                   /db_xref="GOA:Q65PD8"
FT                   /db_xref="InterPro:IPR001943"
FT                   /db_xref="UniProtKB/TrEMBL:Q65PD8"
FT                   /codon_start=1
FT                   /translation="MICQECKERPATFHFTKVINGEKKEMHICEQCAKENSESYSMNES
FT                   GGFSIHNLLSGLLNFDSSFTNSSEAQLFQQPDQVLRCKKCNMTFPEFRKTGRFGCSECY
FT                   KTFHSYITPVLRKVHSGNTVHAGKIPKRIGGNLHVRRQIEALKKELKELIQQEEFEKAA
FT                   NIRDQIRSLEQNLNANKEEED"
FT                   /product="McsA"
```

Abbildung 33: Auszug aus der *B. licheniformis* DSM13 EMBL-Datei mit Beschreibung von RNA- und CDS-f*eatures*

Neben den Genomdaten werden im Laufe des Projektes Daten aus komparativen FASTA-Analysen (Pearson & Lipman, 1988) generiert, die Ähnlichkeiten verschiedener Proteine untereinander beschreiben. Es handelt sich im wesentlichen um FASTA3-Ausgabedateien (Abbildung 34), die zusätzlich zu den essentiellen Informationen, wie den Namen des *hits*, Identitätswert, *e-value* und *bit score*, auch Informationen zu den Trefferpositionen und Lücken enthalten.

```
  2>>>BLi00002 378 AA 378 aa - 378 aa
 vs  db.fasta library

6359671 residues in 22258 sequences
  Expectation_n fit: rho(ln(x))= 5.7624+/-0.000384; mu= 7.7782+/- 0.021
 mean_var=78.7821+/-16.609, 0's: 0 Z-trim: 9  B-trim: 3 in 1/49
 Lambda= 0.144498

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 37, opt: 25, open/ext: -10/-2, width:  16
 Scan time:  2.230
The best scores are:                                      opt bits E(22258)     %_id  %_sim   sw  alen  an0  ax0  pn0  px0  an1  ax1 pn1 px1 gapq gapl  fs
BL00077                                          ( 378) 2354 499.9  1e-141    1.000 1.000 2354  378    1  378    1  378    1  378    1  378   0   0   0
BLi00002                                         ( 378) 2354 499.9  1e-141    1.000 1.000 2354  378    1  378    1  378    1  378    1  378   0   0   0
BSU00020                                         ( 378) 2131 453.4  1e-127    0.905 0.968 2131  378    1  378    1  378    1  378    1  378   0   0   0
BcerKBAB4_0002                                   ( 381) 1838 392.4 2.5e-109    0.759 0.921 1838  381    1  378    1  378    1  381    1  381   3   0   0
BcerKBAB4_2456                                   ( 376) 1028 223.5 1.7e-58     0.425 0.792 1028  379    1  378    1  378    1  376    1  376   1   3   0
BLi02767                                         ( 159)  105 30.9    0.67     0.333 0.667  105   69  281  348    1  378   74  140    1  159   1   2   0
BL02108                                          ( 159)  105 30.9    0.67     0.333 0.667  105   69  281  348    1  378   74  140    1  159   1   2   0

>>>BLi00002, 378 aa vs db.fasta library

  3>>>BLi00003 71 AA 71 aa - 71 aa
 vs  db.fasta library

6359671 residues in 22258 sequences
  Expectation_n fit: rho(ln(x))= 4.7148+/-0.000355; mu= 4.8815+/- 0.020
 mean_var=53.9281+/-11.290, 0's: 2 Z-trim: 7  B-trim: 0 in 0/50
 Lambda= 0.174649

FASTA (3.5 Sept 2006) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 36, opt: 24, open/ext: -10/-2, width:  16
 Scan time:  0.760
The best scores are:                                      opt bits E(22258)     %_id  %_sim   sw  alen  an0  ax0  pn0  px0  an1  ax1 pn1 px1 gapq gapl  fs
BLi00003                                         (  71)  458 122.2 1.8e-29    1.000 1.000  458   71    1   71    1   71    1   71    1   71   0   0   0
BL00078                                          (  71)  458 122.2 1.8e-29    1.000 1.000  458   71    1   71    1   71    1   71    1   71   0   0   0
BSU00030                                         (  71)  416 111.7 2.7e-26    0.900 0.971  416   70    1   70    1   71    1   70    1   71   0   0   0
BcerKBAB4_0003                                   (  70)  302 82.9 1.2e-17     0.687 0.881  302   67    3   69    1   71    2   68    1   70   0   0   0
```

Abbildung 34: Auszug aus einer FASTA Ausgabedatei

Den wesentlichen Bestandteil der experimentellen Daten bilden die Expressionsanalysen von *B. licheniformis* DSM13 in verschiedenen Wachstumsphasen und auf unterschiedlichen Medien (pers. Kommunikation M. Schwarzer, Dissertation 2010). Dazu wurden jeweils zwei Bedingungen auf einem *chip* hybridisiert. Als gemeinsame Referenz wurde das Wachstum von *B. licheniformis* DSM13 auf Glukose gewählt. Die gemessenen Werte wurden nach statistischer Relevanz vorgefiltert.

Zum einen ist eine textuelle Beschreibung des Experimentes und seinen Parametern notwendig. Dazu wurde festgelegt, dass ein Experiment mit folgenden fünf Informationen beschrieben werden soll: Flussrate, Temperatur, *microarray*-Plattform (z. B. *selfmade*, Affymetrix), Kulturtyp (z. B. *fedbatch*, kontinuierlich), Kommentar. Die Informationen werden zeilenweise hintereinander in einer Datei hinterlegt (Abbildung 35) und sollen dann in die Datenbank eingelesen werden können.

```
A  3.5
B  37
C  selfmade
D  batch
E  experiment erfolgreich
```

Abbildung 35: Beispiel für eine Datei, die ein Experiment beschreibt
(A) Flussrate, (B) Temperatur, (C) *microarray*-Plattform, (D) Kulturtyp, (E) Kommentar

Zum anderen sollen die erhaltenen Expressionswerte gespeichert werden. Zur besseren Bewertung der Qualität des Experiments, werden die absoluten Werte der Rot/Grün-Messung jeweils sowohl für den Vordergrund, als auch für den Hintergrund benötigt.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
0	1	2	6	BLi00486	NasF: uroporphyrin-III C-methylt	458	149	308	94	214	705	225	480	88	392	-0,64	0,64	0,67	0,56	17,41
0	1	3	6	BLi00486	NasF: uroporphyrin-III C-methylt	517	147	369	112	257	749	220	528	92	436	-0,52	0,7	0,75	0,74	6,43
0	1	6	6	BLi00304	Mta: transcriptional activator o	1305	151	1153	210	943	1390	224	1166	89	1077	-0,02	0,99	0,98	0,99	1,61
0	1	7	6	BLi00304	Mta: transcriptional activator o	1120	141	978	83	895	1303	221	1082	108	974	-0,15	0,9	0,87	0,95	8,6
0	1	8	6	BLi01927	CotE: spore coat protein (outer)	700	140	559	110	449	868	218	650	77	573	-0,22	0,86	0,91	0,91	5,7
0	1	9	6	BLi01927	CotE: spore coat protein (outer)	681	137	543	142	401	838	206	631	117	514	-0,22	0,86	0,85	0,85	1,51

Abbildung 36: Auszug aus einer Datei, die ein *microarray*-Ergebnis enthält
Bedeutung der einzelnen Spalten: (A)-(D) interne Positionen auf dem *array*; (E) *locus tag*; (F) Annotation, (G) – (P) Rohwerte der Vorder- und Hintergrundauswertungen für Rot- und Grün-Werte; (Q) *log ratio*; (R) *ratio of medians*; (S) *ratio of means*; (T) *rgn ratio*; (U) Abweichung der *ratios* in %

Die Transposonmutagenesedaten lagen bis zur Fertigstellung dieser Arbeit nicht vor und wurden deshalb in den weiteren Schritten nicht berücksichtigt.

5.2 Datenbankentwurf

Auf Basis der in 5.1 beschriebenen Daten wurde ein *Entity-Relationship*-Modell (ER-Modell) (Chen, 1976) entwickelt, das die komplexen Daten und ihre Beziehungen zueinander modelliert. Dieses Modell dient als Brücke zwischen der realen Sicht auf die Daten und der Implementierung als physische Datenbank.

Entitäten sind in diesem Fall spezielle Informationen, die sich aus den Daten ergeben, wie zum Beispiel ein Organismus oder ein *feature*. Jeder dieser Informationen können Attribute hinzugefügt werden, die die Entität näher beschreiben, wie der *organismname*. Der Primärschlüssel identifiziert eine Entität eindeutig. Entitäten stehen in unterschiedlichen Beziehungen (Relationen) zueinander. Ein *feature* kann genau einer DNA-Sequenz zugeordnet werden, aber jede DNA-Sequenz kann mehrere *features* haben. Solch eine Beziehung wird als 1:n – Beziehung beschrieben. Es gibt aber auch 1:1 oder n:m – Beziehungen.

Dem hier entwickelten ER-Modell (Abbildung 37) liegen im Wesentlichen folgende Annahmen zugrunde:

- Es gibt die Entitäten: Organismus, DNA-Sequenz, Feature, Zugriff, Benutzer, Ähnlichkeit, Transkriptom und Experiment jeweils mit ihren typspezifischen Attributen und dem Primärschlüssel

- Es gibt binäre Beziehungen zwischen den Entitäten:
    - Feature <> DNA-Sequenz (n:1): Jedes *feature* wird einer DNA-Sequenz zugeordnet, aber eine DNA-Sequenz kann mehrere *features* haben.
    - DNA-Sequenz <> Organismus (n:1): Jede DNA-Sequenz wird einem Organismus zugeordnet, aber ein Organismus kann mehrere DNA-Sequenzen (*contigs*, *replikons*) haben.
    - DNA-Sequenz <> Zugriff (1:n): Jede DNA-Sequenz hat einen Zugriff bzw. ein Zugriffsrecht, aber ein Zugriffsrecht kann für mehrere DNA-Sequenzen gelten.
    - Zugriff <> Benutzer (n:1): Jedem Benutzer ist ein Zugriffsrecht zugeordnet, aber ein Zugriffsrecht kann mehrere Benutzer haben.
    - Feature <> Annotation (1:n): Jedes *feature* hat eine Annotation, aber eine Annotation kann für mehrere *features* gelten.
    - Feature <> Transkriptom (n:1): Jedem Transkriptionswert kann ein *feature* zugeordnet werden, aber ein *feature* kann mehrere Transkriptionswerte haben.
    - Transkriptom <> Experiment (1:n): Jedes Experiment hat viele Transkriptionswerte, aber ein Transkriptionswert kann nur ein Experiment haben
- Es gibt eine ternäre Beziehung zwischen den Entitäten:
    - Feature <> Feature <>Ähnlichkeit (n:m:1): Zwei *features* haben einen Ähnlichkeitswert zueinander.

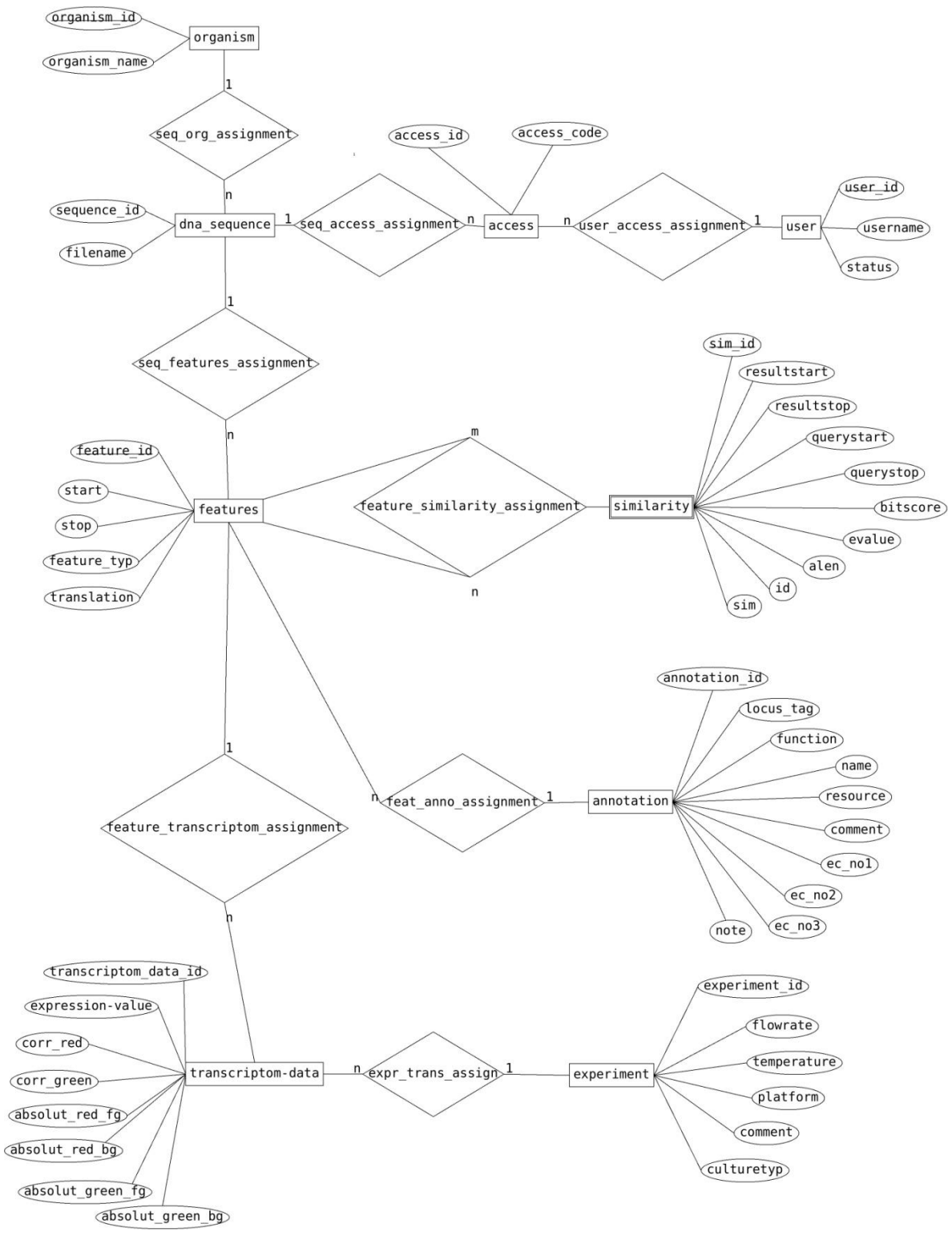


Abbildung 37: ER-Modell der zu entwickelnden Datenbank
Rechtecke entsprechen den Entitäten; Rauten zeigen Relationen; Ovale geben die Attribute der Entitäten an

5.3 Implementierung der relationalen Datenbank

Das *Entitiy-Relationship*-Modell (Abbildung 32) wurde in ein Datenbankschema übertragen (Chen, 1976) und im relationale Datenbanksystem PostgreSQL 8.3 (http://www.postgresql.de/) implementiert.

Jede Entität entspricht dabei einer Tabelle in der Datenbank. Eines der spezifischen Attribute entspricht dem Primärschlüssel oder falls keines der Attribute eindeutig ist, wird eine Spalte mit einer ID hinzugefügt, die die Entität eindeutig macht.

Für 1:n – Beziehungen erhält die Tabelle mit der Kardinalität n eine zusätzliche Spalte, die als Fremdschlüssel den Schlüssel der Tabelle mit der Kardinalität 1 erhält. Bei der ternären n:m:1 – Beziehung muss eine zusätzliche Tabelle erzeugt werden, die einen zusammengesetzten Primärschlüssel erhält, der aus den Primärschlüsseln der einzelnen Entitäten besteht.

Das Ergebnis der Ableitung von Tabellen aus dem ER-Modell ist in Tabelle 19 dargestellt. Die acht Tabellen „Organismus“, „DNA-Sequenz“, „Features“, „Annotation“, „TranskriptomDaten“, „Experiment“, „Ähnlichkeit“ und „FeatureÄhnlichkeit“ wurden erstellt. Sie sind in Tabelle 19 mit ihren jeweiligen Spaltennamen angegeben. Jede Tabelle besitzt mindestens eine ID-Spalte, die die Eindeutigkeit der in der Tabelle zu speichernden Daten sicherstellt. Für die ternäre Beziehung zwischen zwei *features* und ihrer Ähnlichkeit, wurde zusätzlich die Tabelle FeatureÄhnlichkeit-Verknüpfung angelegt. Die IDs der beiden zu verknüpfenden *features* bilden den Fremdschlüssel.

Tabelle 19: Überblick der erzeugten Tabellen im Datenbanksystem
Tabellennamen sind fett markiert; Tabellenspaltennamen sind angegeben; Primärschlüssel sind einfach Fremdschlüssel doppelt unterstrichen

Organismus	**DNA-Sequenz**	**Features**	**Annotation**
Organismus ID	Sequenz ID	Feature ID	Annotation ID
Organismusname	Dateiname	Start	Herkunft
Kürzel	Organismus ID	Stop	Funktion
		Featuretyp	Kommentar
		Sequenz ID	EC Nummer 1
		Annotation ID	EC Nummer 2
		Translation	EC Nummer 3
			Bemerkung
			Locustag
			Featurename
TranskriptomDaten	**Experiment**	**Ähnlichkeit**	**FeatureÄhnlichkeit-Verknüpfung**
TranskriptomDaten ID	Experiment ID	Ähnlichkeit ID	FeatureÄhnlichkeit ID
Expressionswert	Flussrate	Evalue	Feature ID
Feature ID	Plattform	Bitscore	Feature ID
absolutRotFG	Temperatur	Feature ID	
absolutRotBG	Kulturtyp	ProzentÄhnlichkeit	
absolutGrünFG	Kommentar	AlignmentLänge	
absolutGrünBG		QueryStart	
KorrelationRot		QueryStop	
KorrelationGrün		ResultStart	
Experiment ID		ResultStop	

In Kapitel 5.5.2 wird beschrieben, wie eine Beispiel-Datenbank in PostgreSQL 8.3 angelegt werden kann. Im Anhang (Kapitel 9.1) ist das psql-Skript „init_db_2010“ hinterlegt, das die Tabellen anlegt.

5.4 Implementierung der Anwendungsebene

Die abgeleiteten Tabellen (Kapitel 5.3) sind die physische Basis des Datenbanksystems. Um die Datenbank nutzen zu können, muss eine Anwendungsebene implementiert werden, die Methoden zur Datenspeicherung und -abfrage bereitstellt. Die Implementierung erfolgte in Java (www.java.com). Im Anhang befindet sich die Datei DBTools.jar, die den Programmcode enthält. Abbildung 38 zeigt die schematische Darstellung der Java-Klassen in UML (http://www.uml.org) und in Tabelle 20 sind alle Hauptmethoden kurz beschrieben.

Es gibt neun Klassen, die jeweils für spezifische Aufgaben konzipiert sind. Die Objekte der Klasse „Feature“ enthalten alle notwendigen Informationen zu einem *feature*, wie u. a. den *locus tag*, Start- und Stoppposition und ggf. die Übersetzung in Aminosäuren. Da alle kodierenden Eigenschaften der DNA-Sequenz, die mit Start- und Stoppposition angegeben werden können, als *feature* betrachtet werden, sind auch RNAs oder Operons *features*. Feature-Objekte werden von der Klasse „EMBLParser“ erzeugt. Die Klasse „EMBLParser“ enthält die Methoden, die zum Einlesen einer EMBL-Datei und zum Erstellen eines Feature-Objektes notwendig sind. Außerdem gibt es die Klassen „Experiment“, „ExperimentalData“ und „FastaResult“, die ebenfalls Methoden zum Einlesen und Filtern der für die jeweiligen Objekte notwendigen Daten (Kapitel 5.2) breitstellen.

Die zentrale Klasse, die das Speichern von Daten in das Datenbanksystem über einen jdbc:postgresql-Treiber (http://jdbc.postgresql.org) realisiert, ist der „DBconnector“. Für jede der acht Tabellen gibt es in der „DBconnector“-Klasse eine eigene Methode zur Datenspeicherung. Die Methoden insertAnnotation(), insertOrganism(), insertSequence() und insertFeature() werden mit einem Feature-Objekt aufgerufen.

Die insert-Methoden insertExperiment(), insertExperimentalData() und insert FastaResult() benötigen jeweils spezifische Objekte (Experiment, Experimental Data und FastaResult) mit denen sie aufgerufen werden können.

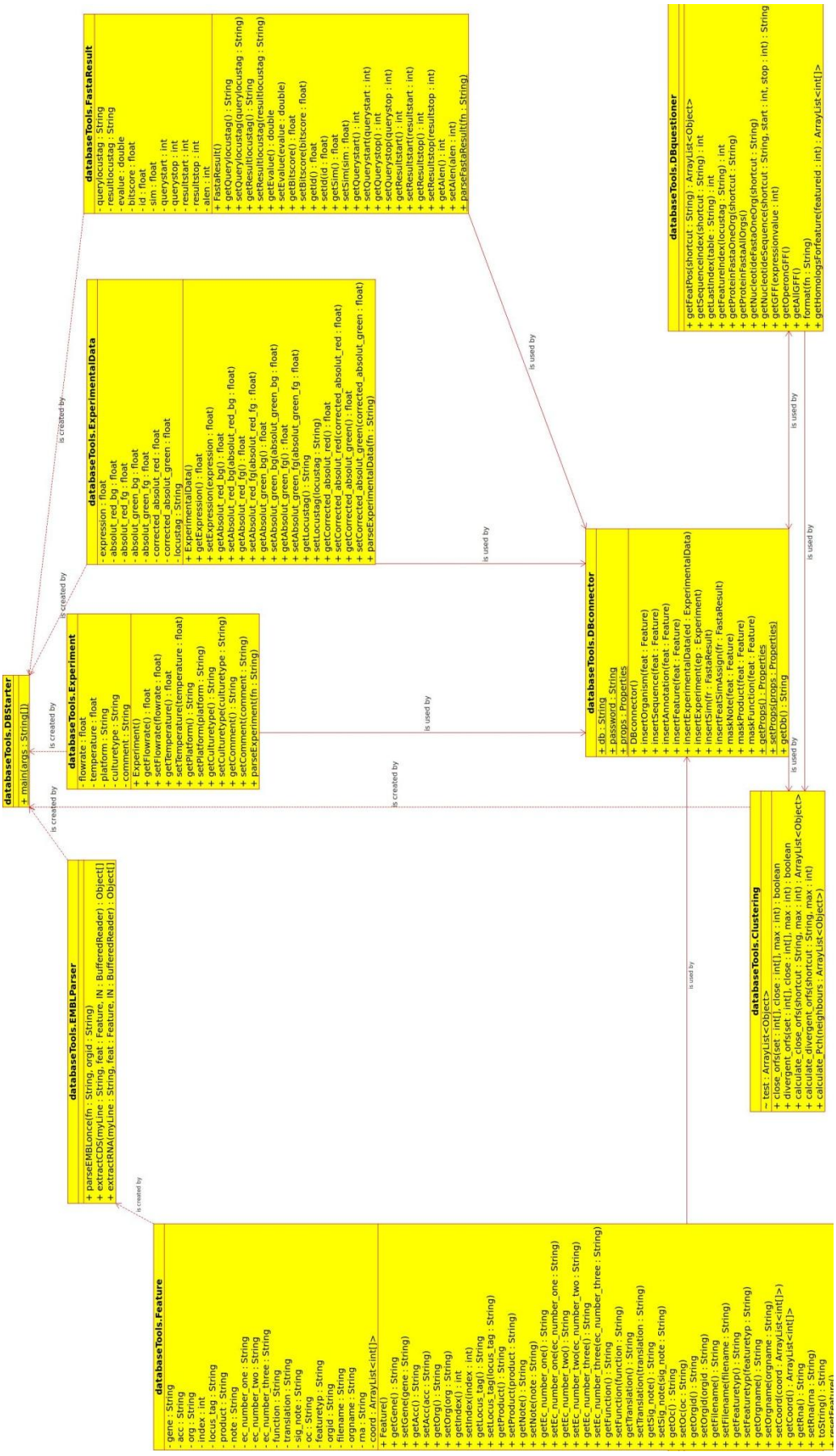


Abbildung 38: UML-Diagramm der entwickelten Java Methoden zum Zugriff auf die Datenbank

Mit der Klasse „Clustering“ werden Methoden bereitgestellt, die direkt mit den *features* auf der DNA-Sequenz arbeiten. Die Methode calculate_close_ORFs() berechnet für einen Organismus und einen vorgegebenen maximalen Abstandswert der Gene Operone. Dies stellt eine rein sequenzbasierte, einfache Operonvorhersage dar. Dabei werden diejenigen benachbarten Gene eines Stranges als Operon bezeichnet, die eine geringere Distanz, als den maximalen Abstandswert zueinander haben.

Tabelle 20: Kurzbeschreibung aller in einer Klasse implementierten Methoden

Klasse	Methodenname	Kurzbeschreibung
Clustering	**close_orfs()**	Prüft, ob zwei ORFs auf dem gleichen Strang nah beieinander sind
	divergent_orfs()	Prüft, ob zwei ORFs auf unterschiedlichen Strängen nah beieinander sind
	calculate_close_orfs()	Berechnet alle benachbarten, co-linearen ORFs und speichert sie als Feature Operon in der DB ab
	calculate_divergent_orfs()	Berechnet alle benachbarten divergenten ORFs und gibt sie in einer Datei aus
	calculate_Pch()	Berechnet Paare von nah benachbarten ORFs in verschiedenen Organismen und gibt sie in einer Datei aus
DBconnector	**insertOrganism()**	Speichert einen Organismus in der DB
	insertSequence()	Speichert eine Sequenz in der DB
	insertAnnotation()	Speichert eine Annotation in der DB
	insertFeature()	Speichert ein Feature in der DB
	insertExperimentalData()	Speichert experimentelle Werte in der DB
	insertExperiment()	Speichert Informationen zu einem Experiment in der DB
	insertSim()	Speichert Informationen zu einer Homologie in der DB
	insertFeatSimAssign()	Legt die Verknüpfung zwischen zwei Features und der Homologie an
DBquestioner	**getFeatPos()**	Liefert ein Java-Objekt mit allen *feature*-Positionen eines Organismus
	getSequenceIndex()	Liefert für einen Organismus die zugehörige SequenzID
	getLastIndex()	Liefert den letzten vergebenen Index für eine Tabelle
	getFeatureIndex()	Liefert für einen *locustag* die FeatureID
	getProteinFastaOneOrg()	Schreibt eine multi-Fasta-Datei mit den Proteinsequenzen für alle CDS eines Organismus
	getProteinFastaAllOrgs()	Schreibt eine multi-Fasta-Datei mit den Proteinsequenzen für alle CDS in der DB
	getNucleotideFastaOneOrg()	Schreibt eine multi Fasta-Datei mit den Nukleotidsequenzen für alle CDS eines Organismus

Klasse	Methodenname	Kurzbeschreibung
DBquestioner	**getNucleotideSequence()**	Liefert einen String mit einer Nukleotidsequenz eines Organismus von frei wählbaren Positionen
	getGFF()	Schreibt eine GFF-Datei mit allen regulierten Genen über einem angegeben Schwellwert
	getOperonGFF()	Schreibt eine GFF-Datei mit allen Operons (Artemis Tag: operon)
	getAllGFF()	Schreibt eine GFF-Datei mit allen regulierten Genen (Artemis Tag: high, low, undef)
	format()	Formatiert eine Fasta-Datei, so dass genau 60 Zeichen in einer Zeile sind
	getHomologsForFeature()	Liefert ein Objekt, dass die Homologen *features* für ein *feature* enthält
EMBLParser	**parseEMBLonce()**	Liest eine EMBL-Datei und speichert alle relevanten Informationen in der DB oder schreibt sie als Dateien
Experiment	**extractCDS()**	Liest alle Informationen für ein Protein-*feature* und speichert sie in einem *feature*-Objekt
	extractRNA()	Liest alle Informationen für ein RNA-*feature* und speichert sie in einem *feature*-Objekt
	parseExperiment()	Liest Experimentinformationen ein und speichert sie in der DB
Experimental Data	**parseExperimentalData()**	Liest Werte eines *microarray*-Experimentes ein und speichert sie in der DB
FastaResult	**parseFastaResult()**	Liest eine FASTA-Result-Datei ein und speichert die Daten in der DB
Feature		Klasse zum Erzeugen von *feature*-Objekten

Der „DBquestioner“ erlaubt das Auslesen von Daten aus der Datenbank. Neben den Standardabfragen der einzelnen Tabellen, wurden vier Methoden entwickelt, die spezifischere Ausgaben erzeugen. GFF-Dateien sind eine Möglichkeit strukturiert Positionen einer Nukleotidsequenz abzuspeichern und eignen sich als Eingabedateien für genomische Visualisierungs-Werkzeuge, wie z. B. Artemis (Rutherford *et al.*, 2000) oder DNAPlotter (Carver *et al.*, 2009). Gene, die in einem Experiment einen gewissen Expressionswert überschreiten werden durch getGFF() als GFF-Datei ausgegeben. Mit Hilfe von getAllGFF() wird die GFF-Ausgabedatei „experiment.gff“ erzeugt, die alle Gene mit Expressionswerte über 5 mit „high“, unter 3 mit „low“ und dazwischen mit undefiniert markiert. Die Methode getOperonGFF() erstellt aus allen *features*, die in der Datenbank mit Featuretyp „Operon“ versehen sind die GFF-Datei „operon.gff“ mit der Markie-

rung „operon“. Mit Hilfe von Artemis oder dem DNAPlotter und einer entsprechend angepassten Options-Datei, die für jede Markierung eine Farbe vergibt, lassen sich die regulierten Gene (Abbildung 47) bzw. Operone (Abbildung 48, 49) visuell darstellen (Kapitel 9.1, „experiment.gff“, „operon.gff“) Dazu werden die GFF-Dateien zur EMBL-Datei des korrespondierenden Organismus hinzu geladen.

5.5 Anwendungsbeispiel *B. licheniformis* DSM13

5.5.1 Verwendete Daten

Um die Fähigkeiten der Datenbank zu demonstrieren, wurden vier Organismen in die DB geladen. Es handelt sich um die EMBL-Dateien von *B. licheniformis* DSM13 (AE017333), *B. licheniformis* ATTC14580 (CP000002), *B. subtilis* 168 (AL0009126) und *B. weihenstephanensis* KBAB4 (CP000903). Die Proteinsequenzen von *B. licheniformis* DSM13 wurden mittels FASTA-Algorithmus mit den übrigen drei Organismen verglichen. Die Ausgabedatei „bl_fasta.result“ stellt das Ergebnis dar. Außerdem wurde ein *microarray*-Experiment (pers. Kommunikation M. Schwarzer, Dissertation 2010) verwendet, das zwei Wachstumsbedingungen vergleicht: Zum einen als Referenz das Wachstum auf Glukose und zum anderen das Wachstum von *B. licheniformis* DSM13 auf einem 7SB-Mix zum Zeitpunkt T1 (Anfang der logarithmischen Phase). Der 7SB-Mix ist ein Medium, das der Sojabohne nachempfunden ist und 7 Aminosäuren enthält. Die Datei „experiment“ beschreibt beispielhaft das Experiment und in der Datei „7SB“ finden sich die nach statistischer Signifikanz vorgefilterten Ergebnisdaten des *microarray*-Experiments. Alle hier genannten Dateien sind im Anhang (Kapitel 9.1) hinterlegt.

5.5.2 Schritt für Schritt Anleitung

1) PostgreSQL 8.3 installieren http://www.postgresql.org
2) Anlegen der Datenbank:

   createdb mydb (evtl. nur als Superuser postgres möglich)
3) Anlegen des DB Schemas:

   psql -d mydb < init_db_2010

4) Automatisches Laden der Testdaten in die Datenbank und Generierung der GFF-Dateien:

   java -jar DBTools.jar

(Starten des JAR-Archivs in dem Ordnern, in dem sich auch die Testdaten (Kapitel 5.5.1) befinden; evtl. Anpassung der Variablen db (*default*: jdbc:postgresql://localhost:5432/mydb“), user (*default*: guest) und password (*default*: password) in DBconnector.java nötig)

5) Anzeigen der Ergebnisdateien „experiment.gff“ und „operon.gff“ mit Artemis bzw. DNAPlotter (Kapitel 9.1)
   Laden der AE017333 EMBL-Datei von *B. licheniformis* DSM13 in Artemis und Hinzufügen der experiment.gff, operon.gff

5.6 Ergebnisse

Abbildungen 39 – 41 zeigen exemplarisch die ersten Zeilen der jeweiligen Tabelle mit den eingefügten Daten. Für die DNA-Sequenz (Abbildung 40 D) wird jeweils nur der EMBL-Dateiname der jeweiligen Datei gespeichert. Abbildung 41 zeigt, dass für das *feature* mit der *featureid* 1 zwölf Orthologe gefunden werden konnten und ihm die *annotationid* 1 sowie *sequenceid* 1 zugeordnet wird. Die *similarity*-Tabelle zeigt die weiteren Details der jeweiligen Orthologen (Abbildung 40 E). Bei dem *feature* mit der *featureid* 1 handelt es sich laut der Annotationstabelle um *dnaA* (Abbildung 39). Abbildung 40 D zeigt, dass der *sequnceid* 1 die *organismid* 1 zugeordnet ist. Abbildung 40 C gibt Aufschluss darüber, dass es sich bei dem entsprechenden Organismus um *B. licheniformis* DSM13 handelt.

Außerdem wurde ein Experiment (Abbildung 40 A) angelegt, das Transkriptomdaten (Abbildung 40 F) enthält, die wiederum bestimmten *features* zugeordnet werden.

	annotationid [PK] serial	resource character	function character	comment character	ecnumber character	note text	locustag character	featurename character varying(100)	ecnumber2 character varying(20)	ecnumber3 character varying(20)
1	1	extract fron	DnaA: initia	no commen	''	no note	BLi00001	dnaA	''	''
2	2	extract fron	DnaN: DNA	no commen	''	no note	BLi00002	dnaN	''	''
3	3	extract fron	YaaA: simila	no commen	''	no note	BLi00003	yaaA	''	''
4	4	extract fron	RecF: DNA r	no commen	''	no note	BLi00004	recF	''	''
5	5	extract fron	YaaB: RBL05	no commen	''	no note	BLi00005	yaaB	''	''
6	6	extract fron	GyrB: DNA g	no commen	''	no note	BLi00006	gyrB	''	''
7	7	extract fron	GyrA: DNA g	no commen	''	no note	BLi00007	gyrA	''	''
8	8	extract fron					BLi00008	16S ribosomal RNA		
9	9	extract fron					BLi00009	tRNA-OTHER		
10	10	extract fron					BLi00010	tRNA-OTHER		
11	11	extract fron					BLi00011	23S ribosomal RNA		
12	12	extract fron					BLi00012	5S ribosomal RNA		
13	13	extract fron	YaaC: simila	no commen	''	no note	BLi00013	yaaC	''	''

Abbildung 39: Tabellenauszug *annotation*

A

	experimentid [PK] serial	flowrate double precision	platform character varying(50)	temperature double precision	culturetyp character varying(50)	comment text
1	1	3.5	selfmade	37	batch	experiment erfolgreich
*						

B

	featuresimid [PK] integer	featureid1 integer	featureid2 integer
1	1	1	4290
2	2	1	8562
3	3	1	12787
4	4	1	11547
5	5	1	17206
6	6	1	10824
7	7	1	7274
8	8	1	2985
9	9	1	1296
10	10	1	5586
11	11	1	16262
12	12	1	9898
13	13	2	4291
14	14	2	8563
15	15	2	12788
16	16	2	15300
17	17	2	2710
18	18	2	7003
19	19	3	4292
20	20	3	8564
21	21	3	12789

C

	organismid [PK] serial	organism_name character varying(100)	shortcut character varying(5)
1	1	Bacillus licheniformis DSM13	BL
2	2	Bacillus licheniformis ATCC 14580	BLI
3	3	Bacillus subtilis subsp. subtilis str. 168	BS
4	4	Bacillus weihenstephanensis KBAB4	BWS
*			

D

	sequenceid [PK] serial	filename character varying(50)	organismid integer
1	1	AE017333	1
2	2	CP000002	2
3	3	AL009126	3
4	4	CP000903	4
*			

E

	similarityid [PK] integer	evalue double precision	bitscore double precision	identity double precision	sim double precision	alen integer	querystart integer	querystop integer	resultstart integer	resultstop integer
1	1	3.8e-169	591.5	1	1	446	1	446	1	446
2	2	1.3e-161	566.5	0.951	0.982	446	1	446	1	446
3	3	1.1e-142	503.7	0.832	0.948	446	1	446	1	446
4	4	0.001	41.4	0.215	0.526	302	2	293	27	305
5	5	0.0055	39	0.328	0.613	119	103	216	123	232
6	6	0.018	37.4	0.257	0.558	226	113	327	5	216
7	7	0.059	35.6	0.217	0.566	152	147	293	164	305
8	8	0.059	35.6	0.217	0.566	152	147	293	165	306
9	9	0.14	34.1	0.268	0.622	127	145	265	118	240
10	10	0.14	34.1	0.268	0.622	127	145	265	118	240
11	11	0.44	32.5	0.347	0.68	75	147	216	126	197
12	12	0.65	31.9	0.273	0.618	110	145	248	118	223

F

	transcriptomdataid [PK] serial	expressionvalue double precision	experimentid integer	featureid integer	absolut_red_fg double precision	absolut_red_bg double precision	absolut_green_fg double precision	absolut_green_bg double precision	corr_red double precision	corr_green double precision
1	1	-0.52	1	477	517	147	749	220	257	436
2	2	-0.02	1	298	1305	151	1390	224	943	1077
3	3	-0.15	1	298	1120	141	1303	221	895	974
4	4	-0.22	1	1891	700	140	868	218	449	573
5	5	-0.22	1	1891	681	137	838	206	401	514
6	6	-0.81	1	371	263	132	440	211	47	145
7	7	0.37	1	2418	3545	144	2865	242	3286	2505
8	8	0.43	1	2418	3349	141	2634	252	3047	2257
9	9	-1.28	1	297	6880	152	16544	225	6630	16238
10	10	-1.28	1	297	6678	151	16092	223	6399	15783
11	11	-2.22	1	3211	1152	147	4897	223	879	4561
12	12	-2.12	1	3211	1186	141	4762	228	909	4401

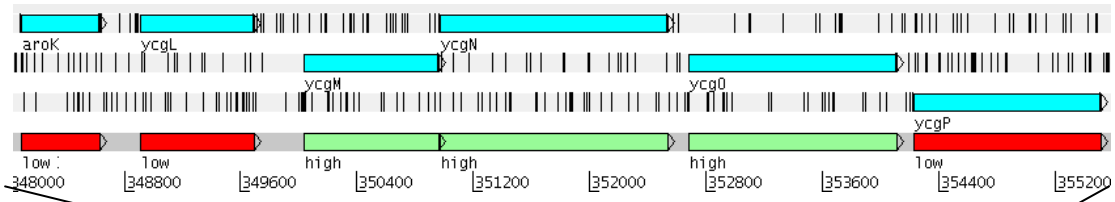
Abbildung 40: Verschiedene Tabellenauszüge aus der DB
(A) *experiment*; (B) *feat-sim-assign*; (C) *organism*; (D) *dna_sequence*; (E) *similarity*; (F) *transcriptom-data*

	featureid [PK] serial	start integer	stop integer	featuretyp character varying(50)	sequenceid integer	annotationid integer	translation text
1	1	311	1651	CDS	1	1	MKNISDLWNQALGQIEKKLSKPSFETWMKSTKAHSLQ
2	2	1828	2964	CDS	1	2	MKFTIQKDRLVESVQDVLKAVSSRTTIPILTGIKIVASDE
3	3	3132	3347	CDS	1	3	MTKTVAIDTEMITLGQFLKLADVIQSGGMAKWFLSEYE
4	4	3364	4476	CDS	1	4	MYIQNLTLSSYRNYERLDLQFENKVNVIIGENAQGKTNI
5	5	4494	4736	CDS	1	5	MYIHLGDDFVVSTREIVAIFDYKAKTSPIVEEFLSKQKQI
6	6	4796	6709	CDS	1	6	MEQQNNYDENQIQVLEGLEAVRKRPGMYIGSTSGKGL
7	7	6900	9368	CDS	1	7	MSEQHKPQVQEVNISQEMRTSFLDYAMSVIVSRALPDV
8	8	9710	11257	RNA	1	8	null
9	9	11358	11434	RNA	1	9	null
10	10	11446	11521	RNA	1	10	null
11	11	11592	14520	RNA	1	11	null
12	12	14644	14768	RNA	1	12	null
13	13	15762	14785	CDS	1	13	MQDKGWKDLKLFYSVETAQHFLEMVYKNQGMEDAK
14	14	15884	17350	CDS	1	14	MWESKFSKEGLTFDDVLLVPAKSEVLPRDVDLSVELTP
15	15	17505	18830	CDS	1	15	MKSKRLKQLIMLIVAFAVTVGAFSPMSTAKAANDPINV
16	16	19017	19901	CDS	1	16	MAQTGTDRVKRGMAEMQKGGVIMDVVNAEQAKIAEI
17	17	19923	20513	CDS	1	17	MLTIGVLGLQGAVREHIRSIEACGAAGKVIKWPEELKEI
18	18	20841	22118	CDS	1	18	MLDVKLLRANFEEIKQKLAHRGEDLSDFDQFEELDTKF
19	19	24562	23294	CDS	1	19	MDIQINAIGALCALVISIFLILKKVPPAYGMMIGALAGGL
20	20	25770	24652	CDS	1	20	MNYLSKELAEEIVERTMSIIRHNINMMDENGVIIASGDF
21	21	26222	25800	CDS	1	21	MWKHFISRLPQDYTLNRPIETGKLQKAEELLGVSFPDE

Abbildung 41: Tabellenauszug *features*

Die Visualisierung der aus der Datenbank generierten GFF-Dateien ist in Abbildung 42 A (Artemis) und 42 B (DNAPlotter) dargestellt.

A



B

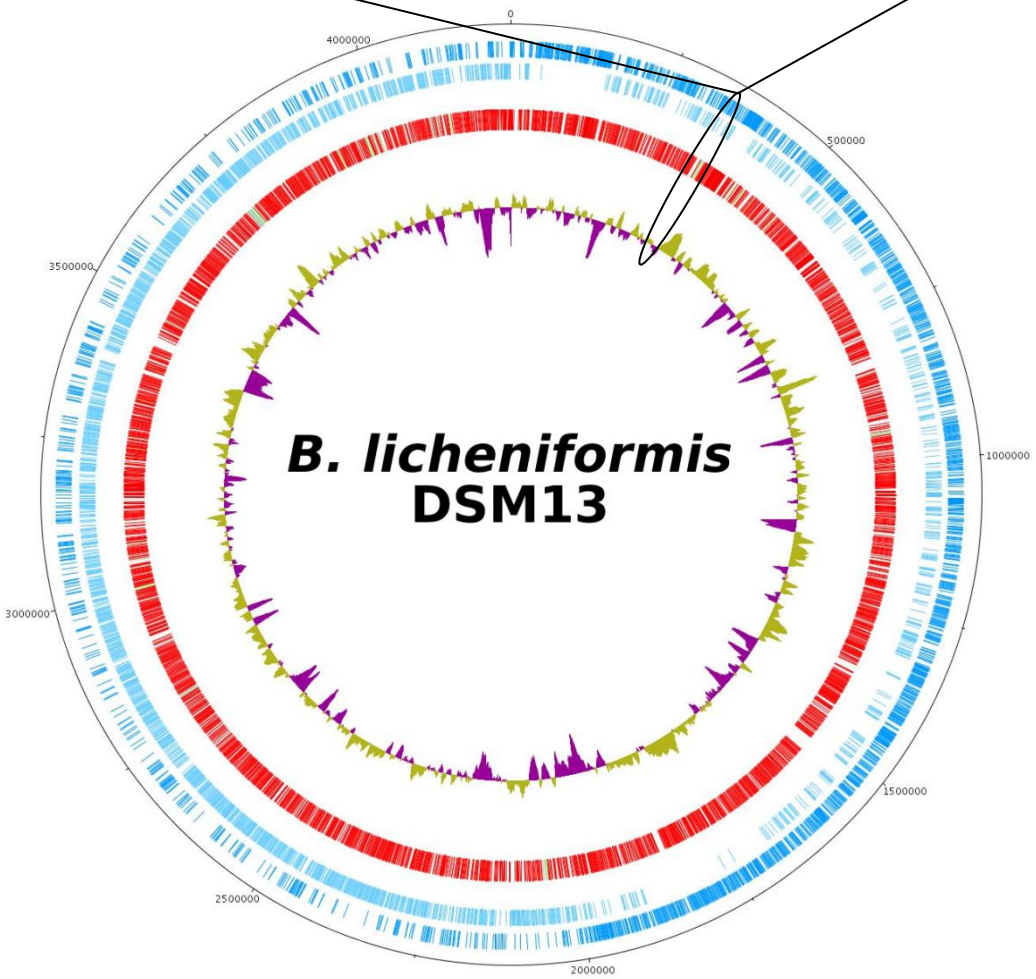


Abbildung 42: Visualisierung der Datei experiment.gff

(A) mit dem DNA Plotter: Gensonne mit den regulierten Genen aus DSM13 vom 7SB-Experiment zum Zeitpunkt T1, hochregulierte Gene (Expressionswert über 5) sind grün, herunter regulierte Gene (Expressionswerte unter 3) rot dargestellt.

(B) mit Artemis : ein Ausschnitt mit hochregulierten Genen ist heraus g*ezoomt* dargestellt.; hochregulierte Gene sind in grün dargestellt; herunter regulierte Gene sind rot markiert

Abbildung 42 (A) zeigt einen Ausschnitt des *B. licheniformis* DSM13 Genoms, der sowohl hoch- (grün) als auch herunterregulierte (rot) Gene darstellt. Die GFF-Datei ist im Anhang hinterlegt (Kapitel 9.1, „experiment.gff“). Insgesamt sind 59 Gene hoch- und 3153 Gene herunter reguliert. Außerdem gibt es 116 Gene, die möglicherweise gemeinsam reguliert sind, aber im Vergleich vom 7SB-Mix zu Glukosebedingungen keine signifikanten Regulationsunterschiede zeigen. Abbildung 42 B zeigt die Visualisierung der Experimentergebnisse für das gesamte Genom.

Der Abgleich mit der BiBaG-Analyse (Kapitel 4.4.1) zeigt, dass die hochregulierten Gene auch in anderen *Bacillus*-Spezies vorkommen (Kapitel 9.1, „BiBaG _Abgleich.xlsx). Außerdem treten sie häufig in Clustern von zwei oder mehr Genen auf, die gemeinsam hochreguliert sind. Diese Cluster sind teilweise auch konserviert. Das längste Cluster umfasst mit fünf Proteinen die ORFs BLi03054-BLi03058 aus *B. licheniformis* DSM13. Sie kodieren für zwei hypothetische Proteine sowie MutM , eine formamidopyrimidin DNA Glycosidase, PolA, die DNA Polymerase I, und PhoR, ein zwei Komponeten-System, das in die Phosphatregulation eingebunden ist . Diese Proteine sind in allen 29 untersuchten *Bacilli* konserviert und liegen außer in *B. halodurans* C-125 ge*cluste*r*t* vor (Kapitel 9.1, „mitPlasmiden/biblast.xlsx“).

Abbildung 43 zeigt einen Ausschnitt aus der „operon.gff“-Datei, die mit Artemis visualisiert wurde. Es sind zwei Operons dargestellt, deren Gene jeweils einen geringeren Abstand als 100 Basen auf dem gleichen Strang haben. Zwischen den beiden Operons befindet sich die RNA BLi00022.

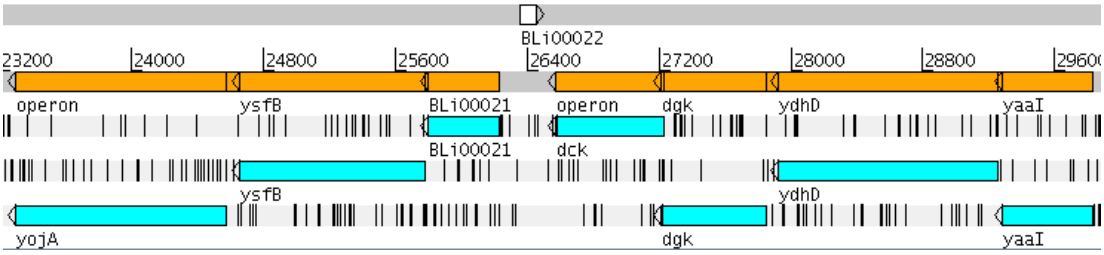


Abbildung 43: Visualisierung der Datei operon.gff in Artemis
Dargestellt sind zwei benachbarte Operons (orange)

Insgesamt konnten mit dieser auf chromosomaler Nachbarschaft beruhenden Methode 2199 Operons identifiziert werden, die in Abbildung 44 dargestellt sind. Das längste Operon ist 26.584 bp lang und kodiert hauptsächlich für Flagellen- und Chemotaxisproteine. Seine chromosomale Position in *B. licheniformis* DSM13 ist 1796486-1823070. Abbildung 28 zeigt die Homologie-Cluster der BiBaG-Analyse (Kapitel 4.4.1) mit fünf *Bacillus*-Stämmen. Die längsten Homologie-Cluster zwischen *B. licheniformis* DSM13, *B. subtilis* 168 und *B. amyloliquefaciens* FZB42 enthalten das hier beschrieben längste Operon (Kapitel 9.1, „bacillus_biblast.xls“). *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-215 zeigen keine Clusterbildung für diese Proteine und weisen einen deutlich sichtbaren Unterschied in der Sequenzähnlichkeit auf (Abbildung 27).

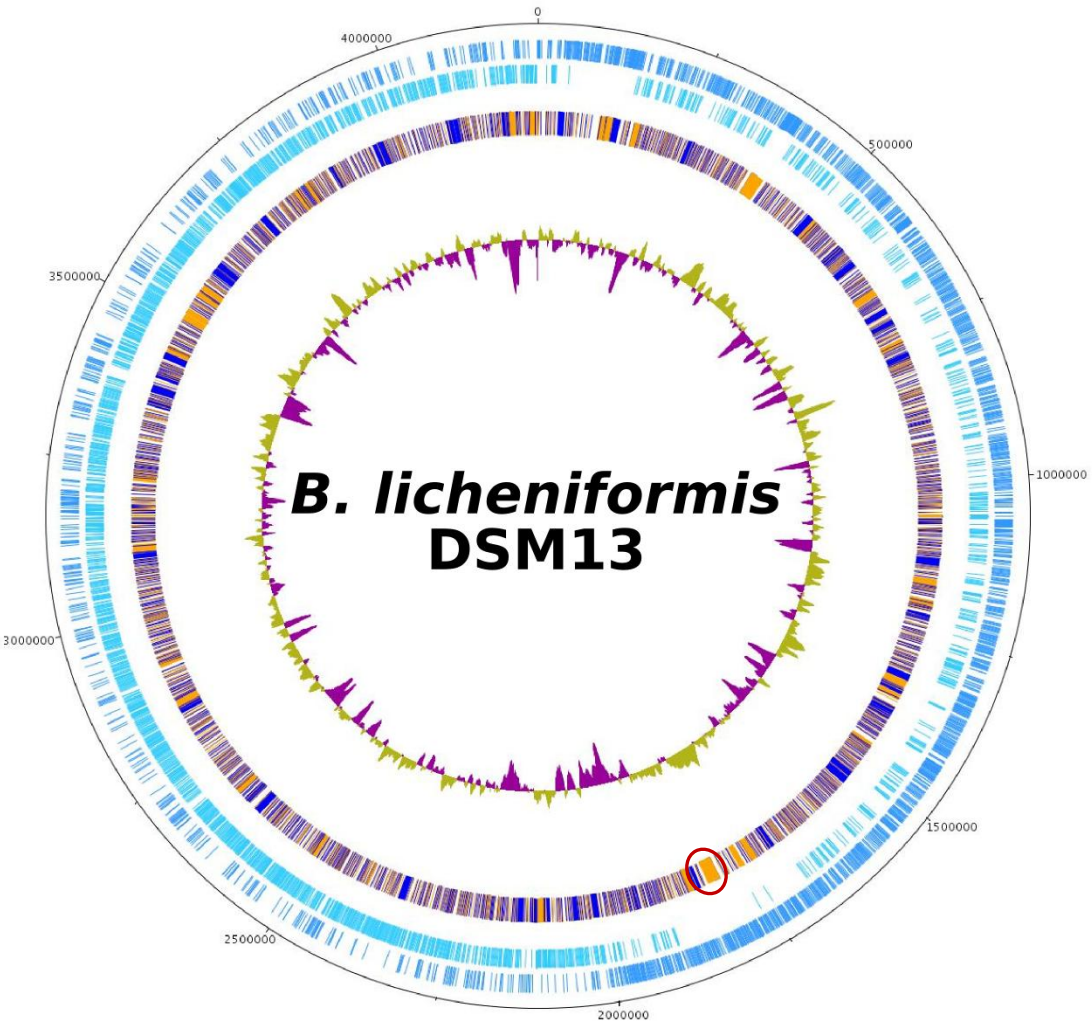


Abbildung 44: Gensonne von *B. licheniformis* DSM13 mit eingezeichneten Operons
Benachbarte Operons sind jeweils abwechselns orange und blau markiert. Rot markiert ist das längste Operon.

6 Diskussion

In den folgenden Unterkapiteln werden die Ergebnisse noch einmal zusammengefasst und im Hinblick auf aktuelle Forschungsergebnisse erläutert. Die komparative Analyse des Kompetenzsystems von *B. licheniformis* DSM13 (Kapitel 6.1) bildet den Ausgangspunkt für weitere Experimente zur Verbesserung der genetischen Manipulation und ist Teil der Basis für das entwickelte *software tool* BiBaG (Kapitel 6.2). Mit Hilfe von BiBaG kann eine spezifische Analyse bestimmter Regulationsmechanismen oder einzelner metabolischer *pathways* ausgehend von Gesamtgenomvergleichen vereinfacht durchgeführt werden. Die entwickelte Datenbank (Kapitel 6.3) bildet die Schnittstelle zwischen den experimentellen, sequenzbasierten und komparativen Datenmengen.

6.1 Komparative Analyse des Kompetenzsystems von DSM13

Die komparative Analyse des Kompetenzsystems zeigt, dass die bekanntermaßen kompetenten Spezies *B. subtilis* 168 (Kunst *et al.*, 1997) und *B. amyloliquefaciens* FZB42 (Chen *et al.*, 2007) alle untersuchten Kompetenz-gene besitzen (Kapitel 3.3.1). Allerdings sind die Kompetenzgene bis auf *ssb* nicht essentiell in *B. subtilis* 168 (Kobayashi *et al.*, 2003). Auch in *B. licheniformis* DSM13, *B. pumilus* SAFR-032 und *B. anthracis* str. Sterne konnten die Mehrzahl der Kompetenzgene, insbesondere die Strukturgene zur Aufnahme von DNA, identifiziert werden. Dieses Ergebnis legt die Vermutung nahe, dass auch andere *Bacillus*-Stämme, die die untersuchten Gene besitzen, natürlich kompetent werden können. An dieser Stelle sei noch darauf hingewiesen, dass nur 10 - 20 % der Zellen einer *Bacillus*-Kultur kompetent werden (Maamar & Dubnau, 2005). Für eine Population hat dies den Effekt, dass sie sich mit einigen Zellen zwar über genetische Veränderungen einen Vorteil verschaffen kann, aber gleichzeitig durch Sporenbildung anderer Zellen nicht davon abhängig ist. Dies ist konsistent mit der Beobachtung, dass sich eine *B. subtilis* - Kultur durch Zelldifferenzierung wie ein multizellulärer Organismus verhalten kann (Chai *et al.*, 2010).

B. licheniformis DSM13 ist im Vergleich zu *B. subtilis* und B. *amyloliquefaciens* nicht bzw. nur in geringem Maße natürlich kompetent (Veith *et al.*, 2004). Die bioinformatischen Ergebnisse legen nahe, dass dies an einer in *comP* integrierten

Transposase liegt. Darüber hinaus fehlen in *B. licheniformis* DSM13 die Regulatoren *rapC* und *phrC*. Für *B. subtilis* 168 ist beschrieben, dass ein *knock out* von *rapC* (Lazazzera *et al.*, 1999) bzw. die Repression von RapC durch CSF (Solomon *et al.*, 1996) zu einer Anreicherung von ComA~P und damit zu einer erhöhten Expression von *srfA* / *comS* führen.

comP ist als evolutionärer *hot spot* beschrieben (Tortosa *et al.*, 2001; Tran *et al.*, 2000). Das *comQXPA*-Cluster ist in einen konservierten und einen stammspezifischen Bereich geteilt, dessen Mitte in *comP* liegt. *comA* und der C-terminale Bereich sind in unterschiedlichen *Bacillus*-Stämmen stark konserviert, wohingegen der N-terminale Bereich von *comP, comX* und *comQ* nicht konserviert ist. Dies deckt sich mit den Ergebnissen der Kompetenzregulationsmodulanalyse verschiedener *B. licheniformis*-Stämme im Vergleich zu *B. subtilis* (Kapitel 3.3.3). Dieser Polymorphismus deutet auf ihre Koevolution hin (Tortosa *et al.*, 2001). In räumlich getrennten Isolaten von *Bacillus*-Stämmen verschiedener Wüstenstandorte konnten vier unterschiedliche Pheromontypen bestimmt werden (Ansaldi *et al.*, 2002). Das Ergebnis konnte für die Untersuchung verschiedener Isolate aus Bodenproben reproduziert werden (Stefanic & Mandic-Mulec, 2009). Damit ist gezeigt, dass verschiedene sehr nah verwandte *Bacillus*-Stämme unterschiedliche Pheromone zur Zell-Zell-Kommunikation nutzen.

Die Vermutung liegt nahe, dass auch innerhalb der *B. licheniformis*-Gruppe unterschiedliche Pheromone verwendet werden. Tabelle 5 zeigt, dass der stammspezifische Bereich des *comQXPA*-Clusters zwischen den isogenischen Stämmen *B. licheniformis* DSM13 und *B. licheniformis* ATTC14580 stark konserviert ist. Allerdings zeigt der *B. licheniformis* F11 *locus* nahezu identische Sequenzen zu den *loci* von *B. licheniformis* DSM13 und ATTC14580. *B. licheniformis* 9945A verwendet aufgrund der geringeren Sequenzähnlichkeiten vermutlich einen anderen Pheromontyp.

Es gibt folglich zwei experimentelle Strategien um die Kompetenz in *B. licheniformis* DSM13 wiederherzustellen. Entweder wird versucht, ein aktives *comP* des gleichen Pheromontyps einzubringen oder es wird später in die Kompetenzbildungskaskade eingegriffen, wodurch das nicht funktionierende *quorum-sensing*-Modul umgangen wird.

Für die erste Strategie ergeben sich wiederum drei unterschiedliche Ansätze. i) Man kann versuchen, die Transposase zu entfernen und so das zerstörte *comP* wieder aktivieren. Nachteil ist, dass die Transposase dazu basengenau herausgeschnitten bzw. zum herausspringen bewegt werden muss, damit es zu keiner Leserahmenverschiebung in *comP* kommt.

ii) Die zweite Möglichkeit besteht darin, das gesamte *comQXPA*-Cluster durch ein funktionierendes Cluster aus einem anderen nah verwandten Stamm zu inserieren. Man könnte dazu auf *B. subtilis* oder *B. amyloliquefaciens* zurückgreifen. Ein idealer Kandidat wäre aber sicherlich eher ein kompetenter *B. licheniformis*-Stamm um stammspezifische Seiteneffekte zu minimieren.

iii) Die dritte Möglichkeit besteht darin, das defekte *comP* durch ein funktionierendes *comP* aus einem anderen *B. licheniformis*-Stamm gleichen Pheromontyps zu ersetzen. Die Arbeitsgruppe Meinhardt hat diesen Ansatz verfolgt und *comP* aus *B. licheniformis* F11 in *B. licheniformis* MW3 (Waschkau *et al.*, 2008) , ein restriktions-negatives Derivat von DSM13, inseriert (Hoffmann *et al.*, 2010). *Bacillus*-Stämme, deren *comP* nicht funktionsfähig ist, zeigen keine Polyglutamatbildung (Nagai *et al.*, 2000). Die Aktivität des in MW3 eingebrachten *comP* wurde über die wiederhergestellte Polyglutamatbildung bestätigt. Überraschenderweise konnte auf diesem Wege die Kompetenz nicht wiederhergestellt werden.

Die zweite experimentelle Strategie besteht in der Überexpression von ComK. Für verschiedene *Bacillus*-Wildtyp-Isolate (Duitman *et al.*, 2007) sowie für einen *B. cereus*-Stamm (Mirończuk *et al.*, 2008) konnte diese Methode erfolgreich angewandt werden. Es wurde eine um das 100-fache erhöhte Ausbildung der Kompetenz gezeigt (Duitman *et al.*, 2007).

Einer anderen Studie zufolge konnte eine Überexpression von ComK zwar für einige *Bacillus*-Wildtyp-Stämme erfolgreich durchgeführt werden, aber unter den gegebenen Versuchsbedingungen nicht in *B. licheniformis* reproduziert werden (Nijland, 2010). Vorteil dieser Methode ist, dass die Stämme nur kurzfristig induziert natürlich kompetent werden und im Anschluss das Überexpressionsplasmid wieder verlieren.

Für *B. licheniformis* MW3 konnte jedoch abweichend von der vorherigen Studie gezeigt werden, dass eine Überexpression von ComK zu einer Steigerung der natürlichen Kompetenz führt (Hoffmann *et al.*, 2010).

Da für die frühen (Hoffmann *et al.*, 2010), als auch für die späten Kompetenzgene in *B. licheniformis* DSM13 eine Aktivität nachgewiesen werden konnte, wurde gefolgert, dass die Verbindung beider Module gestört sein könnte. Neben ComK ist ComS ein entscheidender Faktor in der Kompetenzbildung (Dsouza *et al.*, 1995; Ogura *et al.*, 1999). Zwar konnte für DSM13 bioinformatisch ein kleiner ORF vorhergesagt werden, der Ähnlichkeiten zu *comS* aus *B. subtilis* und auch *B. amyloliquefaciens* aufweist, jedoch gibt es bisher keine Expressionsstudien. Die genomweite, bioinformatische Suche in DSM13 nach einem für ComS codierenden Gen, das möglicherweise an einem anderen chromosomalen Ort lokalisiert ist, brachte keine weiteren Treffer. Gestützt wird die Hypothese, dass *comS* für die mangelnde natürliche Kompetenz verantwortlich ist dadurch, dass es auch innerhalb der *B. licheniformis*-Gruppe Sequenzunterschiede in *comS* zwischen kompetenten und nicht kompetenten Vertretern gibt (Kapitel 3.3.3). Das *comS* aus *B. licheniformis* 9945A weist eine höhere Sequenzähnlichkeit zu den kompetenten *B. subtilis* 168 und *B. amyloliquefaciens* FZB42 auf, als zu den anderen *B. licheniformis* Stämmen.

Um über natürliche Kompetenz genetische Manipulationen in *B. licheniformis* DSM13 vornehmen zu können, bietet sich folglich die Überexpression von ComK an. Eine andere Möglichkeit besteht darin, die genetische Zugänglichkeit nicht über natürliche Kompetenz, sondern über Konjugation vorzunehmen (pers. Komm. M. Rachinger, Dissertation 2010).

Die Abgrenzung von *B. licheniformis* 9945A zu den anderen analysierten Vertretern der *B. licheniformis*-Gruppe in Bezug auf die fehlende Transposase in *comP*, die Sequenzunterschiede im stammspezifischen *comQX-locus* sowie *comS* spiegeln sich auch in der Phylogenie auf Ebene der 16S-rRNA wider (Abbildung 5).

Zwar ist die natürliche Kompetenz für *B. pumilus* SAFR-032 und *B. anthracis* Sterne nicht beschrieben, aber weil die späten Kompetenzgene konserviert sind, ist zu vermuten, dass die Überexpression von ComK ebenso wie in *B. cereus*

(Mirończuk *et al.*, 2008) ebenfalls zu kompetenten Stämmen führen kann. *B. pumilus* SAFR-032 könnte über einen bisher nicht experimentell untersuchten *quorum-sensing*-Mechanismus kompetent werden, da bis auf *comS* alle aus *B. subtilis* bekannten frühen Kompetenzgene konserviert sind. Sollte *B. anthracis* Sterne über die Fähigkeit der natürlichen Kompetenz verfügen, so ist eine Regulation, die nicht an das *quorum-sensing*-Modul aus *B. subtilis* angelehnt ist, sehr wahrscheinlich. Das *comQXPA*-Cluster ist in *B. anthracis* Sterne nicht vorhanden (Kapitel 3.3.1). Eine kürzlich veröffentlichte Analyse der späten Kompetenzgene in 20 komplett sequenzierten *Bacillus*-Stämmen zeigt, dass es in allen diesen Stämmen Homologe zu den in *B. subtilis* beschrieben späten Kompetenzgenen gibt (Kovacs *et al.*, 2009). Dies impliziert, dass die Regulation der Kompetenzgene in den Stämmen der Spezies *B. anthracis* über einen anderen Mechanismus erfolgt, was nicht überrascht, wenn man den völlig anderen Lebenszyklus der pathogenen Spezies im Vergleich zu den Saprophyten der *B. subtilis* - Gruppe in Betracht zieht.

6.2 Genomweite Identifikation von Orthologen

BiBaG wurde zur komparativen Genomanalyse entwickelt und erfüllt drei typische Analyseaufgaben: i) Gesamtgenomvergleiche von ausgewählten, verwandten Genomen zur Identifikation von orthologen Proteinen, ii) die Identifikation von Homologieclustern, iii) die Berechnung von *core* und *pan genomes* in Gruppen von Genomen. BiBaG wurde zur komparativen Analyse und für die Genomannotation in verschiedenen Veröffentlichungen von Genomen angewandt (Klee *et al.*, 2010; Köpke *et al.*, 2010; Liesegang *et al.*, 2010; Schmeisser *et al.*, 2009).

Test der Identifikation von Orthologen

Um die Genauigkeit von BiBaG abschätzen zu können, wurden mehrere Genome mit sich selbst verglichen (Kapitel 4.3). Die Annahme war, dass alle Proteine wechselseitig auf sich selbst *mappen* müssen. Es zeigte sich, dass maximal fünf von 1000 Proteinen nicht *gemappt* werden konnten. Die genaue Analyse der in allen betrachteten Organismen nicht *gemappten* Proteine zeigte, dass es sich um Paraloge handelt. Für *B. licheniformis* DSM13 ist eine Transposase beschrieben, die aus zwei ORFs besteht (Mahillon & Chandler, 1998). Es handelt sich um eine

IS3-Transposase, die in acht nicht codierenden Bereichen und in *comP* inseriert ist (Rey *et al.*, 2004). Die Transposasen sind mit den 16 nicht *gemappten* Proteinen identisch. Dieses Ergebnis unterstreicht die Robustheit von bidirektionalen besten *hits* als Methode zur Orthologenidentfikation. Hulsen und Altenhoff beschreiben in ihren Bewertungen verschiedener Methoden zur Orthologenbestimmung, dass die falsch-positiven Rate bei BBHs sehr gering ist (Hulsen *et al.*, 2006) und BBHs oftmals bessere Ergebnisse liefern, als komplexere Algorithmen (Altenhoff & Dessimoz, 2009).

Gesamtgenomvergleiche

Gesamtgenomvergleiche geben für die verwendeten Organismen einen Überblick über stammspezifische und orthologe Proteine. Die Verteilung dieser Proteine innerhalb des Genoms liefert interessante Einblicke in die Evolution und Genomdynamik verschiedener Spezies (Kolstø, 1997).

Der Vergleich von fünf *Bacillus*-Stämmen (Kapitel 4.4.1) zeigt für *B. licheniformis* DSM13 vier eindeutig identifizierbare stammspezifische Bereiche, die auch einen vom Durchschnitt abweichenden GC-Gehalt haben. Zur Verifikation dieser genomischen Inseln wurde eine IslandViewer-Analyse (Langille & Brinkman, 2009) durchgeführt. Das Ergebnis ist in Abbildung 45 dargestellt. Die berechneten *islands* decken sich mit den per BiBaG identifizierten Bereichen (Abbildung 27). Zusätzlich konnten mit dem IslandViewer sechs weitere *genomic islands* identifiziert werden, die ebenfalls stammspezifisch für *B. licheniformis* DSM13 sind und sich in der komparativen Gensonne (Abbildung 27) bestätigen lassen.

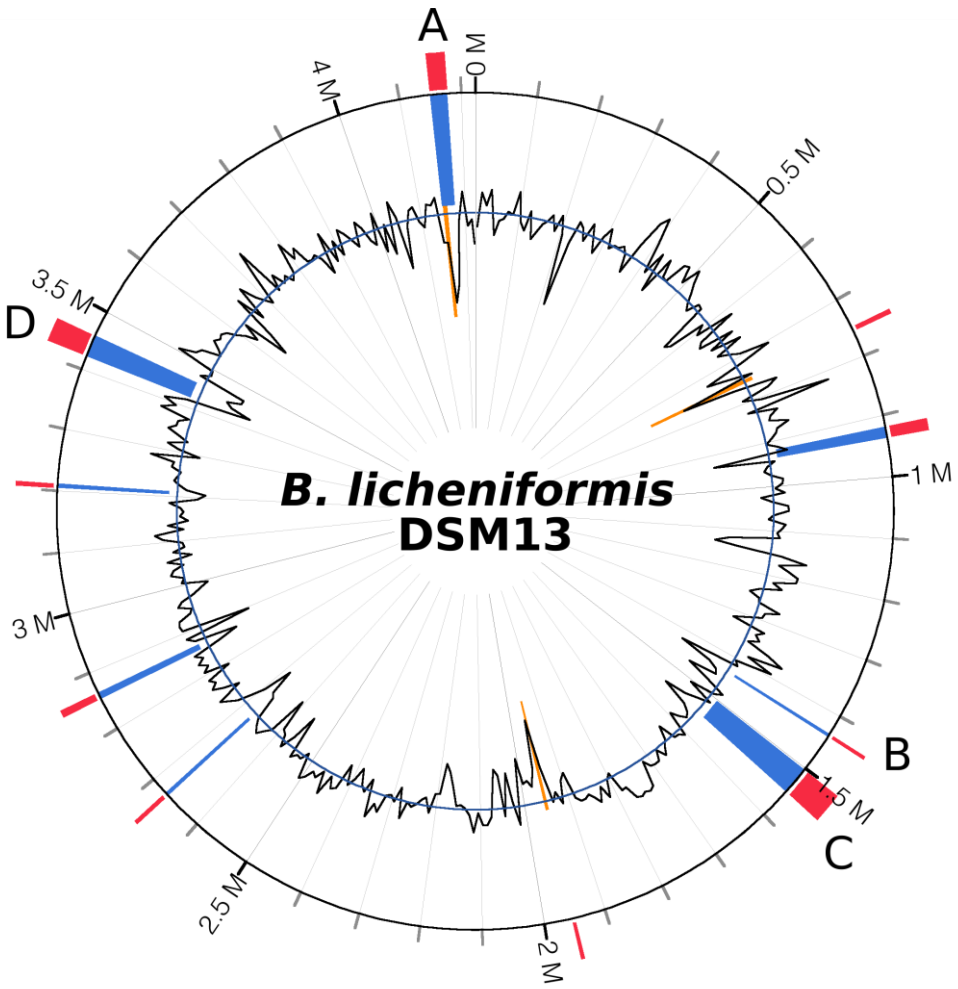


Abbildung 45: Grafische Darstellung der IslandViewer-Analyse
Blau markiert sind die Vorhersagen mit IslandPath-DIMOB, orange sind die Vorhersagen mit SIGI-HMM und rot umfasst die Bereiche, die mit wenigstens eienr dieser Mathode vorhersagt wurden. Mit A-D sind die Bereiche markiert, die auch mittels BiBaG identifiziert werden konnten.

Um verschiedene Organismen taxonomisch einzuordnen werden häufig 16S-rRNA basierte Vergleiche durchgeführt (Kong *et al.*, 2002; Porwal *et al.*, 2009; Woese *et al.*, 1975). Abbildung 46 zeigt einen 16S-rRNA Stammbaum der in Kapitel 4.4.1 verglichenen Stämme. Die phylogenetischen Ergebnisse decken sich mit der komparativen Analyse dieser Stämme. Demnach ist *B. licheniformis* DSM13 am dichtesten mit *B. subtilis* 168 verwandt, gefolgt von *B. amyloliquefaciens* FZB42, *B. pumilus* SAFR-032, *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125. Die Bestimmung von biologischen Verwandtschaftsbeziehungen auf Basis von Orthologenanzahlen liefert allerdings einen evolutionär detaillierten Blick auf die Organismen, da ein größerer Bereich des Genoms mit einbezogen wird.

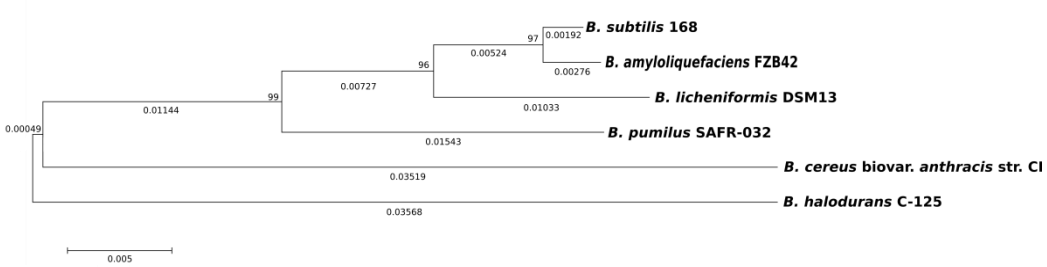


Abbildung 46: 16S-rRNA basierter Stammbaum der sechs komparativ analysierten *Bacilli*
Der horizontale Balken repräsentiert 0,005 Substitutionen pro Nukleotidposition

Abbildung 47 (I) zeigt einen Vergleich von fünf *E. coli*-Stämmen. *E. coli* ist eine sehr homogene Spezies mit erstaunlich hohen Ähnlichkeiten der Orthologen und klar abzugrenzenden *genomic islands* (Mira & Rodríguez-Valera, 2010). *Clostridia* sind hingegen eine sehr heterogene Spezies (Paredes *et al.*, 2005; Sathish & Swaminathan, 2009). Abbildung 47 (II) zeigt die Visualisierung eines Vergleichs von fünf *Clostridia*-Stämmen. Im Bereich um den *origin* ist eine Konservierung deutlich zu erkennen, wohingegen im Bereich um den *terminus of replication*, der sich zwischen den Basen 1,5 Mbp und 3,6 Mbp der Genomsonne befindet, weniger Orthologe vorhanden sind.

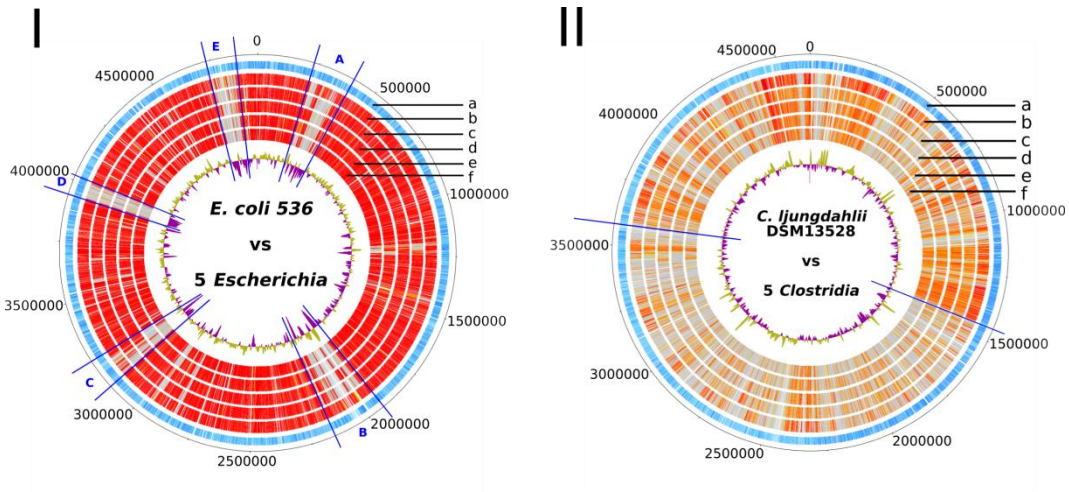


Abbildung 47: Visualisierung der NW-*similarities* für BiBaB-Analysen mit *E coli* 536und *C. ljungdhalii* DSM13528
(I) BiBaG-Analyse von *E. coli* 536 mit fünf anderen *Escherichia*-Stämmen: (a) *E. coli* 536, (b) *E. coli* CFT073, (c) *E. coli* SMS-3-5, (d) *E. coli* APEC O1, (e) *E. coli* O157:H7 EDL933, (f) *E. coli* str. K-12 substr. MG1655; Die *genomic islands* sind mit A-E markiert.
(II) BiBaG-Analyse von *C. ljungdahlii* DSM13528 mit fünf anderen Clostridia: (a) *C. ljungdahlii* DSM13528, (b) *C. kluyveri* DSM 555 , (c) *C. beijerinckii* NCIMB 8052, (d) *C. botulinum* F str. Langeland, (e) *C. acetobutylicum* ATCC 824, (f) *C. tetani* E88

Bacillus-Stämme haben im Vergleich zu *E. coli*-Stämmen auch klar abgegrenzte *genomic islands*, weisen aber nicht so große Ähnlichkeiten in den orthologen Proteinen auf. Die Unterschiede in der absoluten Ähnlichkeit reflektieren die Tatsa-

che, dass bei den Vergleichen auf unterschiedlichen taxonomischen Ebenen gearbeitet wurde. Die *E. coli* - Stämme sind Entitäten innerhalb einer Spezies, während bei den *Bacillus*-Stämmen verschiedene Spezies verglichen wurden. Insgesamt verteilen sich die *Bacillus*-Orthologen ebenso wie in *E. coli* über das gesamte Genom (Abbildung 27) im Gegensatz zu den *Clostridia* (Abbildung 47 (II)).

Homologiecluster

Die Evolution von Spezies basiert auf einem dynamischen Austausch von physiologischen Funktionen, die häufig geclustert in *genomic islands* (Hacker & Carniel, 2001) oder Operons (Bratlie *et al.*, 2010) vorliegen. Homologiecluster können somit einen noch genaueren evolutionären Einblick geben als die Bestimmung von Orthologen, da sie die chromosomale Co-Lokalisation der konservierten Proteine mit einbeziehen.

Abbildung 28 zeigt drei Bereiche in denen die jeweils längsten Homologiecluster der verglichenen Stämme auftreten. Der erste Bereich ist um den Replikationsursprung herum zu finden, der zweite Bereich umfasst den *downstream locus* des PBSX-Prophagen und der dritte Bereich befindet sich im Bereich des Endpunktes der Replikation. Eine hohe Konservierung in der Umgebung des Replikationsursprungs ist zu erwarten, da sich hier die zur Replikation notwendigen Proteine befinden. Die Gene *gyrA* und *gyrB*, die für ein Protein kodieren, das doppelsträngige DNA entspiralisiert, befinden sich in unmittelbarere Nähe zum Replikationsursprung. Aufgrund der speziesübergreifenden Konservierung beider Gene werden sie auch für phylogenetische Vergleiche herangezogen (Chun & Bae, 2000).

Der *downstream* Bereich des *Bacillus* spezifischen PBSX-Prophagen (Steensma *et al.*, 1978) umfasst im wesentlichen Flagellenproteine. Dies erklärt die langen Homologiecluster in *B. subtilis* 168, *B. amyloliquefaciens* FZB42 und *B. pumilus* SAFR-032, wohingegen die *cluster* in diesem Bereich für *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 weniger stark ausgeprägt sind.

Insgesamt fällt eine Abweichung der Clusterlängen von *B. subtilis* 168, *B. amyloliquefaciens* FZB42 und *B. pumilus* SAFR-032 zu *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 auf (Tabelle 14), die die phylogeneti-

sche Verwandtschaft innerhalb der *Subtilis*-Gruppe gegenüber den anderen beiden Vertretern der *Bacillus*-Spezies unterstreicht (Porwal *et al.*, 2009). Allgemein kann festgestellt werden, dass in nah verwandten Stämmen, die Anzahl der *cluster* geringer, dafür aber die Clusterlänge größer ist. Im Vergleich dazu, ist die Clusteranzahl in den entfernter verwandten *B. cereus* biovar. *anthracis* str. CI und *B. halodurans* C-125 größer, dafür aber die einzelnen Clusterlängen geringer.

Abbildung 48 verdeutlicht die bereits diskutierte Homogenität der *E. coli*-Gruppe im Vergleich zur Heterogenität innerhalb der *Clostridia*. Die *E. coli*- Homologiecluster verteilen sich über das gesamte Chromosom. Lücken treten hauptsächlich in den *genomic islands* auf. Außerdem umfassen benachbarte *cluster* überwiegend mehrere Proteine, so dass sich eine Blockstruktur ergibt (Abbildung 48 (I)). Die Homologiecluster in *C. ljungdahlii* sind im oberen Bereich der Gensonne dominant und treten im unteren Bereich häufig nur noch als einzelne konservierte Proteine auf.

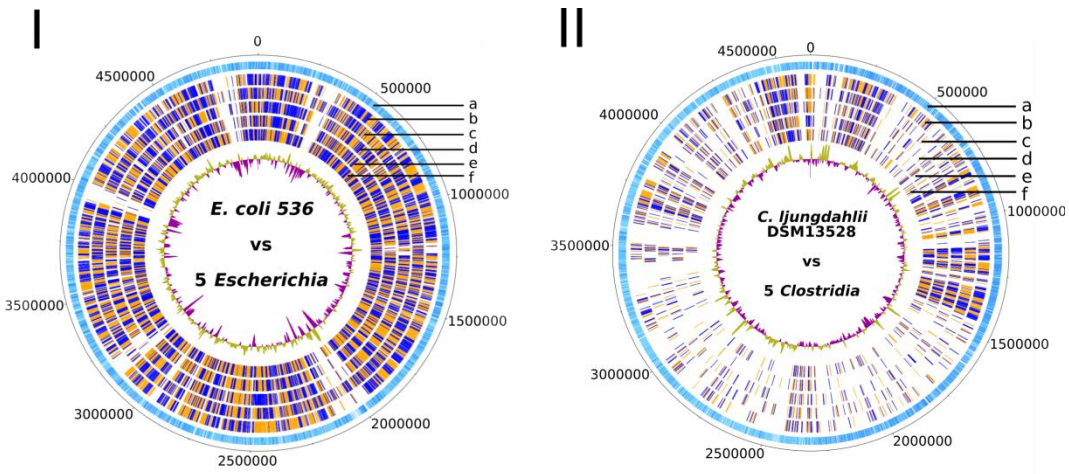


Abbildung 48: Homologiecluster in *E. coli* 536 und *C. ljungdahlii* DSM13528
(I) BiBaG-Analyse von *E. coli* 536 mit fünf anderen *Escherichia*-Stämmen: (a) *E. coli* 536, (b) *E. coli* CFT073, (c) *E. coli* SMS-3-5, (d) *E. coli* APEC O1, (e) *E. coli* O157:H7 EDL933, (f) *E. coli* str. K-12 substr. MG1655
(II) BiBaG-Analyse von *C. ljungdahlii* DSM13528 mit fünf anderen Clostridia: (a) *C. ljungdahlii* DSM13528, (b) *C. kluyveri* DSM 555 , (c) *C. beijerinckii* NCIMB 8052, (d) *C. botulinum* F str. Langeland, (e) *C. acetobutylicum* ATCC 824, (f) *C. tetani* E88

pan und *core genomes*

Die Definition von *pan* und *core genome* ist eng mit dem Spezieskonzept verknüpft (Medini *et al.*, 2005). Einige Stämme, wie z. B. *B. anthracis* lassen sich auf 16S-rRNA-Ebene nicht voneinander unterscheiden. Hier ist es notwendig,

stärker variierende, aber dennoch konservierte Gene zu identifizieren und im sogenannten *multilocus sequence typing* (MLST) einzusetzen (Klee *et al.*, 2006; Maiden, 2006). Üblicherweise werden hierfür *housekeeping genes* verwendet, die in allen betrachteten Organismen vorkommen. Daraus wird eine lange, zusammenhängende Sequenz gebildet, die den Organismus identifizieren soll. Basierend auf diesen *multi locus sequences* wird anschließend ein phylogenetischer Stammbaum berechnet. Es gibt keine Restriktion bezüglich der Art der Sequenz (DNA- oder Aminosäure-Sequenz), so dass *core* Proteine für MLST geeignet sind. Allerdings wurde kürzlich angeregt, für die taxonomische Unterscheidung innerhalb von Spezies auf *pan genome* Bereiche zurückzugreifen, da MLST dafür eine zu geringe Auflösung hat (Hall *et al.*, 2010). Die BiBaG-Vergleiche auf *B. cereus* biovar. *anthracis* str. CI haben gezeigt, dass die Methode geeignet ist, Gene für MLST in einer Gruppe von *B. cereus* / *anthracis*-Stämmen zu identifizieren (Klee *et al.*, 2010).

Die Triple-BiBaG-Analyse liefert erste Hinweise, welche Proteine von drei Organismen geteilt werden und welche stammspezifisch sind. Gleichzeitig liefert sie eine Datenbasis für eine Analyse mit der Fragestellung, wie stark *pan* und *core genome* von der Definition von Ähnlichkeit und damit von *cut off values* abhängen. Abbildung 29 zeigt drei Venn-Diagramme, die zu unterschiedlichen Ähnlichkeitsniveaus von orthologen Genen korrespondieren. Wird das *core genome* auf Basis von allen vorhanden bidirektionalen besten *hits* gebildet, decken sich die Zahlen bis auf geringe Abweichungen mit dem *core genome* basierend auf wenigstens 25 % NW-*similarity*. 25 % Sequenzähnlichkeit sind wenigstens erforderlich, damit die Chance besteht, dass das Protein die gleiche Struktur einnimmt (Sander & Schneider, 1991). Wird ein strengeres Kriterium gewählt, so dass das *core genome* aus denjenigen Proteinen gebildet wird, die wenigstens zu 90 % globaler Sequenzähnlichkeit übereinstimmen, nimmt die Anzahl der *core* Proteine deutlich ab. Dabei muss bedacht werden, dass *B. cereus* biovar. *anthracis* str. CI der *cereus* / *anthracis*-Gruppe und *B. licheniformis* DSM13 sowie *B. amyloliquefacienz* FZB42 der *Subtilis*-Gruppe angehören.

Auffällig ist, dass sich die Anzahlen der Überlappungsbereiche im Venn-Diagramm unterscheiden, abhängig davon aus welcher Sicht der Vergleich durch-

geführt wurde. Diese Diskrepanzen ergeben sich zum einen aus den unterschiedlichen Genomgrößen und damit verbunden aus den unterschiedlichen Proteinanzahlen und zum anderen aus den unterschiedlichen Anzahlen an Paralogen und sich wiederholenden, gleichen Proteindomänentreffern.

Betrachtet man *pan* und *core genomes* stößt man auf die Frage, ob eine Spezies ein offenes oder geschlossenes *pan genome* hat. Ein geschlossenes *pan genome* liegt vor, wenn mit der Hinzunahme eines weiteren Organismus einer Spezies in die komparative Analyse keine weiteren Gene / Proteine hinzugefügt werden, die im bisher betrachteten *pool* noch nicht vorhanden sind (Medini *et al.*, 2005). Im Gegensatz dazu ist ein offenes *pan genome* dadurch definiert, dass mit jedem neu sequenzierten Organismus auch neue Gene / Proteine identifiziert werden können.

Mit der statistischen Analyse von *core* und *pan genomes* (Kapitel 4.4.1) stellt BiBaG die Möglichkeit bereit, für einen Referenzorganismus und weitere Vergleichsorganismen abzuschätzen, ob es sich um ein offenes oder geschlossenes *pan genome* handelt. Für die Analyse von *B. licheniformis* DSM13 und zehn weiteren *Bacillus*-Stämmen zeigte sich, dass das *core genome* ein Plateau erreicht, aber das *pan genome* mit jeder Hinzunahme eines weiteren Organismus stark ansteigt. Die Vermutung liegt nahe, dass die Gattung *Bacillus* ein offenes *pan genome* hat. Da *Bacilli* ubiquitär vorkommen und isoliert werden können, ist diese Beobachtung konsistent damit, dass sie sich. je nach Standort an die dort gegebenen Bedingungen anpassen können. Außerdem ist die Gruppe der *Bacilli* insgesamt sehr heterogen, was den *lifestyle* betrifft (Porwal *et al.*, 2009). Dies lässt auf ein *kleines core genome* schließen, das notwendig ist, um ein *Bacillus* zu sein und auf ein großes *pan genome*, das alle Proteine enthält, die für die jeweilige Nischenadaption akquiriert wurden. Weitere Spezies mit offenen *pan genomes* sind *E. coli* (Konstantinidis & Tiedje, 2005) oder *Prochlorococcus* (Kettler *et al.*, 2007). Absolut gesehen ist zu bedenken, dass es natürlich nur eine endliche Anzahl von Genen bzw. Genomen geben kann und die derzeit verfügbaren Sequenzen nur einen geringen Bruchteil der natürlichen Diversität darstellen (Daniel, 2005). Das heißt, bei hinreichend vielen sequenzierten Genomen werden alle Spezies eine Obergrenze ihres *pan genomes* erreichen müssen.

Betrachtet man hingegen Spezies, die auf einen isolierten Lebensraum beschränkt sind und denen damit die Möglichkeit fehlt, ihren Genpool zu erweitern, findet man Beispiele für geschlossene *pan genomes*. *B. anthracis* kann mit vier sequenzierten Stämmen komplett beschrieben werden (Medini *et al.*, 2005) und auch der intrazelluläre Symbiont *Buchnera aphidicola* hat ein geschlossenes *pan genome* (Medini *et al.*, 2005*;* Tamas *et al.*, 2002*).*

Für die Berechnung der Größe von *pan* und *core genomes* gibt es einige kürzlich erschienene Analysen, die auf eine Spezies fokussiert sind (Lukjancenko (Davenport *et al.*, 2010; Fischer *et al.*, 2010; Lukjancenko *et al.*, 2010) sowie allgemein formuliert sind (Snipen *et al.*, 2009). Es kann gezeigt werden, dass das speziesweite *pan genome* deutlich größer ist, als ein Genom eines einzelnen Vertreters der Spezies. Diese Aussage stimmt mit den hier präsentierten Ergebnissen der komparativen Analysen überein.

Annotationsübertragung mit BiBaG

Mit der Bestimmung von Orthologen ergibt sich auch die Möglichkeit, Annotationsübertragungen vorzunehmen. Nachdem durch die *next-generation-sequencing*-Technologien die Genomsequenzierung enorm beschleunigt wurde, stellt die Annotation den Engpass der sequenzbasierten Analysen dar. Automatische Annotationen basieren häufig auf BLAST-Analysen gegen Proteindatenbanken, wie z. B. Swiss-Prot (Bairoch *et al.*, 2004) oder Trembl (Boeckmann *et al.*, 2003). Nachteil ist, dass es zunächst keine Möglichkeit gibt, diese automatischen Annotationen auf bestimmte Organismengruppen zu beschränken. Dies birgt die Gefahr von Fehlerfortpflanzungen. Ein Protein, das einmal falsch annotiert in einer öffentlichen Datenbank gespeichert wurde und Orthologe in neu annotierten Genomen hat, wird automatisch mit der falschen Annotation versehen (Bork, 2000).

BiBaG erlaubt die Annotationsübertragung innerhalb der gewählten Vergleichsorganismen auf den Referenzorganismus. Über die NW-*similarity*, kann frei gewählt werden, ab welchen globalen Alignmentlängen Annotationen als sicher gelten. Durch die benutzerdefinierte Auswahl der Vergleichsorganismen kann sichergestellt werden, dass die intrinsische Fehlerfortpflanzung minimiert wird.

Mit Hilfe von BiBaG wurden die überarbeiteten Annotationen von *B. subtilis* 168 (Barbe *et al.*, 2009) auf *B. licheniformis* DSM13 und *B. licheniformis* 9945A übertragen. Eine Annotation wurde übernommen, wenn ein bidirektionaler bester *hit* vorliegt und die NW-*similarity* wenigstens 70 % beträgt. Für beide Organismen konnten so mehr als die Hälfte der Proteine annotiert werden.

Eine automatische Annotation ist zwar niemals so gut, wie eine manuelle Annotation von einem Experten, aber eine gesicherte automatische Annotation erleichtert den Annotatoren die Arbeit und spart Zeit.

Für *B. subtilis* 168 sind zahlreiche Bestrebungen bekannt, die Annotationen zu aktualisieren (Barbe *et al.*, 2009) bzw. *community* basiert , kooperativ, laufend zu aktualisieren (Flórez *et al.*, 2009) und darüber hinausgehend navigierbare Stoffwechselwege bereit zu stellen (Lammers *et al.*, 2010).

Abgrenzung zu anderen komparativen Ansätzen

BiBaG ist ein *user*-orientiertes *tool* zur komparativen Analyse von mikrobiellen Genomen. Es erlaubt den Vergleich von einem Organismus mit beliebig vielen anderen Organismen. Über die Einstellungsparameter kann die Programmausgabe sowohl in Bezug auf *cut-offs* als auch farbliche Gestaltung individuell angepasst werden. BiBaG ist für *user* selbst zu installieren und ist auf einer Reihe von unixartigen Systemen getetstet worden. Der Benutzer ist folglich nicht darauf angewiesen, seine Daten auf einen fremden Web-Server zu laden, sondern kann die komparativen Analysen lokal durchführen.

Damit unterscheidet es sich zum Beispiel von den webbasierten *tools* EDGAR (Blom *et al.*, 2009), das ebenfalls Gesamtgenomvergleiche basierend auf bidirektionalen besten *hits* bereitstellt sowie PSAT (Fong *et al.*, 2008) und „*The Microbe Browser*“ (Gattiker *et al.*, 2009), die die Genumgebungen von Homologen analysieren. Auch MOSAIC ist internetbasiert und auf intra-Spezies-Vergleiche spezialisiert (Gattiker *et al.*, 2009). Außerdem werden in der komparativen Genomanalyse weitere Programme mit unterschiedlichen Schwerpunkten eingesetzt. ACT (Carver *et al.*, 2008) ist ebenso wie MUMmer (Kurtz *et al.*, 2004) auf den paarweisen Vergleich zweier Genome fokussiert, während MAUVE (Darling *et al.*,

2004) lineare Gesamtgenom-*alignments* von zwei und mehr Organismen produziert.

Bedeutung der komparativen Analysen für die Stammoptimierung

Im Laufe dieser Arbeit konnten mit komparativen Methoden sowohl stammspezifische Bereiche in *B. licheniformis* DSM13 bestimmt werden, als auch Insertions-/ Deletionstargets identifiziert werden.

In den stammspezifischen, hypothetischen Proteinen steckt ein großes Potential, so dass diese Proteine ebenfalls als zu untersuchende Deletionstargets aufgefasst werden können (Galperin & Koonin, 2004). Bioinformatische Ansätze, die auf regulären Ausdrücken, statt auf BLAST-Vergleichen basieren (Cestari *et al.*, 2006) oder experimentelle Proteinanalysen (z. B. Expressionsstudien, Strukturaufklärungen) (Kuznetsova *et al.*, 2005) bieten die Möglichkeit diese Proteine näher zu charakterisieren.

Als mögliches Insertionstarget konnte das Urease-Operon aus *B. licheniformis* 9945A identifiziert werden. Mit Hilfe dieses *clusters* könnte *B. licheniformis* DSM13 sein Substratspektrum erweitern und auch auf Harnstoff wachsen.

Eine Strategie zur rationalen Stammoptimierung, die auf starker Reduktion des Genoms basiert, ist der Minimalgenom-Ansatz. Für *B. subtilis* 168 führte die Reduzierung des Genoms zu einer erhöhten Proteinproduktivität (Morimoto *et al.*, 2008). Auch in *E. coli*-Stämmen wurden Genomreduzierungen vorgenommen (Mizoguchi *et al.*, 2007). Damit einher geht das *metabolic engineering*, bei dem durch gezielte Insertion von einzelnen Genen oder *pathways* die Optimierung eines Stammes im Sinne von Produktivität, Erweiterung des Substratspektrums oder die verringerte Produktion von Nebenprodukten erreicht werden kann (Nielsen, 2001).

Ein Beispiel für erfolgreiches *metabolic engineering* ist das größte industrielle Produkt, im Sinne von Produktionsvolumen und Verkaufsraten: Bioethanol (Otero & Nielsen, 2010). Die bekannteste Produktionsplattform basiert auf *Saccharomyces cerevisiae*. Die Verwendung von metabolischen Modellen führte

zum einen zu einer erhöhten Produktion von Ethanol bei gleichzeitiger Reduzierung der Produktion des Nebenprodukts Glyzerin um 40 % durch eine Gendeletion (Nissen *et al.*, 2000) und zum anderen zur erhöhten Produktion von Bioethanol durch Insertion eines Gens (Bro *et al.*, 2006). Ein weiteres Beispiel für erfolgreiches *metabolic engineering* stellt die Produktion von 1,3 Propandiol dar (Nakamura & Whited, 2003). Die Erweiterung des natürlichen glycerinbasierten Stoffwechselweges zu einem effizienteren Prozess beinhaltet zahlreiche Optimierungsschritte: i) Wechsel von eines anaeroben zu einem aeroben Prozess, ii) Verbesserung der Substrataufnahme sowie iii) Design und Implementierung einer optimierten Lösung zur Balance des Energiehaushalts unter Berücksichtigung des bakteriellen Wachstums und Produktbildung.

Die Genomsequenzierung und funktionale Genomik bilden die Grundlage für die Erstellung von metabolischen Modellen, die im nächsten Schritt auch *in silico* Simulationen erlauben (Durot *et al.*, 2009). Für zahlreiche biotechnologisch und medizinisch relevante Organismen wurden metabolische Modelle erstellt (Milne *et al.*, 2009). Auch für *B. subtilis* sind *metabolic engineering*-Ansätze bekannt. Es gibt bereits Modelle des Metabolismus (Henry *et al.*, 2009; Oh *et al.*, 2007) sowie des zentralen Stoffwechsels von *B. subtilis* (Goelzer *et al.*, 2008). Mit Hilfe einer vergleichenden Transkriptomanalyse konnte die Riboflavinproduktion optimiert werden (Shi *et al.*, 2009).

Obwohl die systembiologischen Ansätze vielversprechende Methoden zur Bioprozessmodellierung sind (Park *et al.*, 2008), so sind die Modelle keinesfalls vollständig. Die Integration regulatorischer Mechanismen in die metabolischen Modelle ist eine der zukünftigen Herausforderungen (Liu *et al.*, 2010).

6.3 Datenbank

Für das Projekt wird ein Datenbanksystem benötigt. Zahlreiche Datenbanksysteme stehen zur Verfügung. Das kommerzielle Datenbanksystem Phylosopher wird gezielt in der Biotechnologie eingesetzt und erfüllt zahlreiche Aufgaben, die in der Stammentwicklung eine Rolle spielen (www.genedata.com). Darüber hinaus gibt es einige *open source* LIMS, die alle im Labor verwendeten und anfallenden Daten speichern und verwalten (Prilusky *et al.*, 2005; Stocker *et al.*, 2009). Diese

tools sind mit Methoden ausgestattet, die für die hier vorhandenen experimentellen und anderen Daten zu umfangreich und unspezifisch sind. Außerdem gibt es Datenbanksystem, die für spezifische Aufgaben konzipiert sind, z. B. die Bereitstellung von Operons (Mao *et al.*, 2009; Pertea *et al.*, 2009) oder die Speicherung und Bereitstellung von Expressionsexperimenten (Kapushesky *et al.*, 2010; Parkinson *et al.*, 2009). Diese Systeme sind für die im Projekt zu verwaltenden Daten zu spezifisch.

Deswegen wurde eine SQL-Datenbank entwickelt, die für die Speicherung und Analyse der im Projekt anfallenden Daten ausgelegt ist. Insbesondere werden *B. licheniformis* DSM13 zentrierte genomische Informationen, komparative Vergleichsdaten zu anderen *Bacillus*-Stämmen und experimentelle *microarray*-Daten gespeichert. Zusätzlich wurden Methoden entwickelt, die einfache Datenbankabfrage, z. B. zur Vorehrsage von Operons oder zur Erstellung von GFF-Dateien der Operons bzw. regulierten Genen, erlauben. Dadurch wird eine genomweite Visualisierung der Daten ermöglicht (Abbildungen 42-44).

Die Implementierung der in dieser Arbeit entwickelten Datenbank legt es nahe, die sequenzbasierte Operonvorhersage durch die vollständige Integration der übrigen Informationen zu optimieren. Es gibt zahlreiche Ansätze zur Vorhersage von Operons in Prokaryoten (Brouwer *et al.*, 2008). Sie basieren im Wesentlichen auf fünf Kriterien: i) intergenische Distanz, ii) konservierte Gencluster, iii) funktionale Beziehungen, iv) Sequenzelemente und v) experimentellen Bestätigungen. Mit der Datenbank werden die Kriterien i), ii) und v) bereitgestellt. Funktionale Beziehungen lassen sich z. B. über COG (Tatusov *et al.*, 2003), metabolische Stoffwechselwege (Zheng *et al.*, 2002) oder Gene Ontologies (Consortium, 2010) hinzufügen. Mit Sequenzelementen sind transkriptionale Terminatoren, Promotorsequenzen oder Transkriptionsfaktorbindestellen gemeint, für die es ebenfalls bioinformatische Vorhersagemethoden gibt (Ermolaeva *et al.*, 2001; Kingsford *et al.*, 2007; Posch *et al.*, 2010; Werner, 2002). Durch die vollständige Integration dieser Daten wird eine solide Operonvorhersage für *B. licheniformis* DSM13 ermöglicht.

Die Verwendung von experimentellen Daten erhöht die Validität der Vorhersagen. Allerdings beschreibt Brouwer bei einem Vergleich verschiedener Methoden (Brouwer *et al.*, 2008), dass die Operonvorhersage von Moreno Hagelsieb und

Collado-Vides, die nur auf intergenischen Distanzen basiert (Moreno-Hagelsieb & Collado-Vides, 2002), bessere Ergebnisse liefert, als neuere, komplexere Algorithmen.

Für *B. subtilis* gibt es zahlreiche experimentell verifizierte Operons (Sierro *et al.*, 2008), so dass sowohl geeignete Trainingsdaten für die verschiedenen Operonvorhersagemethoden zur Verfügung stehen, als auch Aussagen zur Genauigkeit der Algorithmen getroffen werden können. Außerdem gibt es einige webbasierte-Datenbanken, die für zahlreiche Organismen Operonvorhersagen bereitstellen (Mao *et al.*, 2009; Pertea *et al.*, 2009).

Anstelle von *microarray*-Daten können auch Transkriptomsequenzierungen (Wang *et al.*, 2009) verwendet werden um Operons experimentell zu verifizieren. Mit Hilfe von Transkriptomanalysen ist die exakte Bestimmung von Transkriptionsstarts möglich (Passalacqua *et al.*, 2009). Diese RNA-seq benannte Technologie kann sowohl kleine RNAs (Fröhlich & Vogel, 2009), als auch Operons identifizieren (Sorek & Cossart, 2010). Es konnte gezeigt werden, dass das Transkriptom dynamischer ist, als bisher vermutet, insbesondere wurden komplett neue biologische Features wie *anti sense transcripte* und CRISPR-Elemente identifiziert, die nicht in das klassische Operonkonzept passen. Ein Nachteil der Verwendung von Transkripom-Daten zur Operonbestimmung ist, dass je nach gewählter Versuchsbedingung nicht alle Operons aktiv sind. Transkriptomanalysen wurden bereits für *B. anthracis* (Martin *et al.*, 2010; Passalacqua *et al.*, 2009) und *B. subtilis* (Rasmussen *et al.*, 2009) durchgeführt. Für *B. licheniformis* gibt es zwar *microarray*-Daten (Hornbaek *et al.*, 2004; Nielsen *et al.*, 2010), aber noch keine Transkriptomanalysen.

Grenzen der Einsetzbarkeit der Datenbank sind im derzeitigen Status der Entwicklung eine *multiuser*-Anwendung, da eine passwortgeschütze Nutzerverwaltung noch nicht implementiert wurde. Außerdem müssen einige Zugriffs- und Speichermethoden für die Verwendung von mehr als einem Organismus angepasst werden.

6.4 Ausblick

Die drei Schwerpunkte dieser Arbeit bieten zahlreiche Ansätze zu weiteren Forschungsaufgaben.

Stammoptimierung von B. *licheniformis* DSM13

Mit der Sequenzierung von *B. licheniformis* 9945A steht die Genomsequenz eines Stammes zur Verfügung (pers. Komm. M. Rachinger, Dissertation 2010), der sehr wahrscheinlich natürlich kompetent ist. Bisher ist allerdings nur für die auxotrophen Mutanten M18 und M28 des Stammes eine natürliche Kompetenz beschrieben (Zitat). Eine komparative BiBaG-Analyse könnte weitere Unterschiede zwischen nicht kompetenten und natürlich kompetenten *B. licheniformis*-Stämmen liefern. Entscheidender Vorteil solch einer Analyse gegenüber der hier angewandten Analyse innerhalb der Gattung *Bacillus* (Kapitel 3) ist, dass sie *B. licheniformis* – spezifisch durchgeführt werden kann. Mit BiBaG kann sie genomweit durchgeführt werden und ist nicht auf die Auswahl spezieller Proteine beschränkt.

Im Hinblick auf die vorhergesagten Insertions- /und Deletionstargets zur Stammoptimierung von *B. licheniformis* DSM13 ist eine Basis für weitere experimentelle Arbeiten geschaffen, deren Ergebnisse nach ihrer Durchführung mit den bioinformatischen Vorhersagen rückgekoppelt werden sollten, um die Vorhersagen anzupassen und ggf. zu optimieren. Darüber hinaus bieten die 396 identifizierten stammspezifischen, zumeist hypothetischen Proteine (Kapitel 4.4.1) eine Basis mit bioinformatischen bzw. experimentellen Methoden (*single knock outs*, Proteincharakterisierungen u. ä.) zu untersuchen, was *B. licheniformis* als Spezies im Vergleich zu *B. subtilis* ausmacht.

Ein entscheidender weiterer Schritt ist die Erstellung einer *in silico* metabolischen Karte, z. B. mit Hilfe von PathwayTools (Karp *et al.*, 2010). Da die Annotation von DSM13 inzwischen mehr als sechs Jahre alt ist und auch eine neuere *B. subtilis* 168 Referenzannotation zur Verfügung steht, ist es sinnvoll die in dieser Arbeit automatisch *gemappte* Annotation noch einmal manuell zu überprüfen und dann als Basis für PathwayTools zu verwenden. Ein *in silico* metabolisches Netzwerk bietet durch die Integration von experimentellen Transkriptom, Proteom

und Metabolomdaten die notwendige Grundlage für Stoffwechselsimulationen und damit *metabolic engineering*-Ansätze. Engpässe im gewünschten Sekretionsablauf können so aufgedeckt und durch gezielte genetische Manipulation umgangen werden.

Weiterentwicklung von BiBaG

BiBaG bietet einige Entwicklungsmöglichkeiten, um ein noch leistungsfähigeres *tool* in der komparativen Genomik zu werden. Bisher werden nur die Positionen der Gene des Ausgangsorganismus für die Orthologenbestimmung und Homologiecluster-Analyse in die Vergleiche mit einbezogen. Positionsangaben, die relativ zur chromosomalen Lokalisation der Orthologen bzw. *cluster* in den Vergleichsstämmen sind, würden noch detaillierte Einblicke in die Genomdynamiken der Organismen liefern.

Darüber hinaus hat sich gezeigt, dass BiBaG-Analysen genauere phylogenetische Einordnungen erlauben, als 16S-rRNA Analysen. Die Bereitstellung der z. B. 20 ähnlichsten Proteine in allen analysierten Organismen, würde die Generierung automatisierter MLST-Stammbäume erlauben.

Ohne die Vorteile der lokalen Installation zu verlieren, wäre ein Webinterface hilfreich, das vorberechnete BiBaG-Analysen zur Verfügung stellt. Für Standardvergleiche oder erste Einblicke in bestimmte Organismengruppen, würde dies Rechenzeit einsparen.

Außerdem ist eine Anbindung von BiBaG an die entwickelte Datenbank unerlässlich, um allen Projektbeteiligten ein zentrales Softwaresystem zur Verfügung zu stellen, das komparative, experimentelle und sequenzbasierte *B. licheniformis*-Daten bereitstellt.

Weiterentwicklung der Datenbank

Basierend auf der bereits erwähnten Integration von BiBaG-Analysen in die Datenbank ist eine umfassendere Implementierung der Operonvorhersage möglich, die neben den sequenzbasierten Daten, auch experimentelle Daten, sowie die Homologiecluster mit einbezieht.

In einer zukünftigen Version der DB können auch regulatorische Elemente Berücksichtigung finden, die nicht in das klassische Operonkonzept passen, wie z.B. *small* RNAs oder *antisense* Transkripte.

Außerdem muss die Benutzerverwaltung vollständig implementiert werden um Zugriffsbeschränkungen zu realisieren und damit die Datensicherheit zu gewährleisten.

Sobald die Transposonmutagenesedaten vorliegen sollen sie ebenfalls integriert werden.

7 Zusammenfassung

- Die Kompetenzsystemanalyse zeigte, dass in allen untersuchten *Bacillus*-Stämmen im Vergleich zum bekanntermaßen natürlich kompetenten Modellorganismus *B. subtilis* 168 wenigsten die späten Kompetenzgene konserviert sind.

- Da sowohl die frühen als auch die späten Kompetenzgene in *B. licheniformis* DSM13 experimentell nachweisbar aktiv sind bzw. durch Austausch von *comP* aktiviert werden konnten, besteht die Vermutung, dass *comS*, als Verknüpfung zwischen beiden Regulationsbereichen ursächlich für die mangelnde natürliche Kompetenz ist.

- Das *software tool* BiBaG wurde entwickelt und erfüllt drei Aufgaben im Bereich der komparativen Genomanalyse: i) Gesamtgenomvergleiche von ausgewählten, verwandten Genomen zur Identifikation von orthologen Proteinen, ii) die Identifikation von Homologieclustern, iii) die Berechnung von *core* und *pan genomes* in Gruppen von Genomen.

- Mit Hilfe von BiBaG wurde i) die Genomdynamik zwischen *Bacillus*-/ *E. coli*-/ und *Clostridia*-Stämmen untersucht, ii) eine Annotationsübertragung von *B. subtilis* 168 auf zwei *B. licheniformis*-Stämme vorgenommen, iii) das *pan* und *core genome* von *B. licheniformis* DSM13 näher charakterisiert, iv) *genomic islands* in *B. licheniformis* DSM13 bestimmt, v) eine Insertionstarget-/ sowie eine Deletionstargetliste zur Stammoptimierung von *B. licheniformis* DSM13 erstellt und vi) ein Beitrag zu vier veröffentlichen Genompublikationen geleistet.

- Es wurde eine integrative Datenbank zur Speicherung experimenteller und sequenzbasierter *B. licheniformis* DSM13-Daten entwickelt. Unter Verwendung der gespeicherten Daten ist eine einfache, sequenzbasierte Operonvorhersage sowie die Visualisierung experimenteller Daten möglich.

8 Literaturverzeichnis

Altenhoff, A. & Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**, e1000262.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment tool. *Journal of Molecular Biology* **215**, 403-410.

Ansaldi, M., Marolt, D., Stebe, T., Mandic-Mulec, I. & Dubnau, D. (2002). Specific activation of the *Bacillus* quorum-sensing systems by isoprenylated pheromone variants. *Molecular Microbiology* **44**, 1561-1573.

Ansaldi, M. & Dubnau, D. (2004). Diversifying selection at the *Bacillus* quorum-sensing locus and determinants of modification specificity during synthesis of the ComX pheromone. *Journal of Bacteriology* **186**, 15-21.

Ara, K., Ozaki, K., Nakamura, K., Yamane, K., Sekiguchi, J. & Ogasawara, N. (2007). *Bacillus* minimum genome factory: effective utilization of microbial genome information. *Biotechnology and Applied Biochemistry* **46**, 169-178.

Avery, O. T., Macleod, C. M. & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of experimental medicine* **79**, 137-158.

Bai, U., Mandicmulec, I. & Smith, I. (1993). SinI modulates the activity of SinR, a developmental switch protein of *Bacillus subtilis*, by protein protein interaction. *Genes & Development* **7**, 139-148.

Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* **5**, 39-55.

Barbe, V., Cruveiller, S., Kunst, F. & other authors (2009). From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* **155**, 1758-1775.

Bassler, B. L. & Losick, R. (2006). Bacterially speaking. *Cell* **125**, 237-246.

Binnewies, T. T., Motro, Y., Hallin, P. F. & other authors (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & Integrative Genomics* **6**, 165-185.

Binnie, C., Lampe, M. & Losick, R. (1986). Gene encoding the sigma 37 species of RNA polymerase sigma factor from *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 5943-5947.

Blom, J., Albaum, S., Doppmeier, D., Pühler, A., Vorhölter, F., Zakrzewski, M. & Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10**, 154.

Boeckmann, B., Bairoch, A., Apweiler, R. & other authors (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. P. (1998). Predicting function: From genes to genomes and back. *Journal of Molecular Biology* **283**, 707-725.

Bork, P. (2000). Powers and pitfalls in sequence analysis: The 70% hurdle. *Genome Research* **10**, 398-400.

Bratlie, M., Johansen, J. & Drabløs, F. (2010). Relationship between operon preference and functional properties of persistent genes in bacterial genomes. *BMC Genomics* **11**, 71.

Bro, C., Regenberg, B., Förster, J. & Nielsen, J. (2006). In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metab Eng* **8**, 102-111.

Brouwer, R., Kuipers, O. & van Hijum, S. (2008). The relative value of operon predictions. *Brief Bioinform* **9**, 367-375.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B., Parkhill, J. & Rajandream, M. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672-2676.

Carver, T., Thomson, N., Bleasby, A., Berriman, M. & Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25**, 119-120.

Cestari, I., Haver, N., Barbosa-Silva, A. & Ramirez, M. (2006). PROTOGIM: a novel tool to search motifs and domains in hypothetical proteins of protozoan genomes. *Parasitol Res* **98**, 375-377.

Chai, Y., Norman, T., Kolter, R. & Losick, R. (2010). An epigenetic switch governing daughter cell separation in *Bacillus subtilis*. *Genes Dev* **24**, 754-765.

Chen, P. P.-S. (1976). The Entity-Relationship Model: Toward a Unified View of Data, pp. 9-36. ACM Transactions on Database Systems.

Chen, X. H., Koumoutsi, A., Scholz, R. & other authors (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nature Biotechnology* **25**, 1007-1014.

Chun, J. & Bae, K. (2000). Phylogenetic analysis of *Bacillus subtilis* and related taxa based on partial gyrA gene sequences. *Antonie Van Leeuwenhoek* **78**, 123-127.

Chung, Y. S. & Dubnau, D. (1995). ComC is required for the processing and translocation of ComGC, a pilin-like competence protein of *Bacillus subtilis*. *Molecular Microbiology* **15**, 543-551.

Chung, Y. S. & Dubnau, D. (1998). All seven *comG* open reading frames are required for DNA binding during transformation of competent *Bacillus subtilis*. *Journal of Bacteriology* **180**, 41-45.

Consortium, G. O. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* **38**, D331-335.

Daniel, R. (2005). The metagenomics of soil. *Nat Rev Microbiol* **3**, 470-478.

Darling, A., Mau, B., Blattner, F. & Perna, N. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403.

Davenport, C., Ussery, D. & Tümmler, B. (2010). Comparative genomics of green sulfur bacteria. *Photosynth Res* **104**, 137-152.

Droffner, M. L. & Yamamoto, N. (1985). Isolation of thermophilic mutants of *Bacillus subtilis* and *Bacillus pumilus* and transformation of the thermophilic trait to mesophilic strains. *Journal of General Microbiology* **131**, 2789-2794.

Dsouza, C., Nakano, M. M., Frisby, D. L. & Zuber, P. (1995). Translation of the open reading frame encoded by *coms*, a gene of the srf operon, is necessary for the development of genetic competence, but not surfactin biosynthesis, in *Bacillus subtilis*. *Journal of Bacteriology* **177**, 4144-4148.

Dubnau, D. (1991a). Genetic competence in *Bacillus subtilis*. *Microbiological Reviews* **55**, 395-424.

Dubnau, D. (1991b). The regulation of genetic competence in *Bacillus subtilis*. *Molecular Microbiology* **5**, 11-18.

Duitman, E. H., Wyczawski, D., Boven, L. G., Venema, G., Kuipers, O. P. & Hamoen, L. W. (2007). Novel methods for genetic transformation of natural *Bacillus subtilis* isolates used to study the regulation of the mycosubtilin and surfactin synthetases. *Applied and Environmental Microbiology* **73**, 3490-3496.

Durot, M., Bourguignon, P. & Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* **33**, 164-190.

Ermolaeva, M., White, O. & Salzberg, S. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**, 1216-1221.

Fischer, W., Windhager, L., Rohrer, S., Zeiller, M., Karnholz, A., Hoffmann, R., Zimmer, R. & Haas, R. (2010). Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Res*.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* **19**, 99-&.

Fleischmann, R. D., Adams, M. D., White, O. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Flórez, L., Roppel, S., Schmeisky, A., Lammers, C. & Stülke, J. (2009). A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki SubtiWiki. *Database (Oxford)* **2009**, bap012.

Fong, C., Rohmer, L., Radey, M., Wasnick, M. & Brittnacher, M. (2008). PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* **9**, 170.

Fraser, C. M., Eisen, J., Fleischmann, R. D., Ketchum, K. A. & Peterson, S. (2000). Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases* **6**, 505-512.

Fröhlich, K. & Vogel, J. (2009). Activation of gene expression by small RNA. *Curr Opin Microbiol* **12**, 674-682.

Galperin, M. Y. & Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Research* **32**, 5452-5463.

Gattiker, A., Dessimoz, C., Schneider, A., Xenarios, I., Pagni, M. & Rougemont, J. (2009). The Microbe browser for comparative genomics. *Nucleic Acids Res* **37**, W296-299.

Gerth, U., Wipat, A., Harwood, C. R., Carter, N., Emmerson, P. T. & Hecker, M. (1996). Sequence and transcriptional analysis of *clpX*, a class-III heat-shock gene of *Bacillus subtilis*. *Gene* **181**, 77-83.

Gerth, U., Kruger, E., Derre, I., Msadek, T. & Hecker, M. (1998). Stress induction of the *Bacillus subtilis clpP* gene encoding a homologue of the proteolytic component of the Clp protease and the involvement of ClpP and ClpX in stress tolerance. *Molecular Microbiology* **28**, 787-802.

Gioia, J., Yerrapragada, S., Qin, X. & other authors (2007). Paradoxical DNA Repair and Peroxide Resistance Gene Conservation in *Bacillus pumilus* SAFR-032. *Plos One* **2**.

Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessières, P., Aymerich, S. & Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst Biol* **2**, 20.

Griffith, F. (1928). The significance of pneumococcal types. *Journal of Hygiene* **27**, 113-159.

Hacker, J. & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep* **2**, 376-381.

Hahn, J., Inamine, G., Kozlov, Y. & Dubnau, D. (1993). Characterization of *comE*, a late competence operon of *Bacillus subtilis* required for the binding and uptake of transforming DNA. *Molecular Microbiology* **10**, 99-111.

Hahn, J., Roggiani, M. & Dubnau, D. (1995). The major role of Spo0A in genetic competence is to down-regulate abrB, an essential competence gene. *Journal of Bacteriology* **177**, 3601-3605.

Hall, B., Ehrlich, G. & Hu, F. (2010). Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiology* **156**, 1060-1068.

Hamoen, L. W., Eshuis, H., Jongbloed, J., Venema, G. & van Sinderen, D. (1995). A small gene, designated *coms*, located within the coding region of the 4th amino acid activation domain of *srfA*, is required for competence development in *Bacillus subtilis*. *Molecular Microbiology* **15**, 55-63.

Hamoen, L. W., Kausche, D., Marahiel, M. A., van Sinderen, D., Venema, G. & Serror, P. (2003a). The *Bacillus subtilis* transition state regulator AbrB binds to the-35 promoter region of comK. *Fems Microbiology Letters* **218**, 299-304.

Hamoen, L. W., Venema, G. & Kuipers, O. P. (2003b). Controlling competence in *Bacillus subtilis*: shared use of regulators. *Microbiology-Sgm* **149**, 9-17.

Havarstein, L. S., Hakenbeck, R. & Gaustad, P. (1997). Natural competence in the genus *Streptococcus*: Evidence that streptococci can change pherotype by interspecies recombinational exchanges. *Journal of Bacteriology* **179**, 6589-6594.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-10919.

Henry, C., Zinner, J., Cohoon, M. & Stevens, R. (2009). iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biol* **10**, R69.

Hoffmann, K., Wollherr, A., Larsen, M., Rachinger, M., Liesegang, H., Ehrenreich, A. & Meinhardt, F. (2010). Facilitation of Direct Conditional Knockout of Essential Genes in *Bacillus licheniformis* DSM13 by Comparative Genetic Analysis and Manipulation of Genetic Competence. *Applied and Environmental Microbiology* **76**, 5046-5057.

Hofreuter, D., Odenbreit, S., Henke, G. & Haas, R. (1998). Natural competence for DNA transformation in *Helicobacter pylori:* identification and genetic characterization of the *comB* locus. *Molecular Microbiology* **28**, 1027-1038.

Hornbaek, T., Jakobsen, M., Dynesen, J. & Nielsen, A. (2004). Global transcription profiles and intracellular pH regulation measured in *Bacillus licheniformis* upon external pH upshifts. *Arch Microbiol* **182**, 467-474.

Hulsen, T., Huynen, M. A., de Vlieg, J. & Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology* **7**.

Kaimer, C. & Graumann, P. L. (2010). *Bacillus subtilis* CinA is a stationary phase-induced protein that localizes to the nucleoid and plays a minor role in competent cells. *Archives of Microbiology* **192**.

Kapushesky, M., Emam, I., Holloway, E. & other authors (2010). Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* **38**, D690-698.

Karp, P., Paley, S., Krummenacker, M. & other authors (2010). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* **11**, 40-79.

Kearns, D. B., Chu, F., Branda, S. S., Kolter, R. & Losick, R. (2005). A master regulator for biofilm formation by *Bacillus subtilis*. *Molecular Microbiology* **55**, 739-749.

Kettler, G., Martiny, A., Huang, K. & other authors (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**, e231.

Kingsford, C., Ayanbule, K. & Salzberg, S. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* **8**, R22.

Klee, S., Ozel, M., Appel, B. & other authors (2006). Characterization of *Bacillus anthracis*-like bacteria isolated from wild great apes from Cote d'Ivoire and Cameroon. *J Bacteriol* **188**, 5333-5344.

Klee, S., Brzuszkiewicz, E., Nattermann, H. & other authors (2010). The genome of a *Bacillus* isolate causing anthrax in chimpanzees combines chromosomal properties of *B. cereus* with *B. anthracis* virulence plasmids. *PLoS One* **5**, e10986.

Knippers, R. (2006). *Molekulare Genetik*, 9. überarbeitete Auflage edn. Stuttgart: Thieme.

Kobayashi, K., Ehrlich, S., Albertini, A. & other authors (2003). Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**, 4678-4683.

Kobayashi, K. (2008). SlrR/SlrA controls the initiation of biofilm formation in *Bacillus subtilis*. *Molecular Microbiology* **69**, 1399-1410.

Kolstø, A. (1997). Dynamic bacterial genome organization. *Mol Microbiol* **24**, 241-248.

Kong, Y., Ong, S., Ng, W. & Liu, W. (2002). Diversity and distribution of a deeply branched novel proteobacterial group found in anaerobic-aerobic activated sludge processes. *Environ Microbiol* **4**, 753-757.

Konstantinidis, K. & Tiedje, J. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**, 6258-6264.

Kovacs, A. T., Smits, W. K., Mironczuk, A. M. & Kuipers, O. P. (2009). Ubiquitous late competence genes in *Bacillus* species indicate the presence of functional DNA uptake machineries. *Environmental Microbiology* **11**, 1911-1922.

Kunst, F., Ogasawara, N., Moszer, I. & other authors (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Kuznetsova, E., Proudfoot, M., Sanders, S., Reinking, J., Savchenko, A., Arrowsmith, C., Edwards, A. & Yakunin, A. (2005). Enzyme genomics: Application of general enzymatic screens to discover new enzymes. *FEMS Microbiol Rev* **29**, 263-279.

Köpke, M., Held, C., Hujer, S. & other authors (2010). *Clostridium ljungdahlii* represents a microbial production platform based on syngas. *Proc Natl Acad Sci U S A* **107**, 13087-13092.

Lammers, C., Flórez, L., Schmeisky, A., Roppel, S., Mäder, U., Hamoen, L. & Stülke, J. (2010). Connecting parts with processes: SubtiWiki and SubtiPathways integrate gene and pathway annotation for *Bacillus subtilis*. *Microbiology* **156**, 849-859.

Lander, E. S., Linton, L. M., Birren, B. & other authors (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Langille, M. & Brinkman, F. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664-665.

Lazazzera, B. A., Kurtser, I. G., McQuade, R. S. & Grossman, A. D. (1999). An autoregulatory circuit affecting peptide signaling in *Bacillus subtilis*. *Journal of Bacteriology* **181**, 5193-5200.

Li, L., Stoeckert, C. J. & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189.

Liesegang, H., Kaster, A., Wiezer, A., Goenrich, M., Wollherr, A., Seedorf, H., Gottschalk, G. & Thauer, R. (2010). Complete Genome Sequence of *Methanothermobacter marburgensis*, a methanoarchaeon model organism (07-08-10). *J Bacteriol*.

Lindner, C., Nijland, R., van Hartskamp, M., Bron, S., Hamoen, L. W. & Kuipers, O. P. (2004). Differential expression of two paralogous genes of *Bacillus subtilis* encoding single-stranded DNA binding protein. *Journal of Bacteriology* **186**, 1097-1105.

Liolios, K., Chen, I. M. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M. & Kyrpides, N. C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **38**, D346-D354.

Liu, L., Agren, R., Bordel, S. & Nielsen, J. (2010). Use of genome-scale metabolic models for understanding microbial physiology. *FEBS Lett* **584**, 2556-2564.

Londonovallejo, J. A. & Dubnau, D. (1993). ComF a *Bacillus subtilis* late competence locus, encodes a protein similar to ATP-dependent. *Molecular Microbiology* **9**, 119-131.

Lorenz, M. G. & Wackernagel, W. (1994). Bacterial gene-transfer by natural genetic-transformation in the environment. *Microbiological Reviews* **58**, 563-602.

Lukjancenko, O., Wassenaar, T. & Ussery, D. (2010). Comparison of 61 Sequenced *Escherichia coli* Genomes. *Microb Ecol*.

Luttinger, A., Hahn, J. & Dubnau, D. (1996). Polynucleotide phosphorylase is necessary for competence development in *Bacillus subtilis*. *Molecular Microbiology* **19**, 343-356.

Maamar, H. & Dubnau, D. (2005). Bistability in the *Bacillus subtilis* K-state (competence) system requires a positive feedback loop. *Molecular Microbiology* **56**, 615-624.

Madigan, M. T. & Martinko, J. M. (2006). *Brock Mikrobiologie*, 11. überarbeitete Auflage edn: Pearson Studium.

Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol Mol Biol Rev* **62**, 725-774.

Maiden, M. (2006). Multilocus sequence typing of bacteria. *Annu Rev Microbiol* **60**, 561-588.

Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Res* **37**, D459-463.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**, 387-402.

Martin, J., Zhu, W., Passalacqua, K., Bergman, N. & Borodovsky, M. (2010). *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics* **11 Suppl 3**, S10.

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development* **15**, 589-594.

Milne, C., Kim, P., Eddy, J. & Price, N. (2009). Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnol J* **4**, 1653-1670.

Mira, A., Martín-Cuadrado, Ana B., D'Auria, Giuseppe, & Rodríguez-Valera, F. (2010). The bacterial pan-genome: a new paradigm in microbiology, pp. 45-57. International Microbiology.

Mirończuk, A. M., Kovács, Á. T. & Kuipers, O. P. (2008). Induction of natural competence in Bacillus cereus ATCC14579. *Microbial Biotechnology* **1**, 226-235.

Mizoguchi, H., Mori, H. & Fujio, T. (2007). *Escherichia coli* minimum genome factory. *Biotechnol Appl Biochem* **46**, 157-167.

Moreno-Hagelsieb, G. & Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl 1**, S329-336.

Morimoto, T., Kadoya, R., Endo, K. & other authors (2008). Enhanced recombinant protein productivity by genome reduction in *Bacillus subtilis*. *DNA Research* **15**, 73-81.

Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*, 2. ed. edn. Cold Spring Harbor, New York.

Msadek, T., Kunst, F., Klier, A. & Rapoport, G. (1991). DegS-DegU and ComP-ComA modulator effector pairs control expression of the *Bacillus subtilis* pleiotropic regulatory gene. *Journal of Bacteriology* **173**, 2366-2377.

Msadek, T., Kunst, F. & Rapoport, G. (1994). MecB of *Bacillus subtilis*, a member of the ClpC ATPase family, is a pleiotropic regulator controlling competence gene expression and growth at high temperature. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 5788-5792.

Musher, D. M. (1992). Infections caused by *Streptococcus pneumoniae*: clinical spectrum, pathogenesis, immunity, and treatment. *Clinical Infectious Diseases* **14**, 801-809.

Nagai, T., Tran, L. S. P., Inatsu, Y. & Itoh, Y. (2000). A new IS4 family insertion sequence, IS4Bsu1, responsible for genetic instability of poly-gamma-glutamic acid production in *Bacillus subtilis*. *Journal of Bacteriology* **182**, 2387-2392.

Nakamura, C. & Whited, G. (2003). Metabolic engineering for the microbial production of 1,3-propanediol. *Curr Opin Biotechnol* **14**, 454-459.

Nakano, M. M. & Zuber, P. (1991). The primary role of ComA in establishment of the competent state in *Bacillus subtilis* is to activate expression of *srfA*. *Journal of Bacteriology* **173**, 7269-7274.

Nedenskovsorensen, P., Bukholm, G. & Bovre, K. (1990). Natural competence for genetic-transformation in *Campylobacter pylori Journal of Infectious Diseases* **161**, 365-366.

Needleman , S. B. & Wunsch, C. D. (1970). A general method applicable to search for similarities in amino acid sequence of 2 proteins. *Journal of Molecular Biology* **48**, 443-453.

Nielsen, A., Breüner, A., Krzystanek, M., Andersen, J., Poulsen, T., Olsen, P., Mijakovic, I. & Rasmussen, M. (2010). Global transcriptional analysis of *Bacillus licheniformis* reveals an overlap between heat shock and iron limitation stimulon. *J Mol Microbiol Biotechnol* **18**, 162-173.

Nielsen, J. (2001). Metabolic engineering. *Appl Microbiol Biotechnol* **55**, 263-283.

Nissen, T., Kielland-Brandt, M., Nielsen, J. & Villadsen, J. (2000). Optimization of ethanol production in Saccharomyces cerevisiae by metabolic engineering of the ammonium assimilation. *Metab Eng* **2**, 69-77.

O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research* **33**, D476-D480.

Ogura, M., Ohshiro, Y., Hirao, S. & Tanaka, T. (1997). A new *Bacillus subtilis* gene, *med*, encodes a positive regulator of *comK*. *Journal of Bacteriology* **179**, 6244-6253.

Ogura, M., Liu, L., LaCelle, M., Nakano, M. M. & Zuber, P. (1999). Mutational analysis of ComS: evidence for the interaction of ComS and MecA in

the regulation of competence development in *Bacillus subtilis*. *Molecular Microbiology* **32**, 799-812.

Ogura, M. & Tanaka, T. (2000). *Bacillus subtilis comZ* (*yjzA*) negatively affects expression of *comG* but not *comK*. *Journal of Bacteriology* **182**, 4992-4994.

Oh, Y., Palsson, B., Park, S., Schilling, C. & Mahadevan, R. (2007). Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282**, 28791-28799.

Otero, J. & Nielsen, J. (2010). Industrial systems biology. *Biotechnol Bioeng* **105**, 439-460.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* **1**, 93-108.

Overbeek, R., Larsen, N., Walunas, T. & other authors (2003). The ERGO (TM) genome analysis and discovery system. *Nucleic Acids Research* **31**, 164-171.

Paredes, C., Alsaker, K. & Papoutsakis, E. (2005). A comparative genomic view of clostridial sporulation and physiology. *Nat Rev Microbiol* **3**, 969-978.

Park, J., Lee, S., Kim, T. & Kim, H. (2008). Application of systems biology for bioprocess development. *Trends Biotechnol* **26**, 404-412.

Parkinson, H., Kapushesky, M., Kolesnikov, N. & other authors (2009). ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* **37**, D868-872.

Passalacqua, K., Varadarajan, A., Ondov, B., Okou, D., Zwick, M. & Bergman, N. (2009). Structure and complexity of a bacterial transcriptome. *J Bacteriol* **191**, 3203-3211.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 2444-2448.

Perego, M., Higgins, C. F., Pearce, S. R., Gallagher, M. P. & Hoch, J. A. (1991). The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. *Molecular Microbiology* **5**, 173-185.

Pertea, M., Ayanbule, K., Smedinghoff, M. & Salzberg, S. (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res* **37**, D479-482.

Piazza, F., Tortosa, P. & Dubnau, D. (1999). Mutational analysis and membrane topology of ComP, a quorum-sensing histidine kinase of *Bacillus subtilis* controlling competence development. *Journal of Bacteriology* **181**, 4540-4548.

Porwal, S., Lal, S., Cheema, S. & Kalia, V. (2009). Phylogeny in aid of the present and novel microbial lineages: diversity in Bacillus. *PLoS One* **4**, e4438.

Posch, S., Grau, J., Gohr, A., Keilwagen, J. & Grosse, I. (2010). Probabilistic approaches to transcription factor binding site prediction. *Methods Mol Biol* **674**, 97-119.

Pottathil, M. & Lazazzera, B. A. (2003). The extracellular Phr peptide-rap phosphatase signaling circuit of Bacillus subtilis. *Frontiers in Bioscience* **8**, D32-D45.

Prilusky, J., Oueillet, E., Ulryck, N. & other authors (2005). HalX: an open-source LIMS (Laboratory Information Management System) for small- to large-scale laboratories. *Acta Crystallogr D Biol Crystallogr* **61**, 671-678.

Provvedi, R., Chen, I. & Dubnau, D. (2001). NucA is required for DNA cleavage during transformation of *Bacillus subtilis*. *Molecular Microbiology* **40**, 634-644.

Quinn, C. P. & Dancer, B. N. (1990). Transformation of vegetative cells of *Bacillus anthracis* with plasmid DNA. *Journal of General Microbiology* **136**, 1211-1215.

Rasmussen, S., Nielsen, H. & Jarmer, H. (2009). The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol* **73**, 1043-1057.

Rey, M. W., Ramaiya, P., Nelson, B. A. & other authors (2004). Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biology* **5**.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics* **16**, 276-277.

Richter, B. G. & Sexton, D. P. (2009). Managing and Analyzing Next-Generation Sequence Data. *Plos Computational Biology* **5**.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.

Sanger, F., Nicklen, S. & Coulson, A. R. (1992). DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology* **24**, 104-108.

Sathish, S. & Swaminathan, K. (2009). Genetic diversity among toxigenic clostridia isolated from soil, water, meat and associated polluted sites in South India. *Indian J Med Microbiol* **27**, 311-320.

Schallmey, M., Singh, A. & Ward, O. P. (2004). Developments in the use of *Bacillus* species for industrial production. *Canadian Journal of Microbiology* **50**, 1-17.

Schlothauer, T., Mogk, A., Dougan, D. A., Bukau, B. & Turgay, K. (2003). MecA, an adaptor protein necessary for ClpC chaperone activity. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2306-2311.

Schmeisser, C., Liesegang, H., Krysciak, D. & other authors (2009). *Rhizobium sp.* strain NGR234 possesses a remarkable number of secretion systems. *Appl Environ Microbiol* **75**, 4035-4045.

Schneider, K. B., Palmer, T. M. & Grossman, A. D. (2002). Characterization of *comQ* and *comX*, two genes required for production of ComX pheromone in *Bacillus subtilis*. *Journal of Bacteriology* **184**, 410-419.

Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., Zeigler, D. R. & Dean, D. H. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiology and Molecular Biology Reviews* **62**, 775-+.

Schwartz, R. M. & Dayhoff, M. O. (1978). Improved scoring matrix for identifying evolutionary relatedness among proteins. *Biophysical Journal* **21**, A198-A198.

Serrano, M., Hovel, S., Moran, C. P., Henriques, A. O. & Volker, U. (2001). Forespore-specific transcription of the *lonB* gene during sporulation in *Bacillus subtilis*. *Journal of Bacteriology* **183**, 2995-3003.

Serror, P. & Sonenshein, A. L. (1996). CodY is required for nutritional repression of *Bacillus subtilis* genetic competence. *Journal of Bacteriology* **178**, 5910-5915.

Setlow, P. (2006). Spores of *Bacillus subtilis*: their resistance to and killing by radiation, heat and chemicals. *Journal of Applied Microbiology* **101**, 514-525.

Shi, S., Chen, T., Zhang, Z., Chen, X. & Zhao, X. (2009). Transcriptome analysis guided metabolic engineering of *Bacillus subtilis* for riboflavin production. *Metab Eng* **11**, 243-252.

Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* **36**, D93-96.

Snipen, L., Almøy, T. & Ussery, D. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* **10**, 385.

Solomon, J. M., Lazazzera, B. A. & Grossman, A. D. (1996). Purification and characterization of an extracellular peptide factor that affects two different developmental pathways in *Bacillus subtilis*. *Genes & Development* **10**, 2014-2024.

Sonenshine, A. L., Losick, R. & Hoch, J. A. (2001). *Bacillus Subtilis and Its Closest Relatives: From Genes to Cells*: Asm Press.

Sorek, R. & Cossart, P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* **11**, 9-16.

Spizizen, J. (1958). Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proceedings of the National Academy of Sciences of the United States of America* **44**, 1072-1078.

Steensma, H., Robertson, L. & van Elsas, J. (1978). The occurrence and taxonomic value of PBSX-like defective phages in the genus *Bacillus*. *Antonie Van Leeuwenhoek* **44**, 353-366.

Stefanic, P. & Mandic-Mulec, I. (2009). Social Interactions and Distribution of *Bacillus subtilis* Pherotypes at Microscale. *Journal of Bacteriology* **191**, 1756-1764.

Stocker, G., Fischer, M., Rieder, D., Bindea, G., Kainz, S., Oberstolz, M., McNally, J. & Trajanoski, Z. (2009). iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis. *BMC Bioinformatics* **10**, 390.

Tadesse, S. & Graumann, P. L. (2007). DprA/Smf protein localizes at the DNA uptake machinery in competent *Bacillus subtilis* cells. *Bmc Microbiology* **7**.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A., Eriksson, A., Wernegreen, J., Sandström, J., Moran, N. & Andersson, S. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376-2379.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D. & other authors (2003). The COG database: an updated version includes eukaryotes. *Bmc Bioinformatics* **4**.

Tettelin, H., Riley, D., Cattuto, C. & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11**, 472-477.

Thompson, J., Higgins, D. & Gibson, T. (1994).CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, pp. 4673-4680. Nucleic Acids Res.

Thorne, C. & Stull, H. (1966). Factors affecting transformation of *Bacillus licheniformis*. *J Bacteriol* **91**, 1012-1020.

Tortosa, P., Albano, M. & Dubnau, D. (2000). Characterization of *ylbF*, a new gene involved in competence development and sporulation in *Bacillus subtilis*. *Molecular Microbiology* **35**, 1110-1119.

Tortosa, P., Logsdon, L., Kraigher, B., Itoh, Y., Mandic-Mulec, I. & Dubnau, D. (2001). Specificity and genetic polymorphism of the *Bacillus* competence quorum-sensing system. *Journal of Bacteriology* **183**, 451-460.

Tran, L. S. P., Nagai, T. & Itoh, Y. (2000). Divergent structure of the ComQXPA quorum-sensing components: molecular basis of strain-specific communication mechanism in *Bacillus subtilis*. *Molecular Microbiology* **37**, 1159-1171.

van Sinderen, D., Withoff, S., Boels, H. & Venema, G. (1990). Isolation and characterization of *comL*, a transcription unit involved in competence development of *Bacillus subtilis*. *Molecular & General Genetics* **224**, 396-404.

van Sinderen, D. & Venema, G. (1994). ComK acts as an autoregulatory control switch in the signal transduction route to competence in *Bacillus subtilis*. *Journal of Bacteriology* **176**, 5762-5770.

Veith, B., Herzberg, C., Steckel, S. & other authors (2004). The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *Journal of Molecular Microbiology and Biotechnology* **7**, 204-211.

Venter, J. C. (2001). The sequence of the human genome (vol 292, pg 1304, 2001). *Science* **292**, 1838-1838.

Waldeck, J., Daum, G., Bisping, B. & Meinhardt, F. (2006). Isolation and molecular characterization of chitinase-deficient *Bacillus licheniformis* strains capable of deproteinization of shrimp shell waste to obtain highly viscous chitin. *Applied and Environmental Microbiology* **72**, 7879-7885.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.

Waschkau, B., Waldeck, J., Wieland, S., Eichstadt, R. & Meinhardt, F. (2008). Generation of readily transformable *Bacillus licheniformis* mutants. *Applied Microbiology and Biotechnology* **78**, 181-188.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Weir, J., Predich, M., Dubnau, E., Nair, G. & Smith, I. (1991). Regulation of *spo0h*, a gene coding for the *Bacillus subtilis* sigma H factor. *Journal of Bacteriology* **173**, 521-529.

Werner, T. (2002). Finding and decrypting of promoters contributes to the elucidation of gene function. *In Silico Biol* **2**, 249-255.

Woese, C., Fox, G., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B. & Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**, 83-86.

Zheng, Y., Szustakowski, J., Fortnow, L., Roberts, R. & Kasif, S. (2002). Computational identification of operons in microbial genomes. *Genome Res* **12**, 1221-1230.

9 Anhang

9.1 Inhaltsverzeichnis der Daten-CD

Ordner sind unterstrichen.

Daten
- 3.3.1 Kompetenzsystemvergleiche
  - Nukleotidvergleiche_Subtilis.xlsx, Nukleotidvergleiche_Licheniformis.xlsx
  - Nukleotidvergleiche_Amyloliquefaciens.xlsx,
  - Nukleotidvergleiche_Pumilus.xlsx
  - Nukleotidvergleiche_Sterne.xlsx,
  - Proteinvergleiche_Subtilis.xlsx
  - Proteinvergleiche_Licheniformis.xlsx,
  - Proteinvergleiche_Amyloliquefaciens.xlsx
  - Proteinvergleiche_Pumilus.xlsx,
  - Proteinvergleiche_Sterne.xlsx
- 4 BiBaG
  - 4.3 Test
    - 9945A
      - BiBaG-Ergebnisse mit biblast.xlsx
      - 9945_falsch_negative.xls
    - DSM13
      - BiBaG-Ergebnisse mit biblast.xlsx
      - DSM13_falsch_negative.xls
    - Subtilis
      - BiBaG-Ergebnisse mit biblast.xlsx
      - Subtilis_falsch_negative.xls
  - 4.4.1 Bacilli
    - mitPlasmiden
      - BiBaG-Ergebnisse mit biblast.xlsx
      - DSM13_orthologe_28.xlsx
      - biblast_core_sorted.xlsx
      - pan_10_berechnung.xlsx, pan_10_graph.xlsx
    - ohnePlasmide
      - BiBaG-Ergebnisse mit biblast.xlsx
      - DSM13_orthologe_5.xlsx
      - DSM13_specific.xlsx
    - Triple
      - Triple-Ergebnisse mit biblast.xlsx, biblast1.xlsx, biblast2.xlsx
      - core genomes
        - BAC_common_blast, BAC_common_needle25
        - BAC_common_needle90
        - BAM_common_blast, BAM_common_needle25
        - BAM_common_needle90
        - BL_common_blast, BL_common_needle25
        - BL_common_needle90
    - essential_dsm13_abgleich_tripleBiBaG.xls
    - markierteBereicheDSM13.xls
    - markierteCluster.xls
  - 4.4.2 Annotation
    - 9945A
      - BiBaG-Ergebnisse mit biblast.xlsx
      - 9945_annotated.xls
      - 9945_biblast_filter70.xls
    - DSM13
      - BiBaG-Ergebnisse mit biblast.xlsx
      - DSM13_annotated.xls

DSM13_biblast_filter70.xls
common.xls
unique.xls
4.4.3 Deletionstargets
Subtilis
BiBaG-Ergebnisse mit biblast.xls
deleted_regions.subtilis.xls (rot = deleted regions, blau = seconddeletions)
essential_subtilis.xls
GFF+EMBL
BL_cluster.gff
BS_deletedregions.gff
BS_seconddeletions.gff
BS_essentiell.gff
AL009126
DSM13
DSM13_biblast_bearbeitet.xls
essential_dsm13.xls
mapped_deletionregions_enhanced.xls
mapped_firstdeletions.xls
mapped_seconddeletions_including_first.xls
GFF+EMBL
BL_essential.gff
BL_first_deletions.gff
BL_second_deletions_including_first.xls
BL_deletions_enhanced.gff
AE017333
4.4.4 Insertionstargetbestimmung
BAM
BiBaG-Ergebnisse mit biblast.xls
möglicheInsertionsTargets_BAM.xls
BLI
BiBaG-Ergebnisse mit biblast.xls
möglicheInsertionsTargets_BAM.xls
BSU
BiBaG-Ergebnisse mit biblast.xls
möglicheInsertionsTargets_BAM.xls
GBK
Ergebnisse mit biblast.xls
möglicheInsertionsTargets_BAM.xls
5 DB
Anleitung zur Erstellung der DB und Ergebnisgenerierung.doc
init_db_2010
BiBaG_Abgleich.xlsx
Eingabedaten
7SB
bl_fasta_result
experiment
AE017333, AL009126, CP000002, CP000903
Ausgabedaten
experiment.gff
operon.gff
Software
BiBaG
BiBaG.zip
DBTools
DBTools.jar
artemis mit angepasster *options*-Datei
DNAPlotter mit angepasster *options*-Datei
README

9.2 Genumgebungen der Kompetenzgene

Die Abbildungen 49 – 62 zeigen die Genumgebungen, die im Vergleich von *B. subtilis* 168 und *B. licheniformis* DSM13 keine Insertions- oder Deletionsereignisse aufweisen.

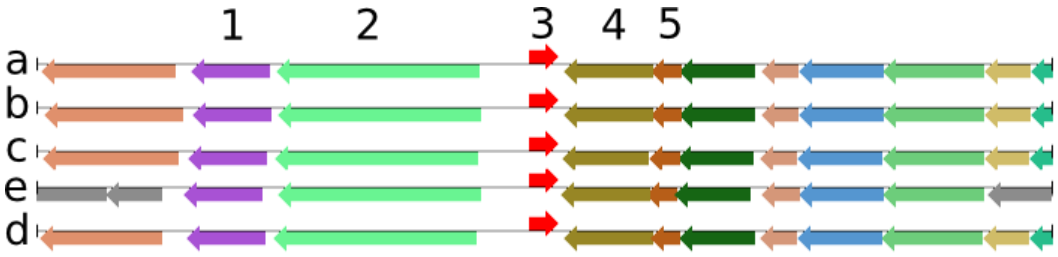


Abbildung 49: Genumgebung von *abrB*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *yabD*, (2) *metS*, (3) *abrb*, (4) *yabC*, (5) *yazA*

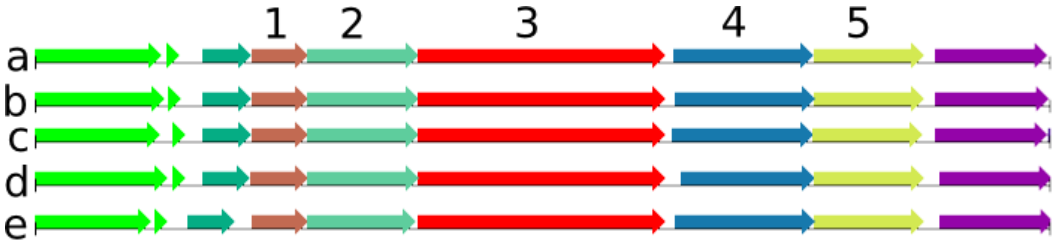


Abbildung 50: Genumgebung von *clpC*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *mcsA*, (2) *mcsB*, (3) *clpC*, (4) *radA*, (5) *yacK*

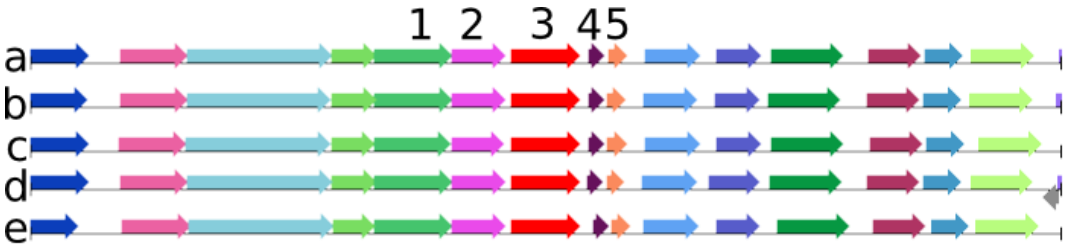


Abbildung 51: Genumgebung von *sigH*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *yacO*, (2) *yacO*, (3) *sigH* , (4) *rpmGB*, (5) *secE*

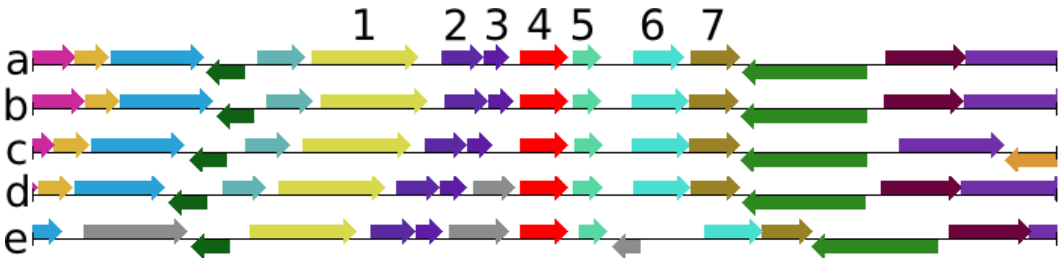


Abbildung 52: Genumgebung von *ylbF*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *ylbC*, (2) *ylbD*, (3) *ylbE*, (4) *ylbF*, (5) *ylbG*, (6) *ylbH*, (7) *ylbI*

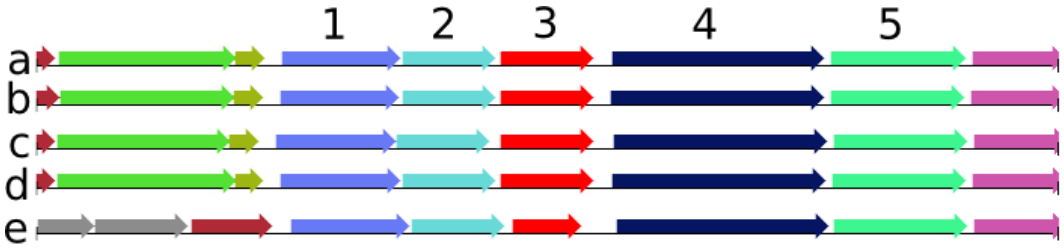


Abbildung 53: Genumgebung von *smf*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *sucC*, (2) *sucD*, (3) *smf*, (4) *topA*, (5) *gid*

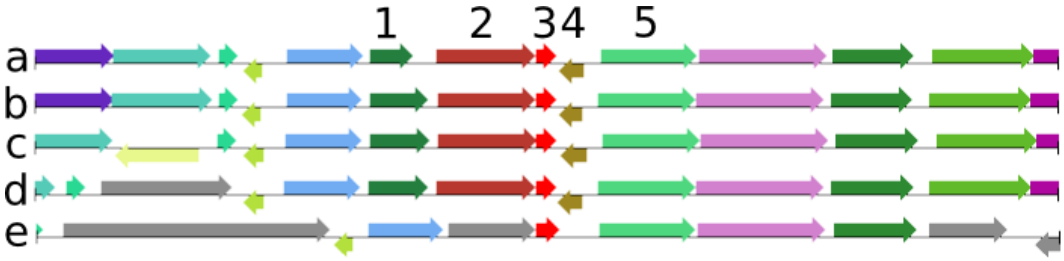


Abbildung 54: Genumgebung von *comZ*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *yjaV*, (2) *med*, (3) *comZ*, (4) *yjzB*, (5) *fabHA*

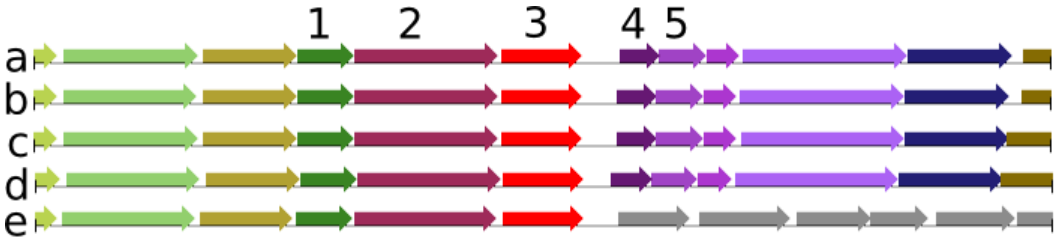


Abbildung 55: Genumgebung von *codY*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *clpQ*, (2) *clpY*, (3) *codY*, (4) *flgB*, (5) *flgC*

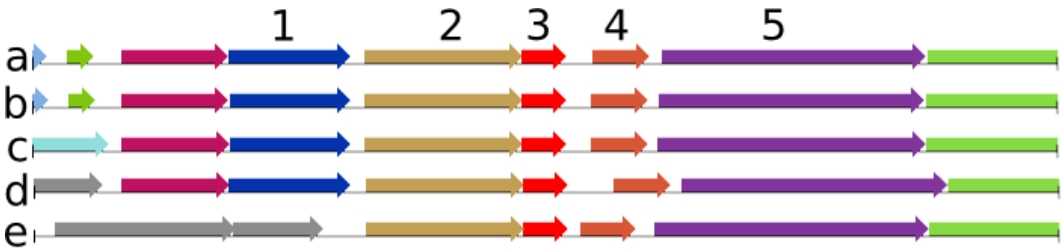


Abbildung 56: Genumgebung von *ymcA*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *kbl*, (2) *ymcB*, (3) *ymcA*, (4) *cotE*, (5) *mutS*

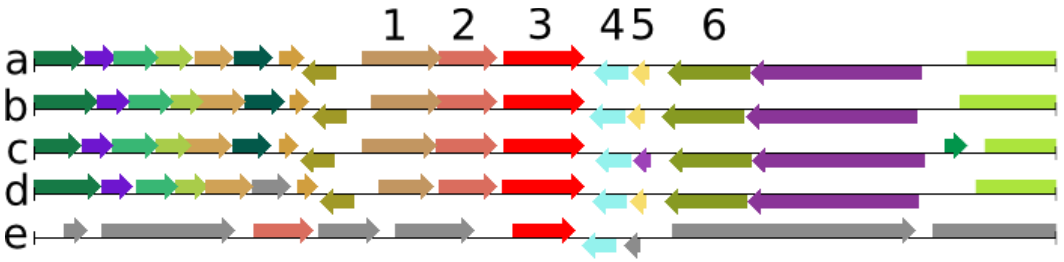


Abbildung 57: Genumgebung von *sinI*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *yqxM*, (2) *sipW*, (3) *tasA*, (4) *sinR*, (5) *sinI*, (6) *yghG*

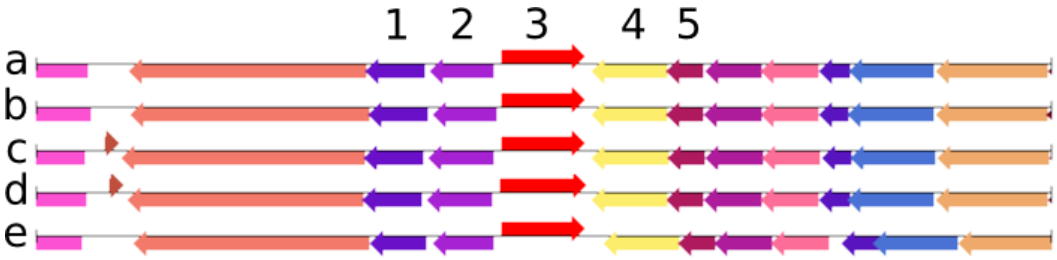


Abbildung 58: Genumgebung von *comER*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *comeB*, (2) *comEA*, (3) *comER*, (4) *yqeM*, (5) *yqeL*

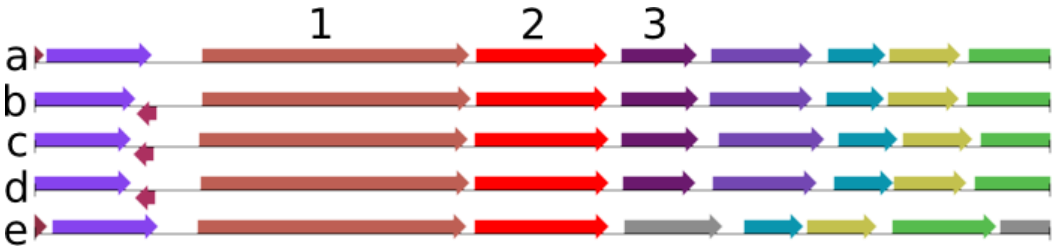


Abbildung 59: Genumgebung von *comC*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *folC*, (2) *comC*, (3) *spoIIB*

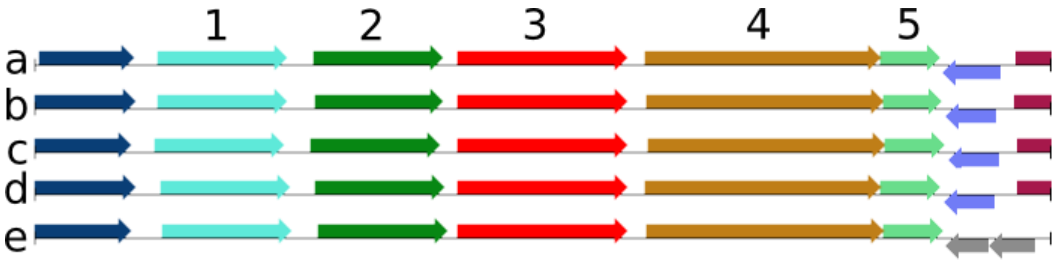


Abbildung 60: Genumgebung von *lonB / clpX*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *tig*, (2) *clpX*, (3) *lonB*, (4) *lonA*, (5) *ysxC*

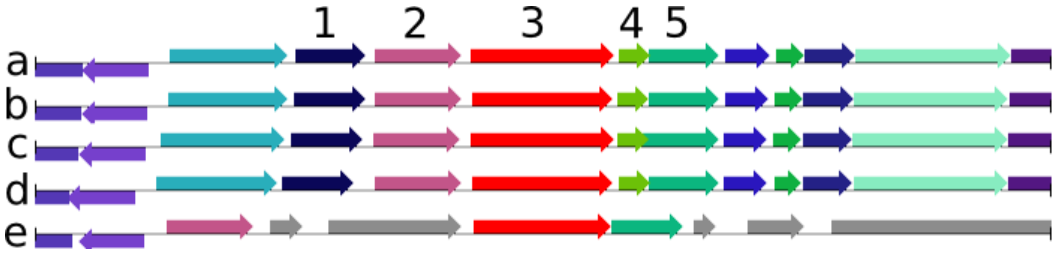


Abbildung 61: Genumgebung von *comCF*
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *degU*, (2) *yviA*, (3) *comFA*, (4) *comFB*, (5) *comFC*

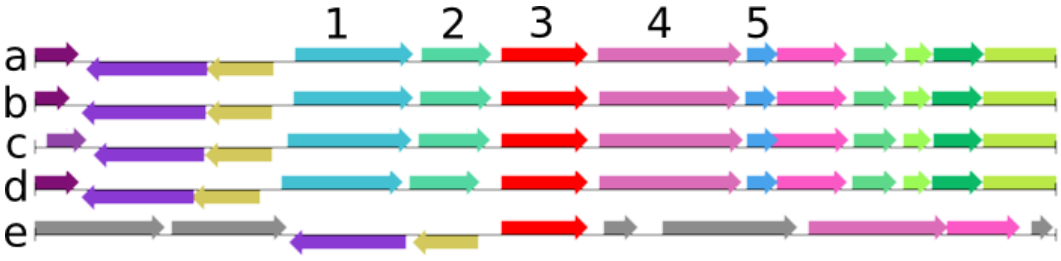


Abbildung 62: Genumgebung des *deg*-Operons
Zuordnung der Genumgebungen zu den Organismen: (a) *B. subtilis* 168, (b) *B. licheniformis* DSM13, (c) *B. amyloliquefaciens* FZB42, (d) *B. pumilus* SAFR-032 , (e) *B. anthracis* str. Sterne; Zuordnung der Gene: (1) *degS*, (2) *degU*, (3) *yviA*, (4) *comFA*, (5) *comFB*

Danksagung

Mein besonderer Dank gilt Prof. Wolfgang Liebl für die Überlassung dieses Themas, die Ermöglichung des selbstständigen Arbeitens und seine Reisebereitschaft zu den Jahresfortschrittsberichten und der finalen Prüfung.

Mein spezieller Dank gilt meinem Anleiter und Mentor Dr. Heiko Liesegang, der mir nicht nur die Welt der Mikroorganismen und seine Faszination dafür näher brachte, sondern auch jederzeit mit regem Interesse und Diskussionsbereitschaft am Fortgang der Arbeit beteiligt war. Danke, dass du auch immer mich als Mensch und meine persönliche Entwicklung im Blick hattest.

Weiterhin möchte ich mich bei den Kooperationspartnern der Firma Henkel, Dr. Maurer, Dr. Evers und Dr. Bongaerts für die Zusammenarbeit bedanken.

Mein Dank gilt außerdem Prof. Burkhard Morgenstern für die Übernahme des Korreferats.

Den alten und neuen Mitgliedern der *Bacillus*-Gruppe möchte ich für die fachspezifischen Anregungen und Diskussionen danken. Insbesondere gilt mein Dank „meinem Mitdoktoranden“ und „Leidensgenossen“ Michael Rachinger, der niemals Müde wurde, mir mikrobiologische Zusammenhänge näher zu erläutern. Ich habe die Zusammenarbeit mit dir und unseren gemischten Kompetenzen als sehr wert- und reizvoll erlebt. Marco Schwarzer danke ich an dieser Stelle für die Überlassung seiner *microarray*-Daten und die umzugsbedingten leider selten gewordenen (Fach)-Gespräche. Sonja (Volland), ich bin froh, dass wir uns kennenlernen durften. Du bist nicht nur eine Bereicherung für die *Bacillus*-Gruppe…

Weiterhin gilt mein Dank dem gesamten G2L für das tolle Arbeitsklima in mehr als 3 Jahren Zusammenarbeit. Danke, für euer Vertrauen in mich und meine bioinformatischen Arbeiten. Katrin und Sandra möchte ich für die immer aufmunternden Worte, insbesondere während der Kaffepausen, danken. Dem Bioinformatik-Office danke ich besonders für eine freundschaftliche Arbeitsatmosphäre.

Außerdem möchte ich mich bei der AG Daniel für die erfolgreiche Zusammenarbeit bedanken. Hier gilt mein besonderer Dank den (noch) Doktoranden Christiane und Birgit für die Flur- und Kaffeegespräche.

Fürs Korrekturlesen gilt mein spezieller Dank Heiko, Christiane, Rüdiger und Susanne, die die Arbeit mit besonderer Sorgfalt durchgelesen haben.

Weiterhin möchte ich mich bei meiner guten Freundin Marie bedanken, die mich seit so vielen Jahren unterstützt, aufmuntert und mit jeder Begegnung mein Leben ein wenig bunter macht.

Ein riesiges Dankeschön an meine Freunde und Weggefährten in Hamburg, Bremen, Lüneburg, Hannover, Braunschweig, Göttingen, Oldenburg und Peine, die mich in allen Krisenzeiten auffangen und/ oder mir immer wieder wahlweise eine nette Begegnung, ein leckeres Essen, ein Bett, eine Tischtennishalle oder alles auf einmal bescheren.

Meiner Familie möchte ich dafür danken, dass ich da bin, wo ich heute bin. Danke, dass Ihr mir immer freie Hand gelassen habt, egal wie weit weg mich Studium und Promotion geführt haben und dass ihr an mich glaubt. Ein großes Dankeschön auch an meine Lieblings-Cousine Susanne dafür, dass es dich wieder in meinem Leben gibt, du hinter mir stehst und mir Kraft gibst.

Lebenslauf

14.05.1983	Geburt in Peine, deutsche Staatsbürgerschaft
1989-1993	Besuch der Eichendorff Grundschule, Peine
1993-1995	Besuch der Orientierungsstufe Bodenstedtschule, Peine
1995-2002	Besuch des Gymnasiums am Silberkamp, Peine
06/ 2002	Abitur
10/ 2002 – 03/ 2005	Diplom-Informatik-Studium an der Universität Hamburg
03/ 2005	Vordiplom in Informatik
04/ 2005 - 06/ 2007	Studium Diplom-Bioinformatik am Zentrum für Bioinformatik in Hamburg; Bioinformatische Diplomarbeit mit dem Titel: „Statistische Untersuchung der Genotyp-Phänotyp-Korrelation mit Hilfe von chip-basierten Techniken“
06/ 2007	Erlangung des Diploms in Bioinformatik
07/ 2007	Beginn der Promotion am Institut für Mikrobiologie und Genetik der Georg-August-Universität im Göttinger Genomlabor mit dem Titel: „Komparative Genomanalyse zur Stammoptimierung produktionsnaher *Bacillus*-Stämme“