STRUCTURAL CHARACTERIZATION OF THE LYSOSOMAL 66.3 KDA PROTEIN AND OF THE DNA REPAIR ENZYME MTH0212 BY MEANS OF X-RAY CRYSTALLOGRAPHY

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

vorgelegt von
Kristina Lakomek
aus
Göttingen

Göttingen 2009

D 7

Referent: Herr Prof. Dr. Ralf Ficner
Abteilung Molekulare Strukturbiologie
Institut für Mikrobiologie und Genetik
Georg-August-Universität Göttingen

Korreferent: Herr Prof. Dr. Kai Tittmann
Abteilung Bioanalytik
Albrecht-von-Haller-Institut für Pflanzenwissenschaften
Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 28.04.2009

TABLE OF CONTENTS

X) 1. ZUSAMMENFASSUNG 1

X) 2. SUMMARY 4

X) 3. DANKSAGUNGEN 7

PART I STRUCTURAL CHARACTERIZATION OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE BY MEANS OF X-RAY CRYSTALLOGRAPHY

I) 1. INTRODUCTION 8

1.1. **Lysosomal Proteins** **8**

1.1.1. The lysosomal compartment 8

1.1.2. Functions of lysosomal proteins 8

1.1.3. Transport of lysosomal proteins to the lysosome accompanied by co- and post-translational modifications 10

1.1.4. Identification of novel lysosomal proteins in proteomics approaches ... 11

1.2. **The 66.3 kDa Protein:**
Bioinformatic analysis and molecular characterization **12**

1.3. **X-ray Crystallography** **15**

1.3.1. Crystallographic methods 15

1.3.2. Sulfur SAD Phasing 16

1.4. **Structure Based Function Prediction** **18**

2. OBJECTIVES 19

3. DE NOVO SULFUR SAD PHASING OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE 19

3.1. **Objectives and authors` contributions** **19**

3.2. **Publication "*De novo* sulfur SAD phasing of the lysosomal 66.3 kDa protein from mouse"** **20**

3.3. **Supplementary Material** **30**

4. INITIAL INSIGHT INTO THE FUNCTION OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE BY MEANS OF X-RAY CRYSTALLOGRAPHY 33

4.1. **Objectives and authors` contributions** **33**

4.2. **Submitted manuscript "Initial insight into the function of the lysosomal 66.3 kDa protein from mouse by means of X-ray crystallography "** **34**

5. DISCUSSION 87

6. CONCLUSION AND FUTURE PERSPECTIVES 94

7. REFERENCES 96

Part II Structural Characterization of the DNA Repair Enzyme Mth0212 from the Thermophilic Archaeon Methanothermobacter Thermoautotrophicus Alone as well as in Complex with Different Substrate DNAs

8. Introduction **101**

8.1. Damages in DNA **101**

8.1.1. Causes of DNA damage 102

8.1.2. Kinds of DNA damage 102

8.1.3. 2`-Deoxyuridine residues arising in DNA 103

8.2. DNA Damage Repair **103**

8.2.1. The base excision repair (BER) pathway 103

8.2.2. Uracil DNA glycosylases generating apurinic / apyrimidinic (AP) sites 104

8.2.3. AP endonucleases 104

8.2.4. The exonuclease III homologue Mth0212 from the thermophilic archaeon *Methanothermobacter thermoautotrophicus* 106

9. Objectives 112

10. 3`-5` Exo Competes with 2`-Deoxyuridine Endonuclease Activity: Structures of the Archaeal ExoIII Homologue Mth0212 Without and in Complex with Different Substrate DNAs Shed Light on the Unique Combination of DNA Repair Activities 112

10.1. Objectives and authors` contributions 113

10.2. Manuscript in preparation “3`-5` Exo Competes with 2`-Deoxyuridine Endonuclease Function in Mth0212-DNA Complex Structures” 113

11. Supplemental Results and Discussion 186

12. Conclusion and future perspectives 205

13. References 206

APPENDIX 1 **Abbreviations** 209

APPENDIX 2 **Curriculum Vitae** 212

X) 1. ZUSAMMENFASSUNG

Diese Doktorarbeit gliedert sich in zwei Abschnitte, welche beide die strukturelle Charakterisierung von Makromolekülen mittels Röntgenkristallographie beinhalten. Abschnitt I befasst sich mit dem lysosomalen 66.3 kDa Protein aus Maus, dessen zelluläre Funktion bislang nicht bekannt ist. In diesem Zusammenhang sind die erweiterte Anwendung zur Verfügung stehender kristallographischer Methoden sowie die anschließende Analyse der Struktur des 66.3 kDa Proteins beschrieben. Im Gegensatz dazu ist im zweiten Teil das bakterielle DNA-Reparatur-Enzym Mth0212 dargestellt. Der Fokus liegt auf der detaillierten Strukturanalyse des Proteins in seiner Apo-Form sowie im Komplex mit verschiedenen DNA-Substraten und deren Vergleich mit homologen Enzymen. Auf ein kristallographisches Problem – eine sog. “Verzwilligung” – wird nur kurz eingegangen.

Die vorliegende Doktorarbeit führte zu den nachfolgend aufgeführten Manuskripten dreier Publikationen. Davon ist eine veröffentlicht (1), eine weitere zur Veröffentlichung eingereicht (2) und eine dritte verblieben in Überarbeitung. Der veröffentlichte Artikel (1) und das eingereichte Manuskript (2) sind in Abschnitt I enthalten. Teil II dagegen befasst sich mit den Ergebnissen, die in Manuskript (3) veröffentlicht werden sollen.

1. K. Lakomek, A. Dickmanns, U. Mueller, K. Kollmann, F. Deuschl, A. Berndt, T. Luebke and R. Ficner* (2009) *De novo sulfur SAD phasing of the lysosomal 66.3 kDa protein from mouse*, Acta Cryst. D65, 220-228. (* corresponding author)
2. Lakomek K., Dickmanns A., Kettwig M., Ficner R.*, Luebke T., *Initial insight into the function of the lysosomal 66.3 kDa protein from mouse by means of X-ray crystallography*, submitted (* corresponding author)
3. Lakomek K., Dickmanns A., Ciirdaeva E., Schomacher L., Fritz H.-J., Ficner R.*, *3`-5` Exo Competes with 2`-Deoxyuridine Endonuclease Function in Mth0212-DNA Complex Structures*, in preparation (* corresponding author)

I) DE NOVO STRUKTURAUFKLÄRUNG UND -ANALYSE DES LYSOSOMALEN 66.3 KDA PROTEINS AUS MAUS

In aktuellen Sub-Proteomanalysen, die auf Mannose-6-Phosphat-Reste tragende Proteine gerichtet waren, wurde das 66.3 kDa Protein als neues lösliches Protein der lysosomalen Matrix identifiziert. Das 66.3 kDa Protein aus Maus und sein menschliches Ortholog p76 wurden anschließend ausführlicher charakterisiert. Das Mausprotein wird als glykosyliertes 75 kDa Präproprotein synthetisiert und in ein 28 kDa und ein 40 kDa Fragment prozessiert. Trotz bioinformatischer Analysen und molekularer Charakterisierung blieben sowohl die Art des Reifungsprozesses als auch die physiologische Funktion bislang unbekannt.

Um diese Fragestellung zu klären, wurde das 66.3 kDa Protein kristallisiert und seine Struktur mittels eines Schwefel-SAD (S-SAD) - Experiments aufgeklärt. Die Expression erfolgte als C-terminal mit einer Histidin-Affinitätssequenz markierte Variante in einer menschlichen Fibrosarkom-Zelllinie. Anschließend wurden die aus einer Polypeptidkette bestehende 66.3 kDa – Form und die Zweikettenvariante, die aus dem 28 kDa und dem 40 kDa - Fragment zusammengesetzt war, bis zur Homogenität gereinigt, konnten jedoch nicht voneinander getrennt werden. Deswegen wurde dieses Gemisch zur Kristallisation eingesetzt und führte zur Entstehung einzeln wachsender Kristalle. Sie gehörten der monoklinien Raumgruppe C2 an. Das Protein, das für die in der vorliegenden Arbeit dargestellten Studien eingesetzt wurde, wurde in dem beschriebenen Reinheitsgrad von Prof. Dr. Torben Lübke (Georg-August-Universität Göttingen) zur Verfügung gestellt.

Die Struktur wurde mittels S-SAD-Phasierung unter Verwendung von Daten mit einem R_{anom}/ $R_{p.i.m.}$ – Verhältnis von 1,1 aufgeklärt. In der Regel wurde ein R_{anom} / $R_{p.i.m.}$ – Verhältnis von 1,5 als notwendig erachtet, erst in letzter Zeit wurden auch einige wenige Strukturen mit Werten von 1,1 oder 1,2 erfolgreich gelöst. Die verfeinerte Substruktur der anomal streuenden Atome enthielt 21 intrinsische Schwefelatome sowie ein Xenonatom mit geringem Besetzungsgrad, das während einer Inkubation mit Xenongas in einer hydrophoben Tasche des Proteins festgehalten worden war. Der Beitrag des Xenonatoms zum anomalen Signal wurde im Vergleich zu demjenigen der Schwefelatome analysiert und war zu vernachlässigen. Somit ist die Struktur des 66.3 kDa Proteins eine der größten bisher mit Hilfe eines S-SAD-Experiments bestimmten Struktur und eine von nur wenigen erfolgreichen S-SAD-Aufklärungen in einer monoklinen Raumgruppe. Im Rahmen der Versuche zur Lösung des kristallographischen Phasenproblems wurden von weiteren Kristallen Datensätze aufgenommen. Sie stellten sich als interessant heraus, da sie unterschiedliche Stadien des Reifungsprozesses des 66.3 kDa Proteins aufzeigen. Die Struktur des 66.3 kDa Proteins wurde zu einer maximalen Auflösung von 1.8 Å verfeinert. Sie zeigt, dass die infolge eines proteolytischen Schnitts entstehenden Fragmente miteinander assoziiert bleiben.

Die Kristallstrukturen weisen eine signifikante Ähnlichkeit des 66.3 kDa Proteins zu mehreren bakteriellen Hydrolasen auf. Die zentrale geschichtete αββα - Anordnung sowie ein N-terminaler Cysteinrest des 40 kDa - Fragments (Cys249) ordnet das 66.3 kDa Protein der strukturell definierten Superfamilie der N-terminalen Nukleophil (Ntn) – Hydrolasen zu. Die Ähnlichkeit zu den bakteriellen Enzymen legt sowohl eine hydrolytische Aktivität gegenüber nicht-peptidischen Amidbindungen als auch einen autokatalytischen Schritt im Verlauf des Reifungsprozesses des 66.3 kDa Proteins nahe. Infolge der Spaltung der Peptidbindung zwischen den Aminosäureresten Serin 248 und Cystein 249 wird eine tiefe Tasche für potentielle Substrate zugänglich, an deren Grund sich das fakultative aktive Zentrum des 66.3 kDa Proteins befindet. Folglich scheint die gezielte Aktivierung des 66.3 kDa Proteins über einen autoproteolytischen Mechanismus zu erfolgen.

II) STRUKTURANALYSE DES EXOIII - HOMOLOGS MTH0212 ALLEIN UND IM KOMPLEX MIT VERSCHIEDENEN SUBSTRAT-DNAS

Das Exonuklease III - Homolog des thermophilen Archaeons *Methanothermobacter thermoautotrophicus*, Mth0212, weist eine einmalige Kombination von Aktivitäten zur DNA-Reparatur auf. Zusätzlich zur klassischen 3`-5`-Exonuklease-Aktivität ist Mth0212 in der Lage, doppelsträngige DNA auf der 5'-Seite eines 2'-Desoxyuridinrests zu schneiden. Das Vorkommen von 2'-Desoxyuridin in DNA ist eine häufige Schadensart der DNA, und aufgrund seiner promutagenen Eigenschaft ist eine zuverlässige Reparatur zur korrekten Aufrechterhaltung der genetischen Information unerlässlich.

Gewöhnlich wird 2'-Desoxyuridin durch die aufeinander folgende Aktivität einer Uracil-DNA-Glykosylase (UDG) und einer AP-Endonuklease entfernt. UDGs schneiden die Base Uracil heraus und erzeugen so einen Nukleotidrest ohne Pyrimidin– oder Purinbase, eine basenlose Stelle (engl.: abasic site = apyrimidinic / apurinic site = AP site). Bislang sind nur wenige Organismen bekannt, denen ein Homolog der UDG-Superfamilie fehlt - unter ihnen *M. thermoautotrophicus*.

Vor kurzem wurde gezeigt, dass beide oben aufgeführten initialen Schritte der 2`-Desoxyuridin-Reparatur (UDG, AP-Endonuklease) von Mth0212 übernommen werden, jedoch in einem einzigen katalytischen Prozess stattfinden. Die Nuklease schneidet das Phosphodiester-Rückgrat direkt und vermeidet auf diese Weise die Entstehung einer abasischen Stelle, die toxischer für die Zelle ist als der ursprüngliche DNA-Schaden.

Um einen tieferen Einblick in den Mechanismus der 2`-Desoxyuridin-Erkennung zu erlangen und um zu verstehen, wie die verschiedenen von Mth0212 katalysierten nukleolytischen Aktivitäten in einem einzigen aktiven Zentrum erfolgen können, wurden Kristallstrukturen des Enzyms in seiner Apo-Form sowie im Komplex mit verschiedenen Substrat-DNA-Molekülen analysiert.

Dazu wurden sowohl Wildtyp-Protein als auch nach rationalen Überlegungen entworfene Mutanten zur Co-Kristallisation mit DNA-Oligonukleotiden unterschiedlicher Kettenlänge, Sequenz und mit variablen Arten an 5`- und 3`-Enden verwendet. Die Protein-DNA-Komplexe unterscheiden sich in ihrer Zusammensetzung und der relativen Orientierung der Makromoleküle zueinander. Sie zeigen intermolekulare Kontakte über die gesamte Interaktionsoberfläche der Nuklease. Die Strukturen wurden mittels „Molekularem Ersatz" gelöst und zu einer maximalen Auflösung im Bereich von 1.2 bis 3.1 Å verfeinert.

Die fünf Apo- und neun Komplex-Strukturen wurden sowohl untereinander als auch mit homologen Enzymen verglichen, die nur die biochemischen Hauptmerkmale dieser Nuklease-Familie aufweisen. Obwohl alle in dieser Arbeit dargestellten Komplexstrukturen Mth0212 in seiner namensgebenden exonukleolytischen Funktion beschreiben und keinen direkten Einblick in den Mechanismus der zusätzlichen Uridin-Endonuklease-Aktivität bieten, zeigen sie eine mögliche Erklärung für die einmalige Kombination der Aktivitäten zur DNA-Reparatur auf. Höchstwahrscheinlich führen winzige strukturelle Unterschiede in den drei spezifischen DNA bindenden Schleifenstrukturen zu dem erweiterten Substratspektrum von Mth0212. Die Einfügung einer Argininseitenkette in die Basenstapelung der DNA-Doppelhelix (Arg209), die in menschlicher UDG beobachteten Interaktionen ähnelt, zusammen mit einem Lysin, einem Serin und zwei Asparaginresten in der Substrat-Bindetasche (Lys125, Ser171, Asn114, Asn153) haben vermutlich Schlüsselfunktionen bei der Erkennung eines 2`-Desoxyuridinrests.

X) 2. SUMMARY

This PhD thesis is divided into two parts both dealing with structural characterization of macromolecules by means of X-ray crystallography. Part I concerns the lysosomal 66.3 kDa protein from mouse. It describes the expanded use of available crystallographic methods for structure determination and the analysis of the structure representing the protein of so far unknown function. In contrast, in part II the archaeal DNA repair enzyme Mth0212 is investigated. The crystallographic problem of twinning is only touched and the main focus lies on the detailed structural analysis of the protein alone as well as in complex with different substrate DNAs and subsequent comparison with homologous enzymes.

This PhD thesis resulted in the following manuscripts of three publications of different status: published (1), submitted (2), in preparation (3). The published article (1) and the submitted manuscripts (2) are dealt with in Part I, whereas Part II corresponds to the manuscript in preparation (3).

1. K. Lakomek, A. Dickmanns, U. Mueller, K. Kollmann, F. Deuschl, A. Berndt, T. Luebke and R. Ficner* (2009) *De novo sulfur SAD phasing of the lysosomal 66.3 kDa protein from mouse*, Acta Cryst. D65, 220-228. (* corresponding author)
2. Lakomek K., Dickmanns A., Kettwig M., Ficner R.*, Luebke T., *Initial insight into the function of the lysosomal 66.3 kDa protein from mouse by means of X-ray crystallography*, submitted (* corresponding author)
3. Lakomek K., Dickmanns A., Ciirdaeva E., Schomacher L., Fritz H.-J., Ficner R.*, *3`-5` Exo Competes with 2`-Deoxyuridine Endonuclease Function in Mth0212-DNA Complex Structures*, in preparation (* corresponding author)

I) DE NOVO STRUCTURE SOLUTION AND ANALYSIS OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE

Recently, sub-proteome studies dealing with mannose 6-phosphate containing proteins revealed the 66.3 kDa protein as a novel soluble protein of the lysosomal matrix. Subsequently, the murine 66.3kDa protein and its human orthologue p76 were characterized in more detail. The mouse orthologue has been shown to be synthesized as a glycosylated 75 kDa preproprotein, which is processed into 28 kDa and 40 kDa fragments. Despite bioinformatics approaches and molecular characterization, the mode of maturation as well as the physiological function of the 66.3 kDa protein have so far remained unknown. In order to tackle these questions, the 66.3 kDa protein was crystallized. The structure determination process by means of sulfur SAD phasing is presented in the following.

After expression in a human fibrosarcoma cell line, the C-terminally His-tagged one chain 66.3 kDa variant and the double chain form consisting of a 28 kDa and a 40 kDa fragment were purified to homogeneity, but could not be separated during the purification procedure. Thus this mixture was used for crystallization, single crystals were obtained. They belong to the monoclinic space group C2. The structure was solved by means of sulfur SAD phasing using data with an R_{anom} / $R_{p.i.m.}$ ratio of 1.1. The refined substructure of anomalous scatterers turned out to comprise twenty-one intrinsic sulfur atoms and one xenon atom with a very low occupancy which had been caught in a hydrophobic pocket during a xenon soak. The contribution of the single xenon atom to the anomalous signal was analyzed in comparison to that of the sulfur atoms and found to be negligible. Thus, its structure is one of the largest solved by sulfur SAD (S-SAD) phasing so far and one of a few successful S-SAD phase determinations using crystals of a monoclinic space group. In the course of solving the crystallographic phase problem, additional data sets were collected which turned out to be of interest as they represent different states of the maturation process of the 66.3 kDa protein. The structure was refined to a maximum resolution limit of 1.8 Å. The structures demonstrate that the fragments of the proteolytic cleavage process stay associated. The crystal structures reveal a significant similarity of the 66.3 kDa protein to several bacterial hydrolases. The core αββα sandwich fold and a cysteine residue at the N-terminus of the 40 kDa fragment (C249) classify the 66.3 kDa protein as a member of the structurally defined N-terminal nucleophile (Ntn) hydrolase superfamily suggesting a hydrolytic activity on non-peptide amide bonds. The similarity to these bacterial proteins also implies an autocatalytic maturation of the lysosomal 66.3 kDa protein. Upon cleavage between serine 248 and cysteine 249, a deep pocket becomes solvent accessible which harbors the putative active site of the 66.3 kDa protein.

II) STRUCTURAL ANALYSIS OF THE EXOIII HOMOLOGUE MTH0212 ALONE AND IN COMPLEX WITH DIFFERENT SUBSTRATE DNAS

The Exonuclease III homologue of the thermophilic archaeon *Methanothermobacter thermoautotrophicus,* Mth0212, displays a unique combination of DNA repair activities. In addition to a 3`-5`exonuclease activity, it is capable of nicking double-stranded DNA at the 5'-side of a 2'-deoxyuridine residue. The occurrence of 2'-deoxyuridine in DNA is a frequent kind of DNA damage, and due to its pre-mutagenic character a reliable repair is crucial for the correct maintenance of the genomic information. Commonly, 2'-deoxyuridine is removed by the consecutive action of a uracil DNA glycosylase (UDG) creating an apyrimidinic/apurinic site (AP site) and of an AP endonuclease. So far, only a few organisms are known to lack a homologue of the UDG superfamily - among them *M. thermoautotrophicus*. Recently it was shown that both initial steps of 2`-deoxyuridine repair are taken over by Mth0212, but catalyzed in a single step. The nuclease directly cuts the phosphodiester backbone avoiding the emergence of an AP site which is even more toxic than the original base damage. In order to get a deeper insight into the recognition of 2`-deoxyuridine and to understand how the different nucleolytic activities of Mth0212 can be accomplished in a single active site, crystal structures of the nuclease alone as well as in complex with different substrate DNAs were analyzed. The wild-type and rationally designed mutants were used for co-crystallization with DNA oligonucleotides varying in length, sequence and kind of 5`- and 3`- ends. The protein-DNA complexes differ in their composition and in the relative orientation of the macromolecules to each other revealing Mth0212-DNA contacts across the whole interaction surface of the enzyme. The structures could be solved by means of 'Molecular Replacement' and were refined to a resolution ranging from 1.2 to 3.1 Å.

Although all complex structures represent Mth0212 in its eponymous exonucleolytic function and give no direct insight into the mechanism of the additional uridine endonuclease activity, the comparison of the five apo and nine complex structures with each other and with homologous enzymes indicated putative explanations for the unique combination of DNA repair activities. Most likely tiny structural differences result in the expanded substrate spectrum of Mth0212 compared with ExoIII homologues only exhibiting the biochemical hallmarks of this nuclease family. The insertion of an arginine side chain (Arg209) into the DNA helical base stack which resembles interactions observed in human uracil DNA glycosylase in concert with a lysine, a serine and two asparagine residues in the substrate binding pocket (Lys125, Ser171, Asn114, Asn153) are supposed to play key roles in 2`-deoxyuridine recognition.

X) 3. DANKSAGUNGEN

Die vorliegende Arbeit wurde von September 2005 bis März 2009 in der Abteilung für Molekulare Strukturbiologie an der Georg-August-Universität Göttingen angefertigt. Nette Menschen haben mich während dieser Zeit unterstützt und die vorliegende Arbeit in dieser Form erst ermöglicht.

Herzlich bedanken möchte ich mich bei

Herrn Prof. Dr. Ralf Ficner	für die Möglichkeit an diesen sehr interessanten Fragestellungen arbeiten zu können, die Anleitung zu selbständigem wissenschaftlichen Arbeiten und eine exzellente Arbeitsumgebung. Ohne seinen Optimismus, seine kontinuierlich freundliche und hervorragende Unterstützung und sein stetes Interesse am positiven Fortgang dieser Arbeit wäre sie in dieser Form nicht möglich gewesen.
Herrn Dr. Achim Dickmanns	für seine außerordentliche Hilfe bei methodischen Fragen und viele wertvolle Hinweise, die Durchsicht der Manuskripte sowie dieser Arbeit und insbesondere für seinen unerschöpflichen Enthusiasmus
Herrn Prof. Dr. Kai Tittmann	für seine sofortige Bereitschaft zur Übernahme des Korreferats. Vielen Dank!
Herrn Prof. Dr. Hans-Joachim Fritz	sowie den ehemaligen und derzeitigen Doktoranden in seiner Abteilung: Lars Schomacher, Swetlana Ber und Elena Ciirdaeva für die sehr gute und nette Zusammenarbeit bezüglich des Mth0212-Projekts mit vielen hilfreichen Diskussionen
Herrn Prof. Dr. Torben Lübke	sowie den ehemaligen und derzeitigen Doktoranden in seiner Abteilung: Katrin Kollmann, Florian Deuschl und Matthias Kettwig für die aufgeschlossene und erfolgreiche Kooperation im Rahmen des 66.3 kDa Protein - Projekts
Prof. Dr. Oliver Einsle	für Hilfe bei kristallographischen und Computer-Problemen. An dieser Stelle möchte ich auch besonders Herrn Dr. Markus Rudolph, Herrn Dr. Tim Grüne, Frau Dr. Regine Herbst-Irmer und Prof. Dr. George Sheldrick und allen Betreuern an den Synchrotons danken.
ALLEN ehemaligen und jetzigen Mitarbeitern der Abteilung Molekulare Strukturbiologie	für die hervorragende Arbeitsatmosphäre, hilfreiche Diskussionen sowie diverse „Synchrotontrips“ und …... Ein besonderer Dank gilt denjenigen, die mir von Anfang an stets hilfreich zur Seite standen, allseits gute Laune verbreitet haben und einen außerordentlichen Anteil dazu beigetragen haben, die letzten Jahre zu einer sehr angenehmen Zeit zu machen: Annette Berndt, Jens Brinkmann, Chrissoula Greulich, Dr. Maren Hoffmann, Sohail Khoshnevis, Angela Kopp, Dr. Denis Kudlinzki, Winfried Lendeckel, Peer Lukat, Dr. Sina Möhlmann, Thomas Monecke, Anja Pomowski, Dr. Carmen Rotte, Nadine Schmidt, Dr. Jana Schmitzova, Eike Schulz, Teresa Senge, Dr. Julia Wittmann, Dr. Daniel Wohlwend.
Kollegen, Freunden	, die hier eigentlich noch gesondert genannt werden müssten
meiner Familie	für die fortwährende Unterstützung in jeder Hinsicht, insbesondere meinen Eltern und meinem Bruder sowie meiner Großmutter und unserem Hund Skipper, die mich lange Zeit begleitet haben. Dankeschön!

PART I

STRUCTURAL CHARACTERIZATION OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE BY MEANS OF X-RAY CRYSTALLOGRAPHY

1. INTRODUCTION

1.1. LYSOSOMAL PROTEINS

1.1.1. The lysosomal compartment

A characteristic feature of eukaryotic cells is their compartmentalization into membrane-bordered organelles such as the Golgi apparatus, the endoplasmic reticulum and lysosomes resulting in a spatial separation of the divergent reactions that occur in the cell. Lysosomes derive from the Golgi apparatus as small vesicles and develop via early and late endosomes. They were shown to contain a set of about 60 hydrolases and associated proteins which are essential for the cell as reflected by the manifestation of severe diseases in the absence of the enzyme activities.

1.1.2. Functions of lysosomal proteins

Most lysosomal proteins are responsible for the degradation of macromolecules or even whole organelles, which are derived from diverse sources including bacterial cells, virus particles, complexes and single molecules (Fig. I-1). The pathway by which the substances enter the lysosomes depends on their origin. While intracellular substrates are ingested by autophagy via specific vacuoles (de Duve & Wattiaux, 1966), extracellular material is received either by receptor- and clathrin-coated vesicle mediated endocytosis (de Duve, 1983; Sleat et al., 2007; Sleat et al., 2008; reviewed in Lübke et al., 2009), by pinocytosis in which cytosolic droplets with extracellular fluid are nonspecifically engulfed or by phagocytotic pathways, which involve the formation of a phagosome and have been predominantly observed in macrophages and granulocytes in the course of the unspecific immune defense (Haas, 2007).

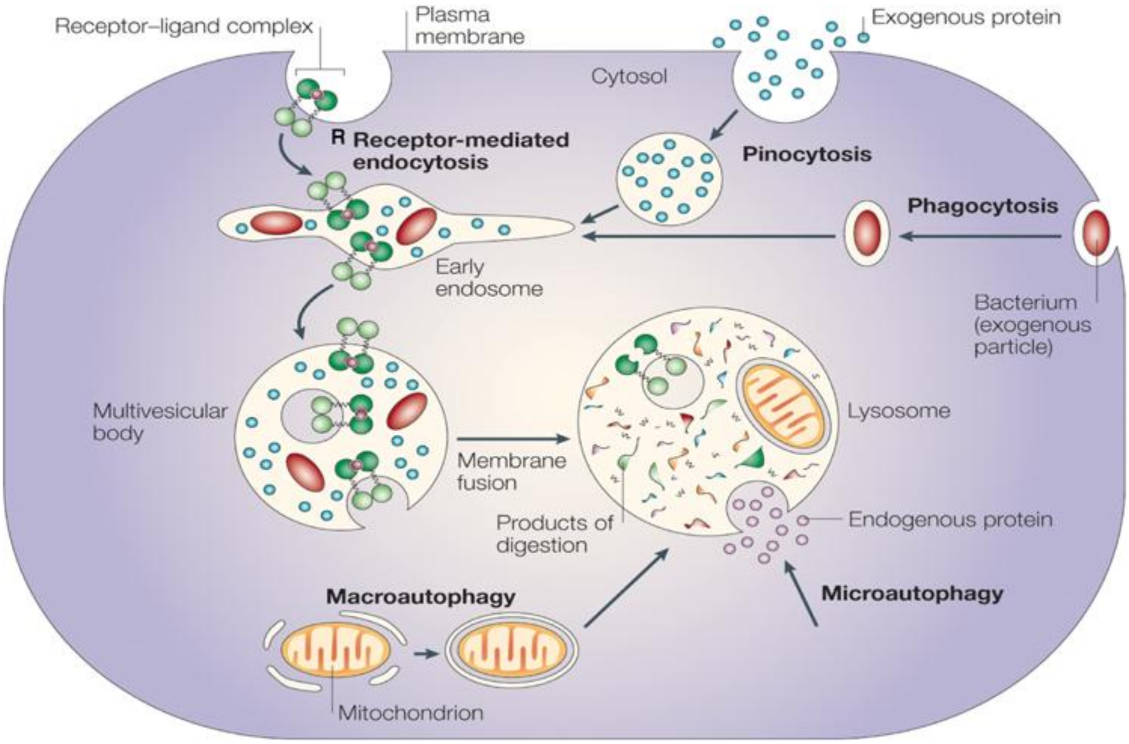


Fig. I-1. Overview of the digestive processes mediated by lysosomes: endocytosis, pinocytosis, phagocytosis and autophagy (Ciechanover, 2005).

The majority of the acid hydrolases required for degradation of ingested molecules / material are soluble and located in the lumen which is referred to as lysosomal matrix (de Duve, 1969; Kornfeld & Mellman, 1989). Their pH optimum amounts to about pH 4.8 reflecting the acidic character of the lysosomal compartment (Ohkuma & Poole, 1978). According to their function, these enzymes are classified as phosphatases, sulfatases, glycosidases, nucleases, proteases, lipases or phospholipases.

Typically the lack of function of lysosomal proteins causes severe pathogenic phenotypes termed "lysosomal storage diseases" since they are associated with the accumulation of undigested molecules in the lysosomal compartment (reviewed in Scriver et al., 2001). Pathophysiological processes are not only caused by defects in the degradation pathways, which regarding lysosomes ubiquitously come to mind, but also by dysfunction of proteins involved in more recently discovered biosynthetic-secretory pathways. Thereby, the lack of active enzymes which are crucial for neuroprotection (Cravatt et al., 2001) as well as for the modulation of hormones and bioactive lipids implicated e.g. in tissue homeostasis and inflammation (Capasso et al., 2001; Izzo et al., 2001; Feulner et al., 2004, reviewed in Hansen et al., 2000) have been related to the development of Alzheimer disease (Nixon & Cataldo, 2006), anorexia, tumor metastasis and propagation (Fehrenbacher & Jaattela, 2005; Garcia et al., 1996; reviewed in Kos & Lah, 1998).

1.1.3. Transport of lysosomal proteins to the lysosome accompanied by co- and post-translational modifications

Lysosomal proteins are commonly synthesized into the lumen of the rough endoplasmic reticulum (rER). At the lumenal side of the membrane, a precursor oligosaccharide is attached as a single entity to asparagine residues by the membrane-bound oligosaccharyl transferase. The N-linked oligosaccharide consists of 14 sugar moieties and is in most cases transferred co-translationally. It is trimmed in the ER prior to the passage of the newly synthesized protein into the Golgi apparatus. In the *cis* Golgi network, further additions and modifications of sugars can occur, e.g. mannose 6-phosphate (M6P) residues are attached exclusively to N-linked oligosaccharides and therein only to selected mannose moieties. Based on the final composition of the oligosaccharide, N-glycans can be subdivided in three types, namely the "high mannose", the "hybrid" or the "complex" type. All of them exhibit a common pentasaccharide core consisting of two N-acetylglucosamine (NAG) and three M6P moieties. The latter are recognized by two M6P receptors (MPRs) as a sorting signal for the transport in clathrin-coated vesicles budding from the *trans* Golgi network (Fig. I-2). The vesicles fuse with late endosomes, which exhibit a slightly acidic interior (pH ~ 6) leading to the dissociation of the MPRs from the transported protein as well as to the release of phosphate from the M6P residues. Subsequent gradual maturation of late endosomes results in the development of lysosomes with a pH of 4-5 and a specific set of proteins. Thus, most lysosomal enzymes are directed to lysosomes by M6P residues, which have been post-translationally generated at the N-glycans on the proteins` surface. Additionally, the oligosaccharides e.g. serve as markers during protein folding and make the decorated proteins more stable due to the reduced access of proteases to the protein.

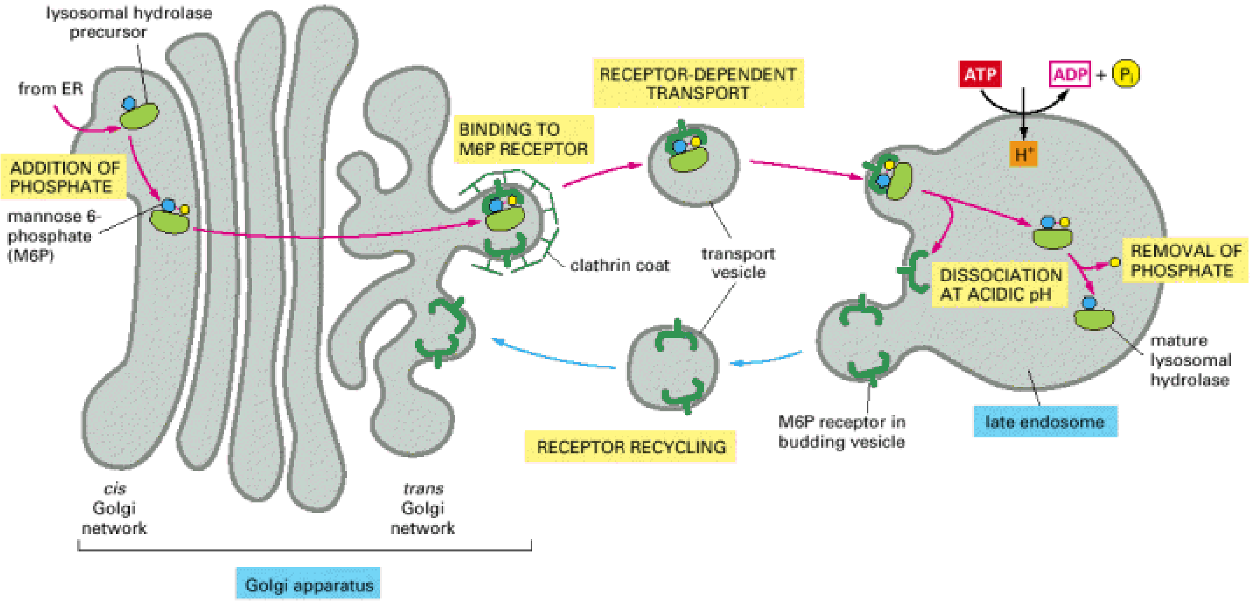


Fig. I-2. The transport of newly synthesized lysosomal proteins to lysosomes mediated by a mannose-6-phosphate (M6P) receptor (Alberts et al., 2002).

1.1.4. Identification of novel lysosomal proteins in proteomics approaches

In order to reveal novel lysosomal proteins and thus novel lysosomal functions, several sub-proteomic studies dealing with soluble lysosomal proteins have been carried out recently (Sleat, Wang et al., 2006; Sleat, Zheng et al.; 2006; Kollmann et al.; 2005; Sleat et al., 2008); reviewed in Sleat et al., 2007). They made use of M6P as a characteristic feature of most lysosomal proteins as follows. Proteins are expressed in cell lines which are deficient in main MPRs so that they are not correctly targeted to lysosomes, but channeled into the alternative vesicular transport pathway, exocytosis, and thus secreted into the medium. Subsequent purification therefore starts with ammonium sulfate precipitation and finally results in concentrated cell extracts. These enriched fractions are subjected to affinity chromatography using a column on which a MPR mixture is attached to the base material. After successive washing steps also including glucose-6-phosphate in order to remove unspecifically bound molecules, proteins are eluted by the addition of free mannose-6-phosphate. They are separated by means of 2-D SDS gel electrophoresis and subsequently analyzed using peptide mass fingerprints in concert with Edman sequencing and bioinformatics. If comparison with identical sequences beyond known lysosomal proteins is of negative outcome, further experiments have to be carried out to unambiguously show the lysosomal localization. For reference, several marker proteins of the lysosomal matrix are monitored throughout the procedure.

In several lysosomal proteomic studies from mouse, rat and human, beyond others, the 66.3 kDa protein was identified as a putative soluble lysosomal protein (Lübke et al., 2009).

1.2. THE 66.3 KDA PROTEIN:

Bioinformatic analysis and molecular characterization

The 66.3 kDa protein is conserved among vertebrates (Fig. I-3). The sequence identity between the 66.3 kDa protein and some orthologues is described in detail in the submitted manuscript (chapter 4).

Following their identification in proteomics approaches, the murine 66.3 kDa protein and its human orthologue p76 were characterized in more detail regarding their lysosomal localization, processing and glycosylation (Deuschl et al., 2006; Jensen et al., 2007). The 66.3 kDa protein is synthesized at the rER as a glycosylated preproprotein with an apparent molecular mass of 75 kDa. After the co-translational removal of the N-terminal signal peptide, the respective proprotein is sorted to the lysosomal compartment and further processed into a 28 kDa N-terminal and a 40 kDa C-terminal fragment (Deuschl et al., 2006) (Fig. I-4).

The 40 kDa fragment might be further processed (Deuschl et al., 2006) (Fig. I-5a) into a 25 kDa N- and a 15 kDa C-terminal fragment (unpublished data T.L., by Edman digest). A similar processing was described for the human orthologue p76 resulting in a 32 kDa N-terminal fragment and a 45 kDa C-terminal fragment (Jensen et al., 2007). The authors suggested an additional maturation step for the 40 kDa fragment from mouse as well due to the detection of a C-terminal 27 kDa fragment of non-secreted protein by Western analysis.

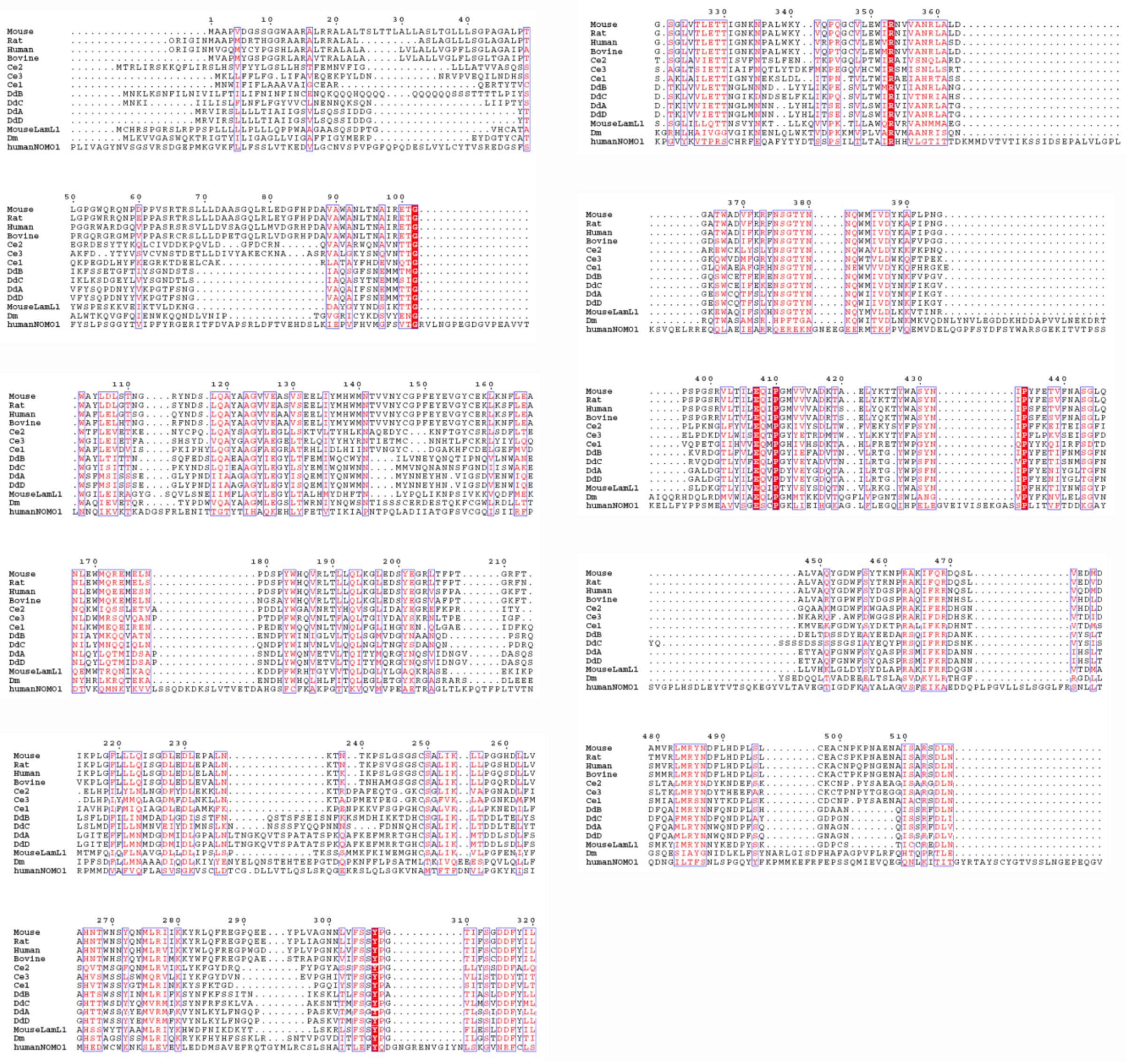


Fig. I-3. Alignment of the amino acid sequences of the 66.3 kDa protein and its homologues using CLUSTALW v. 2.04 (Chenna et al., 2003) and ESPRIPT (Gouet et al., 1999). Ce = *Caenorhabditis elegans*, Dd = *Dictyostelium discoideum*, Dm = *Drosophila melanogaster*.

Such limited proteolysis during the maturation procedure has been observed for many lysosomal proteins and commonly leads to their activation (Hasilik, 1992).

For purification, a C-terminally RGS-His$_6$-tagged derivative of the mouse 66.3 kDa protein was expressed in the human fibrosarcoma cell line HT1080 and due to secretion into the medium first subjected to ammonium sulfate precipitation. Subsequent purification steps included Ni$^+$-NTA affinity and HPLC anion exchange chromatography. Limited proteolysis shed light on the post-translational maturation and revealed Cys249 as the N-terminal residue of the 40 kDa fragment.

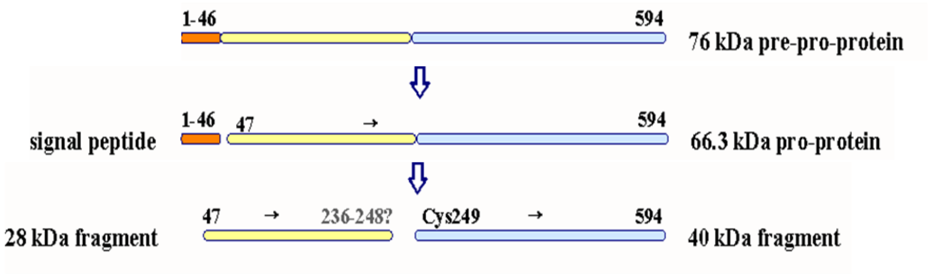


Fig. I-4. Processing of the lysosomal 66.3 kDa protein from mouse.

By means of peptide:N-glycosidase F (PNGase) all five potential N-glycosylation sites of the mouse orthologue have been shown to be used upon expression (Figs. I-5b, c).

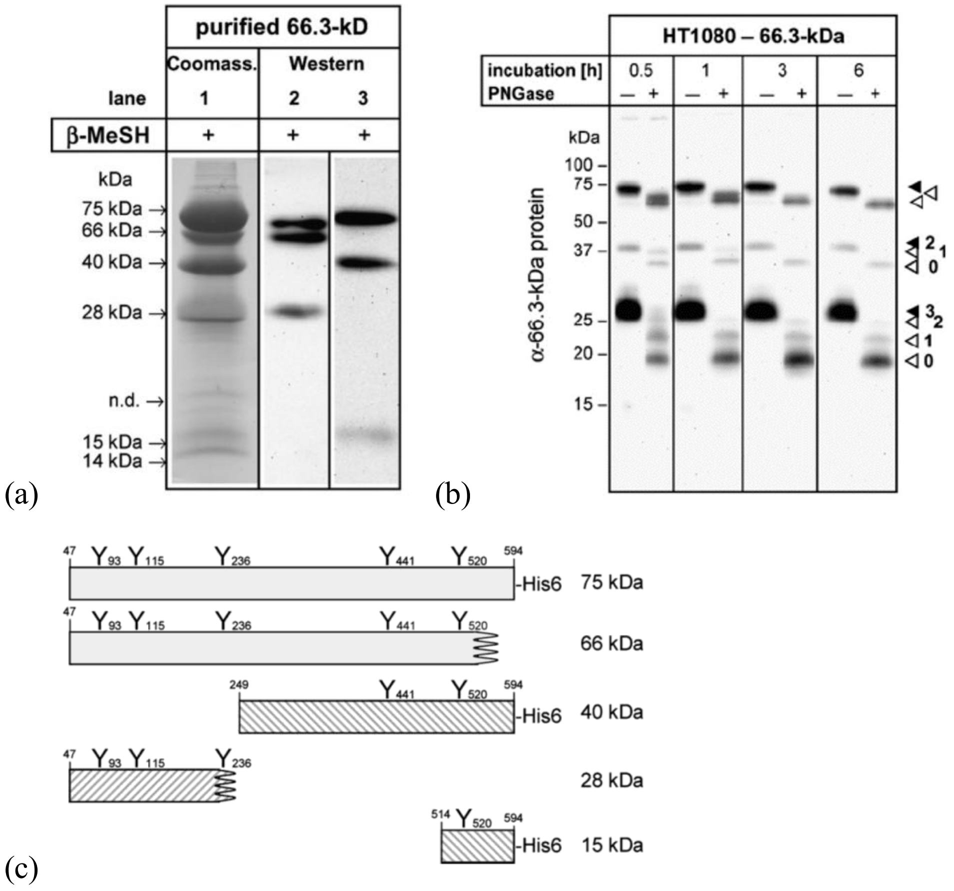


Fig. I-5. Molecular forms of the 66.3 kDa protein from mouse (Deuschl et al., 2006). (a) SDS-PAGE and Western blot analysis of the purified 66.3 kDa protein under reducing conditions. Coomass. = SDS polyacrylamide gel stained with Coomassie Brilliant Blue, β-MeSH = β-mercaptoethanol. lane 2: 66.3 kDa protein antiserum, lane 3: monoclonal antibody against the C-terminal His_6-tag. (b) HT1080 cells stably expressing the 66.3 kDa protein after treatment with PNGase analyzed by Western blotting using the 66.3 kDa protein antiserum. Filled / open arrowheads: glycosylated / (partially) deglycosylated forms with the number of their N-glycans indicated on the right. (c) Schematic representation of the polypeptides of the purified 66.3 kDa protein and their N-glycosylation sites.

However, since neither bioinformatics analysis nor the detailed molecular characterization of the mouse lysosomal 66.3 kDa protein and its human orthologue p76 have provided any hint regarding the activity and the physiological function, the problem was approached by the determination of the three-dimensional structure of the mouse 66.3 kDa protein. Due to the fact, that no structure of a protein with sufficient similarity of the amino acid sequence level has been available, the phases were determined experimentally (chapter 3). Based on subsequent analysis as well as database and literature search, the 66.3 kDa protein belongs to the superfamily of N-terminal nucleophile (Ntn) hydrolases. Despite the lack of a significant sequence similarity there is a close resemblance to several bacterial hydrolases regarding the protein fold and active site residues providing initial insight into its catalytic activity as well as putative substrates. In concert with structures of the 66.3 kDa protein yielded from different purification batches the structural homology to characterized Ntn hydrolases suggested a mechanism of the enzyme`s activation involving autocatalytic proteolysis.

1.3. X-RAY CRYSTALLOGRAPHY

1.3.1. Crystallographic methods

The complete characterization of a specific macromolecule includes the determination of its three-dimensional structure. For this purpose, several methods are available. They differ in their feasibility for the macromolecules of different sizes and yield information to variable resolution limits. Single particle electron microscopy (EM) e.g. is suitable only for molecules and complexes with a molecular weight (MW) of more than 200 kDa and provides information to a resolution of < 2 Å. While EM requires a frozen sample, for the recently developing Small Angle X-ray Scattering (SAXS) complex solutions can be used. Application and results are similar with regard to the macromolecular size and resolution. The molecular weight of a compound considered favorable for a *de novo* structure determination in solution at atomic resolution by means of Nuclear Magnetic Resonance (NMR) spectroscopy lies in the range of 30-40 kDa. Thus, the commonly used method for studies of single macromolecules or protein-protein complexes as well as complexes between proteins and nucleic acids with a MW of more than 40 kDa is X-ray crystallography. Since atomic resolution can be achieved, it provides detailed information about intra- and intermolecular contacts such as hydrogen-bonding interactions and residue conformations.

For the 66.3 kDa protein for example, X-ray crystallography was used as an alternative strategy to bioinformatics in concert with its molecular characterization to gain initial insight into its function and indeed enabled the assignment to an enzyme superfamily with a common substrate class and similar activation and catalytic mechanisms.

A prerequisite for the determination of a crystal structure is an initial model providing the so-called starting phases. Together with the amplitudes which are directly derived from measured data they define the structure. If with regard to a protein structure determination a model of a structurally similar protein is available, this can serve to calculate initial phases. If this does not apply, experimental phases have to be determined. Commonly, either the protein is derivatized with selenomethionine replacing methionine residues or native protein crystals are subjected to soaks with different heavy atom compounds or halide salts. However, the first strategy does not always yield soluble protein or exhibits an incomplete labeling - especially observed for expression in cell cultures - and can fail in crystallization. Soaking procedures can break the used crystals and thus make them unusable for X-ray diffraction experiments. In such cases, in particular applying for proteins that require both, expensive and time-consuming expression and purification procedures which exhibit an increased challenge for protein labeling and prevent extensive screening of heavy atom / halide crystal soaks such as for the 66.3 kDa protein, "Sulfur Single Anomalous Dispersion" (S-SAD) phasing can be used as an alternative strategy to derive initial phases.

S-SAD makes use of the weak anomalous scattering of intrinsic S atoms in unlabeled proteins. Since this method is currently developing and due to recent achievements it might become routine in the next future and encouraging for standard protein structure determinations.

1.3.2. Sulfur SAD phasing

Currently, several high brilliance third generation synchrotron beamlines support the energy range required for long wavelength phasing applications (about 1.7-2.5 Å) (Djinovic-Carugo et al., 2005). Thus, the number of SAD phasing experiments using only weak anomalous scatterers has increased (Ramagopal et al., 2003). Most sulfur SAD data have been collected at tunable synchroton beamlines in order to make use of the appreciably larger $f''(S)$ (Fig. I-6) (Brown et al., 2002; Gordon et al., 2001; Liu et al., 2000; Ramagopal et al., 2003; Weiss, 2001; Weiss et al., 2004), although Cu Kα (1.54 Å) is suitable for sulfur SAD phasing as well (Dauter et al., 1999; Sekar et al., 2004; Roeser et al., 2005). Often successful phasing requires the additional presence of further weak anomalous scatterers such as Ca^{2+} and Cl^{-} ions which commonly are refined to significantly higher occupancies than the sulfur sites (Dauter et al., 1999; Yang & Pflugrath, 2001; Debreczeni, Bunkoczi, Ma et al., 2003, Roeser et al., 2005).

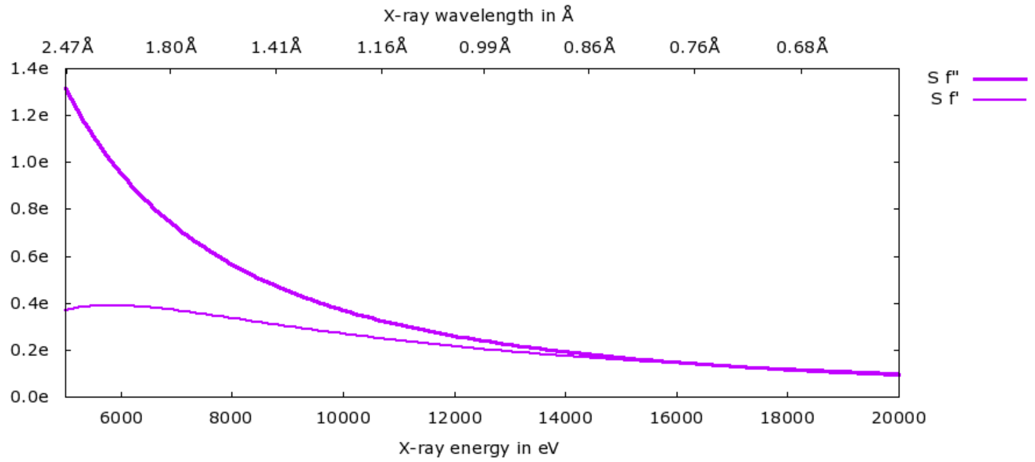


Fig. I-6. Plot of theoretical f' and f" values for sulfur (http://skuld.bmsc.washington.edu/scatter/AS_form.html).

Only for a few structure determinations singularly the anomalous signal of intrinsic sulfur was sufficient (Debreczeni, Bunkoczi, Girmann et al., 2003; Debreczeni, Girmann et al., 2003, Yang & Pflugrath, 2001). The largest structure solved by true sulfur SAD phasing so far, has been a 69 kDa protein from *Thermus thermophilus*. However, in this case the protein crystallized in the high symmetry space group $P2_12_12$ and due to two molecules in the asymmetric unit enabled the use of non-crystallographic symmetry (NCS) (Watanabe et al., 2005). These two characteristics are representative for the majority of crystals suitable for S-SAD phasing. Only rarely proteins crystallizing in low symmetry space groups could be solved by S-SAD phasing and these examples deal with molecules of 20-33 kDa molecular weight (Ramagopal et al., 2003; Dong et al., unpublished, PDB-ID 1YNB, deposited 2005). Most likely, the failure of S-SAD experiments in low symmetry space groups is based on the high multiplicity required fur successful S-SAD phasing, which is especially difficult to achieve for monoclinic crystals, since radiation damage during data collection becomes problematic. In particular in lower dose collecting modes – automatically used at longer wavelengths – radiation damage increases.

1.4. STRUCTURE BASED FUNCTION PREDICTION

In a recent publication with the promising title “Structure-based activity prediction for an enzyme of unknown function” Hermann and co-workers reported on the successful prediction of the cellular role of the protein Tm0936 from *Thermotoga maritima* by means of X-ray crystallography in combination with docking studies and subsequent experimental verification (Hermann et al., 2007). Similarly, the function of a member of the enolase superfamily was derived by docking experiments (Song et al., 2007). In the latter case, the calculations were based on a homology model instead of a crystal structure. However, for the 66.3 kDa protein no structure of a protein with sufficient similarity with respect to the amino acid sequence level has been available.

Despite the two successful structure-guided function predictions of the protein Tm0936 (Hermann et al., 2007) and the enolase (Song et al., 2007), the automated structure-based function prediction is still under development (reviewed in Redfern et al., 2008). The Protein Structure Initiative (PSI) e.g. set out to extend the ability to add structural information to the increasing number of genomic sequences. For most projects, it aims to derive the function of a given protein from the structure without additional experimental data (Watson et al., 2007). In these cases, significant global structural similarity is used in order to transfer functional annotations from close structural homologues, which cannot be identified only based on the amino acid sequences due to the absence of significant similarity on this level. Among others, the methods CE (combinatorial extension) (Shindyalov et al., 1998), STRUCTAL (Kolodny et al., 2005), CATHEDRAL (Redfern et al., 2007), FAT-CAT (Ye et al., 2003), SSM (secondary structure matching) (Krissinel et al., 2004) and DALI (Holm et al., 1996) have been applied in this context. Several recently developed methods such as ProKnow (Pal et al., 2005), Annolite (Marti-Renom et al., 2007) and PHUNCTIONER (Pazos et al., 2004) use comparisons of structural motives and sequence methods in addition to the global structure and assign confidence values for functional assignments. In contrast, PDBSITE (Ivanisenko et al., 2005), MSDSITE (Golovin et al., 2005) as well as PROFUNC (Laskowski et al, 2005) and TEMPURA (Porter et al., 2004) use only comparisons of binding or catalytic sites and thus are also feasible for proteins, where no protein of global structural similarity is known so far. The methods differ in their source of information. While PDBSITE and MSDSITE are based on functional annotations by the authors of a structure, the latter search the Catalytic Site Atlas (CSA) (Porter et al., 2004), a database of hand-curated catalytic residue assembles.

2. OBJECTIVES

The 66.3 kDa protein has recently been identified in a proteomics approach as a novel soluble protein of the lysosomal matrix and has been shown to be processed into a 28 and a 40 kDa fragment. Despite detailed molecular characterization, its cellular function has so far remained unknown. Since bioinformatics inference of function failed as well, this problem was tackled by means of X-ray crystallography. The determination of its three-dimensional structure and subsequent analysis were aimed to provide initial insight into the physiological function as well as a potential catalytic activity and respective substrates. A further, although minor, intention was to understand the putative assembly of the proteolytic fragments and the function and mechanism of the maturation process.

3. DE NOVO SULFUR SAD PHASING OF THE LYSOSOMAL 66.3KDA PROTEIN FROM MOUSE

3.1. OBJECTIVES AND AUTHORS` CONTRIBUTIONS

The following publication describes the *de novo* structure determination of the 66.3 kDa protein from mouse at a resolution of 2.4 Å. The main focus is on the applied methodical crystallographic procedures, since the successful sulfur SAD phasing of the 66.3 kDa protein expanded the use of this method with regard to both, the molecular weight of the studied macromolecule and the low symmetry of the space group (C2 monoclinic), an important parameter classifying a crystal. Thus, this article was selected by the editors of Acta Cryst. D, Ted Baker and Zbyszek Dauter, to be included as a short summary in the newsletter 17#1 of the International Union of Crystallography (IUCr). This newsletter shortly summarizes selected articles recently published in each of the eight IUCr journals. It is published quarterly and distributed to 587 libraries and 17,000 crystallographers and other interested individuals in 102 countries and posted on the IUCr website (http://journals.iucr.org/services/newsletter/newsletter-articles.html). The highlighted articles become open access for three months from the date of publication of the respective Newsletter. Additionally, the derived structure of the 66.3 kDa protein was chosen to be emphasized as the “BESSY structure of the month” on the homepage of the synchrotron BESSY, Berlin, Germany (http://www.mx.bessy.de/structures/index.shtml) presumably by its update in summer (2009).

Protein expression in a human fibrosarcoma cell line and the purification procedure up to the anion exchange chromatography step inclusively were performed by the laboratory of Prof. Dr. Torben Lübke (Center of Biochemistry and Molecular Cell Biology, Department Biochemistry II, GZMB, Georg-August University Göttingen). Initial crystallization hits were obtained in trials set up by Annette Berndt and Dr. Achim Dickmanns (Department of Molecular Structural Biology, Institute of Microbiology and Genetics, GZMB, Georg-August University Göttingen). My contributions under supervision of Prof. Dr. Ralf Ficner concern crystal optimization including the addition of gel filtration as the final purification step, data collection with the help of Dr. Uwe Mueller (BESSY GmbH, Macromolecular Crystallography group, Berlin, Germany) as well as structure determination, refinement and analysis (PDB-ID 3FBX).

3.2. PUBLICATION "DE NOVO SULFUR SAD PHASING OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE"

Acta Crystallographica Section D
**Biological
Crystallography**
ISSN 0907-4449

De novo sulfur SAD phasing of the lysosomal 66.3 kDa protein from mouse

Kristina Lakomek,[a] Achim Dickmanns,[a] Uwe Mueller,[b] Katrin Kollmann,[c] Florian Deuschl,[c] Annette Berndt,[a] Torben Lübke[c] and Ralf Ficner[a]*

[a]Department of Molecular Structural Biology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Justus-von-Liebig-Weg 11, D-37077 Göttingen, Germany, [b]BESSY GmbH, Macromolecular Crystallography Group, Albert-Einstein-Strasse 15, D-12489 Berlin, Germany, and [c]Center of Biochemistry and Molecular Cell Biology, Department of Biochemistry II, Georg-August University Göttingen, Heinrich-Dueker-Weg 12, D-37073 Göttingen, Germany

Correspondence e-mail: rficner@gwdg.de

Received 24 October 2008
Accepted 9 December 2008

PDB Reference: lysosomal 66.3 kDa protein, 3fbx, r3fbxsf.

The 66.3 kDa protein from mouse is a soluble protein of the lysosomal matrix. It is synthesized as a glycosylated 75 kDa preproprotein which is further processed into 28 and 40 kDa fragments. Despite bioinformatics approaches and molecular characterization of the 66.3 kDa protein, the mode of its maturation as well as its physiological function remained unknown. Therefore, it was decided to tackle this question by means of X-ray crystallography. After expression in a human fibrosarcoma cell line, the C-terminally His-tagged single-chain 66.3 kDa variant and the double-chain form consisting of a 28 kDa fragment and a 40 kDa fragment were purified to homogeneity but could not be separated during the purification procedure. This mixture was therefore used for crystallization. Single crystals were obtained and the structure of the 66.3 kDa protein was solved by means of sulfur SAD phasing using data collected at a wavelength of 1.9 Å on the BESSY beamline BL14.2 of Freie Universität Berlin. Based on the anomalous signal, a 22-atom substructure comprising 21 intrinsic S atoms and one Xe atom with very low occupancy was found and refined at a resolution of 2.4 Å using the programs *SHELXC*/*D* and *SHARP*. Density modification using *SOLOMON* and *DM* resulted in a high-quality electron-density map, enabling automatic model building with *ARP*/*wARP*. The initial model contained 85% of the amino-acid residues expected to be present in the asymmetric unit of the crystal. Subsequently, the model was completed and refined to an $R_{\rm free}$ factor of 19.8%. The contribution of the single Xe atom to the anomalous signal was analyzed in comparison to that of the S atoms and was found to be negligible. This work should encourage the use of the weak anomalous scattering of intrinsic S atoms in SAD phasing, especially for proteins, which require both expensive and time-consuming expression and purification procedures, preventing extensive screening of heavy-atom crystal soaks.

1. Introduction

Lysosomes are membrane-bordered organelles in eukaryotic cells that contain a set of about 60 acid hydrolases and associated proteins that are responsible for the digestion of various macromolecules and even whole organelles derived from various sources by endocytosis, autophagy and other trafficking pathways (Sleat *et al.*, 2008; reviewed by Sleat *et al.*, 2007; Lübke *et al.*, 2008). Typically, the lack of function of lysosomal proteins causes severe pathogenic phenotypes collectively referred to as 'lysosomal storage diseases' which are associated with the accumulation of undigested molecules in the lysosomal compartment (reviewed by Scriver *et al.*,

© 2009 International Union of Crystallography
Printed in Singapore – all rights reserved

2001). In addition to their degradative function, lysosomal enzymes are involved in many pathophysiological processes such as cancer (Fehrenbacher & Jaattela, 2005), tumour metastasis and propagation (Garcia *et al.*, 1996; Kos & Lah, 1998), neuroprotection (Cravatt *et al.*, 2001) and Alzheimer's disease (Nixon & Cataldo, 2006), as well as in the modulation of hormones and bioactive lipids implicated, for example, in tissue homeostasis, inflammation (Capasso *et al.*, 2001; Izzo *et al.*, 2001; Feulner *et al.*, 2004; reviewed by Hansen *et al.*, 2000) and anorexia (Saftig *et al.*, 1995).

Most lysosomal enzymes are directed to lysosomes by a mannose 6-phosphate (M6P) residue that is attached to their *N*-glycans post-translationally, recognized by M6P receptors (MPRs) at the trans-Golgi network and subsequently targeted to the lysosomes. Based on the composition of the oligosaccharide, the *N*-glycans can be subdivided in three types, namely the 'high-mannose', 'hybrid' and 'complex' types, all of which exhibit a common pentasaccharide core that consists of three different mannose residues and two *N*-acetylglucosamine (NAG) moieties. In order to reveal novel lysosomal proteins and thus novel lysosomal functions, several subproteomic studies dealing with soluble lysosomal proteins have recently been carried out (Sleat, Wang *et al.*, 2006; Sleat, Zheng *et al.*; 2006; Kollmann *et al.*, 2005; Sleat *et al.*, 2008; reviewed by Sleat *et al.*, 2007). Among others, the 66.3 kDa protein has been identified as a putative soluble lysosomal protein in several lysosomal proteomic studies from mouse, rat and human (Lübke *et al.*, 2008). Subsequently, the murine 66.3 kDa protein and its human orthologue p76 were characterized in more detail regarding their lysosomal localization, processing and glycosylation (Deuschl *et al.*, 2006; Jensen *et al.*, 2007). A C-terminally RGS-His$_6$-tagged derivative of the mouse 66.3 kDa protein was stably expressed at high levels in the human fibrosarcoma cell line HT1080, which secreted the protein into the medium. It was then purified by a combination of affinity and ion-exchange chromatography (Deuschl *et al.*, 2006). The 66.3 kDa protein is synthesized at the rough ER as a glycosylated precursor with an apparent molecular weight of 75 kDa and is further processed into a 28 kDa N-terminal fragment and a 40 kDa C-terminal fragment (Deuschl *et al.*, 2006). Such an extensive endosomal/lysosomal maturation by limited proteolysis is a common step towards the final activation of lysosomal hydrolases (Hasilik, 1992). However, despite bioinformatics studies and molecular characterization, the physiological function of the 66.3 kDa protein so far remains unknown. In order to obtain initial insights into its activity, we set out to determine the crystal structure of the 66.3 kDa protein from mouse. Since no structure of a protein with sufficient similarity at the amino-acid sequence level is available, the phases were determined experimentally.

Currently, several high-brilliance third-generation synchrotron beamlines are available worldwide that support the required energy range for long-wavelength phasing applications (Djinovic Carugo *et al.*, 2005). Thus, the use of SAD phasing using only weak anomalous scatterers has increased (Ramagopal *et al.*, 2003) since the first protein structure was determined by sulfur SAD in 1981 (Hendrickson & Teeter, 1981). Most sulfur SAD data have been collected on tunable synchroton beamlines in order to make use of the appreciably larger f'' of sulfur at longer wavelengths in the range between 1.7 and 2.5 Å (Brown *et al.*, 2002; Gordon *et al.*, 2001; Liu *et al.*, 2000; Ramagopal *et al.*, 2003; Weiss, 2001; Weiss *et al.*, 2004), although Cu $K\alpha$ is also feasible for sulfur SAD phasing (Dauter *et al.*, 1999; Sekar *et al.*, 2004; Roeser *et al.*, 2005). However, successful phasing often relies on the additional presence of several different weak anomalous scatterers such as Ca^{2+} and Cl^- ions (Dauter *et al.*, 1999; Yang & Pflugrath, 2001; Debreczeni, Bunkóczi, Ma *et al.*, 2003). Only rarely are intrinsic S atoms the sole source of the anomalous signal (Debreczeni, Bunkóczi, Girmann *et al.*, 2003; Debreczeni, Girmann *et al.*, 2003; Yang & Pflugrath, 2001). When further heavy-atom sites were determined in addition to the sulfur sites, the former refined to significantly higher occupancies; for example, occupancies of 1.0 and 0.99 for two calcium ions compared with 0.43 for the highest occupancy for a sulfur site in the formylglycine-generating enzyme (Roeser *et al.*, 2005). To our knowledge, the largest structure solved by true sulfur SAD phasing to date has been that of TT0570 from *Thermus thermophilus*, with a molecular weight of 69 kDa. However, this protein was crystallized in space group $P2_12_12$ with two molecules in the asymmetric unit (Watanabe *et al.*, 2005). Like TT0570, the majority of crystals suitable for sulfur SAD phasing experiments belong to high-symmetry space groups. In contrast, only a limited number of successful sulfur SAD phasings of proteins that crystallize in low-symmetry space groups have been described for monoclinic crystals; for example, xylanase from *Thermoascus aurantiacus* (molecular weight 33 kDa; Ramagopal *et al.*, 2003) and the hypothetical protein AF1432 (molecular weight 20 kDa; PDB code 1ynb; A. Dong, T. Skarina, A. Savchenko, E. F. Pai & A. Edwards, unpublished work) crystallized in space groups $P2_1$ and $C2$, respectively. In this work, we demonstrate on the basis of the *de novo* phasing of the 66.3 kDa protein that sulfur SAD phasing is also feasible for a larger protein that crystallizes in a low-symmetry space group such as $C2$.

2. Materials and methods

2.1. Protein purification and crystallization

The 66.3 kDa protein was purified as described by Deuschl *et al.* (2006). For crystallization, size-exclusion chromatography was added as a final purification step (10 m*M* Tris–HCl pH 8.0; Superdex 200 HR 10/300, GE Healthcare). The molecular-weight markers were purchased from Fermentas (SM0431; Fermentas, St Leon-Rot, Germany). The purified protein, which consisted of a mixture of the 66.3 kDa full-length protein and the 40 kDa fragment as well as the 28 kDa fragment, was crystallized using the sitting-drop vapour-diffusion method at 293 K. The 66.3 kDa protein was crystallized in a drop composed of 1.6 µl protein solution with a concentration of 23 mg ml^{-1} and 2.0 µl reservoir solution containing 12%(*w*/*v*) PEG 4000, 200 m*M* ammonium acetate and 100 mM sodium acetate/acetic acid pH 4.6. The final pH of the reservoir solution was determined to be 5.0.

2.2. Data collection and processing

For data collection, the crystal was transferred into a cryo-protecting solution that consisted of 16%(w/v) PEG 4000, 130 mM ammonium acetate, 60 mM sodium acetate/acetic acid pH 4.6 and 10%(v/v) glycerol. Subsequently, the 66.3 kDa protein crystal was derivatized with xenon in a gas chamber for 4 min under a gas pressure of 2.8 MPa and then directly flash-cooled in liquid nitrogen before mounting in the cryo-stream. The data set was collected on BESSY beamline BL14.2, which was equipped with an SX165 detector (Rayonics LLC, Illinois, USA). The crystal was mounted at a distance of 50 mm and the beam stop was adjusted to a distance of 14 mm from the detector. 1120 images were recorded at a wavelength of 1.900 Å from a single crystal in 1.0° oscillations with 3.2 s exposure time at 100 K using 0.03 mm aluminium foil between the X-ray beam and the crystal in order to reduce the beam intensity and radiation damage. The images were processed with *DENZO* and *SCALEPACK* as implemented in *HKL*-2000 (HKL Research, Inc., Charlottesville, Virginia, USA). The scaled data were analysed with *XPREP* (Bruker AXS Inc., Madison, Wisconsin, USA). $R_{\rm anom}$ and $R_{\rm p.i.m.}$ were determined with *SCALA* (Collaborative Computational Project, Number 4, 1994) based on the data scaled using *SCALEPACK* without merging the original indices.

2.3. Structure determination and model building

The anomalous completeness listed in Table 1 corresponds to the output from *SHELXC* (Sheldrick, 2008). *autoSHARP* (de La Fortelle & Bricogne, 1997) served as the platform for structure determination using data in the resolution range 32.11–2.40 Å. *SHELXC*/*D* (Sheldrick, 2008) and *SHARP* (de La Fortelle & Bricogne, 1997) were used for substructure determination and for positional refinement and phase calculations, respectively. The phases obtained were further improved by solvent flattening and histogram matching using *SOLOMON* and *DM* (Collaborative Computational Project, Number 4, 1994) as implemented in *SHARP*. A free-atom model was built into the electron-density map by *ARP*/*wARP* (Perrakis *et al.*, 1999) as implemented in *SHARP* using data to a resolution of 2.40 Å and a solvent content of 56.4%. Subsequent auto-tracing of the amino-acid chain using the *warpNtrace* procedure of *ARP*/*wARP* resulted in an initial model of high quality containing 475 amino acids (of 559 expected amino acids) in 20 chains with a connectivity index of 0.92. Subsequently, the model was completed manually with *Coot* (Emsley & Cowtan, 2004) and refined using *REFMAC*5 from the *CCP*4 suite (Murshudov *et al.*, 1997; Collaborative Computational Project, Number 4, 1994) to R factors of $R_{\rm work}$ = 15.6% and $R_{\rm free}$ = 19.8% (see Table 1), with r.m.s. deviations of 0.013 Å and 1.44° for bond lengths and angles, respectively. Stereochemical analysis of the refined structure was performed with *PROCHECK* (Laskowski *et al.*, 1993). 520 of the 559 amino-acid residues (Val63–Thr238 and Cys249–Pro592), five N-acetylglucosamine moieties, 294 water molecules, three glycerol molecules, three polyethylene glycol molecules, one sodium ion, one acetate ion and one Xe atom with an occupancy of 10% were included in the final structure. The protein atoms exhibit average B factors of only 23.9 Å^2 for the amino-acid residues and of 49.2 Å^2 for the atoms of the sugar moieties attached to Asn115, Asn236, Asn441 and Asn520, while the average B factors of the water molecules and the other solvent molecules were 33.7 and 51.4 Å^2, respectively. Cys249 was oxidized in some molecules forming the crystal as visualized in the electron density. Therefore, this residue was modelled as a cysteine sulfonic acid with an occupancy of 0.5 for all side-chain O atoms.

Table 1
Crystallographic data and refinement statistics.

Values in square brackets and parentheses are for the lowest and highest resolution shells, respectively.

Wavelength (Å)	1.900
X-ray source/detector	BL14.2/SX165
Resolution range (Å)	[50–5.17] 50.0–2.40 (2.49–2.40)
Space group	$C2$
No. of molecules in ASU	1
Unit-cell parameters (Å, °)	a = 148.80, b = 89.67, c = 64.95, α = 90.0, β = 98.7, γ = 90.0
Measured reflections	751575
Unique reflections	32026
Rejected reflections	195 (~0.01%)
Multiplicity of reflections	[23.4] 23.0 (21.2)
Completeness (%)	[99.7] 98.9 (97.5)
Anomalous completeness† (%)	[98.3] 98.7 (97.0)
Mosaicity (°)	0.76
$R_{\rm merge}$‡ (%)	[3.8] 8.3 (37.3)
Average $I/\sigma(I)$	[102.9] 48.8 (10.2)
Maximum $\Delta F''/\sigma(\Delta F)$	1.81
f'/f''§ (electrons)	0.37/0.82
$R_{\rm Cullis}$ (anomalous)¶	0.95
Anomalous phasing power†	0.474
Expected no. of S atoms	22
Heavy-atom sites found in ASU	22 (21 S, 1 Xe)
Amino acids in ASU (expected/placed)	559/520
$R_{\rm work}$ (%)	15.6
$R_{\rm free}$ (%)	19.8

† According to *SHELXC* with resolution bins ∞–8.0 and 2.6–2.4 Å. ‡ $R_{\rm merge} = 100 \times \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$. § Theoretical values for sulfur at a wavelength of 1.9000 Å used by *SHARP*. ¶ $R_{\rm Cullis}$ (acentric reflections) = $\sum E/\sum \Delta F$, where E is the residual lack-of-closure error.

The Harker section in Fig. 3 was prepared using the program *XPREP*. Figs. 2 and 4(b)–4(f) were prepared using *CCP4mg* (Potterton *et al.*, 2004), while Fig. 4(a) was prepared with *Coot*.

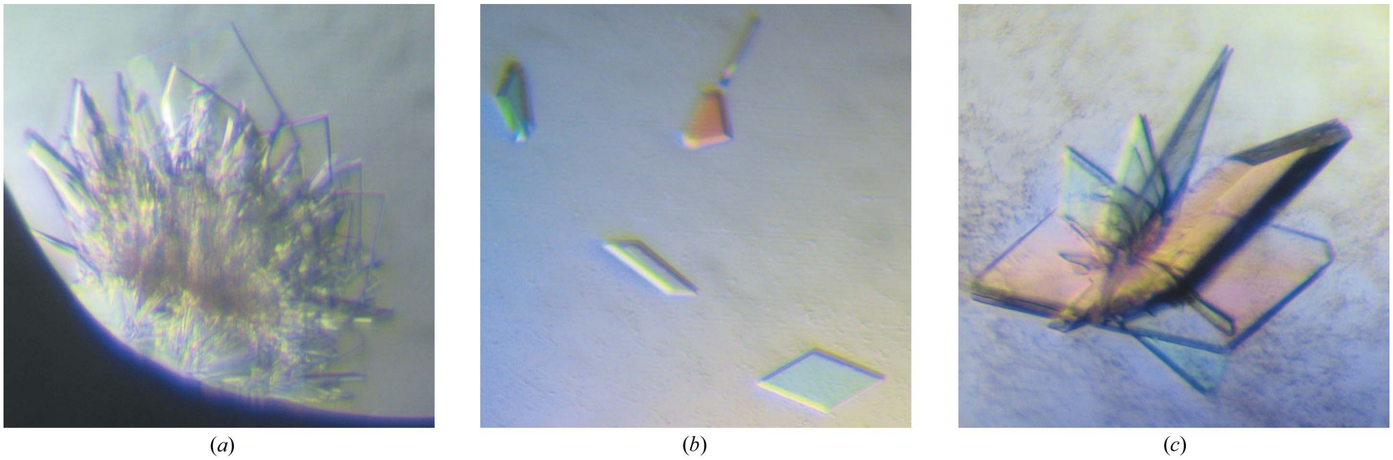
3. Results and discussion

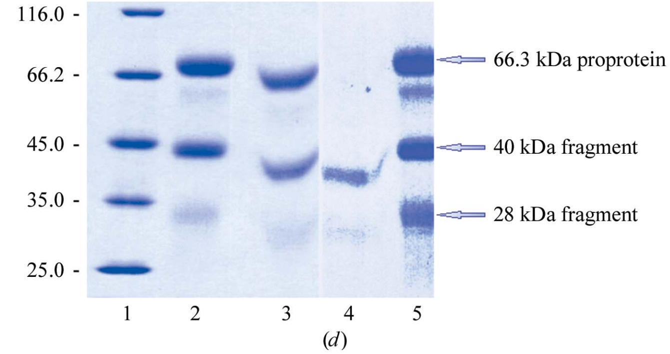
3.1. Purification and crystallization of the 66.3 kDa protein

The glycosylated 66.3 kDa protein from mouse was produced by overexpression as a C-terminally RGS-His$_6$-tagged derivative in a human fibrosarcoma cell line (HT1080) and purified as described elsewhere (Deuschl *et al.*, 2006). Since the 66.3 kDa proprotein and the 28 and 40 kDa fragments could not be separated during the purification procedure, this mixture was used for crystallization. Initial crystals were obtained under acidic conditions near the physiological pH of lysosomes (pH 4.6) using PEG 4000 and ammonium acetate as precipitants. Initial crystals grew in compact clusters and could not be separated (Fig. 1*a*). Extensive variation of the reservoir solution did not improve crystal growth, although exchanging the ammonium acetate for diammonium hydrogen citrate resulted in single crystals (Fig. 1*b*). However, the crystals exhibited a thin plate-like shape and were also not suitable for data collection. Only optimization of the protein concentration and variation of the protein:reservoir ratio in addition to a final size-exclusion chromatography step in the purification procedure resulted in crystals of suitable quality. The supplementary gel filtration seemed to be most important in order to obtain crystals that could be separated despite growing in clusters. Such crystals are shown in Fig. 1(*c*). It is most likely that this optimization is based on the removal of small amounts of aggregated protein that elute in the void volume of the column. The crystals obtained were separated from adhering precipitate in a cryoprotecting solution which contained both an increased PEG concentration and 10%(*v*/*v*) glycerol to prevent ice formation. Since the protein mixture used for crystallization contained three different polypeptide chains, the crystals were dissolved in order to gain information about their content. Owing to the sensitivity of the thin plate-like crystals towards breakage, only two washing steps in reservoir solution were performed. SDS–PAGE analysis of the dissolved crystals and staining with Coomassie Brilliant Blue R/G revealed only one prominent band corresponding to the 40 kDa fragment and an additional faint signal of about 28 kDa corresponding to the second fragment, as shown in Fig. 1(*d*). However, weak staining of the 28 kDa band was also observed for the purified protein solution and the weak signal seemed to be caused by the crystal content rather than by precipitate that had remained stuck to the crystal despite washing. The single-chain form represented by the band at around 66 kDa was unambiguously not present in the crystal. The complete absence of the 66.3 kDa form was also observed after incubation for two weeks at 293 K of crystallization drops from which no crystals were obtained, suggesting that it is processed autocatalytically into the stable 28 and 40 kDa fragments over time. Both the 66.3 kDa protein and the fragments of the solution used for crystallization displayed a higher apparent molecular weight in SDS–PAGE than expected owing to the presence of five glycosylation sites which were all utilized upon expression as demonstrated by two different methods: deglycosylation by peptide:*N*-glycosidase F (PNGase F) treatment (Deuschl *et al.*, 2006) and structural analysis (see below). However, the addition of PEG 4000 resulted in a lower apparent molecular weight of the proteins from the dissolved crystal, as can be seen in lane 4 of Fig. 1(*d*) by comparison with the solution of the purified protein in 10 m*M* Tris–HCl pH 8.0 (lanes 2 and 5). Therefore, this solution (lane 2) was mixed with an appropriate amount of reservoir solution (lane 3) in order to serve as a reference for the dissolved crystal.

3.2. Data collection and structure determination

After incubation in cryobuffer, the crystal containing both fragments was subjected to pressurization with xenon in a


(*a*) (*b*) (*c*)



(*d*)

Figure 1
Crystallization of the murine 66.3 kDa protein. (*a*) Initial crystals of the 66.3 kDa protein obtained using protein purified without size-exclusion chromatography grew in compact clusters. (*b*) Single but small crystals were obtained by the exchange of ammonium acetate for diammonium hydrogen citrate. (*c*) Larger separable crystals of suitable quality only formed after the addition of gel filtration as a final purification step. (*d*) SDS–PAGE of the 66.3 kDa protein crystals. The gel was stained with Coomassie Brilliant Blue R/G. Lane 1, molecular-weight markers; lanes 2 and 5, protein solution used for crystallization (10 m*M* Tris–HCl pH 8.0); lane 3, protein solution mixed with reservoir solution to achieve similar final conditions and concentrations as in the crystallization drop; lane 4, dissolved crystals.

xenon gas chamber (Hampton Research, Aliso Viejo, USA) at BESSY, Berlin. An initial data set was collected on beamline BL14.2 using a wavelength of 1.700 Å in order to use the anomalous signal of Xe atoms for experimental phasing. However, no significant anomalous signal was detected by *SHELXC* (data not shown). Therefore, a second data set was collected at a wavelength of 1.900 Å using the same crystal in order to determine the phases by means of sulfur SAD (see Table 1). 1120 images of 1.0° were taken to ensure high multiplicity of the data. The crystals belonged to the monoclinic space group $C2$, with unit-cell parameters a = 148.8, b = 89.7, c = 65.0 Å, β = 98.7° and one molecule in the asymmetric unit. An average redundancy of 23.0 was obtained after 1120° of rotation, before radiation damage became a serious problem as indicated by an increase in the R factor. The damage was confirmed during model building, when negative peaks became visible in the $F_{\rm o} - F_{\rm c}$ difference map for the carboxyl groups of some glutamate and aspartate side chains. The data were processed with *HKL*-2000 and cut off at a resolution of 2.40 Å. The scaled data were analyzed using *XPREP*. As indicated by the $R_{\rm p.i.m.}$ value of 2.4%, the data were of reasonable quality and enabled us to solve the structure by means of sulfur/Xe SAD phasing using the xenon-derivatized crystal. Data statistics are summarized in Table 1. The crystal displayed a mosaicity of 0.76° and a unit-cell volume of about 857 000 Å^3. The $d''/\mathrm{sig}(d'')$ value output by *SHELXC* as an estimate of the anomalous signal amounted to a maximum of 1.81 in the inner resolution shell (∞–8.0 Å) and dropped below the threshold (0.80) at 3.0 Å, indicating the presence of only moderate anomalous signal. Nevertheless, using *SHELXC*/*D* and *SHARP* as implemented in *autoSHARP*, 22 heavy-atom sites could be identified in the asymmetric unit and refined to reliable occupancies (0.36–0.75; see §2). As suggested by SDS–PAGE analysis of dissolved crystals, structure determination proved that the crystals were composed of the 40 kDa and the 28 kDa fragments in a 1:1 molar ratio. Thus, all cysteine residues of the processed 66.3 kDa protein except for Cys157 and all methionines were represented by significant peaks in the anomalous difference map (six and 15 heavy-atom sites, respectively). The heavy-atom sites are shown in Fig. 2(*a*) together with the C$^{\alpha}$ ribbon as well as the cysteine and methionine side chains of the final model. Two disulfide bridges are formed in the 66.3 kDa protein, namely between Cys147 and Cys157 and between Cys497 and Cys500. Since a peak for Cys157 only appears in the anomalous map at a contour level below 3.5σ and thus the respective S atom has not been detected by *SHELXD*, the disulfide bond between Cys147 and Cys157 only corresponds to a single significant peak around Cys147 in the anomalous map and to one heavy-atom position. In contrast, the second disulfide bond (between Cys497 and Cys500) is represented by individual heavy-atom sites refined to similar occupancies (0.36 and 0.42, respectively) and by two separate peaks in the anomalous map as shown in Fig. 2(*b*). At the resolution of 2.4 Å used for finding the heavy-atom sites, disulfide bridges with a typical distance of 2.1 Å between the S$^{\gamma}$ atoms are commonly not represented by individual peaks but appear together as so-called 'supersulfurs' (Sheldrick, 2008). The detection of individual peaks for the Cys497–Cys500 disulfide bridge and a peak corresponding to only one S$^{\gamma}$ atom of the other cysteines suggest that both S—S bonds have been partially reduced during protein purification and crystallization owing to the absence of a reducing agent or by radiation damage during data collection.

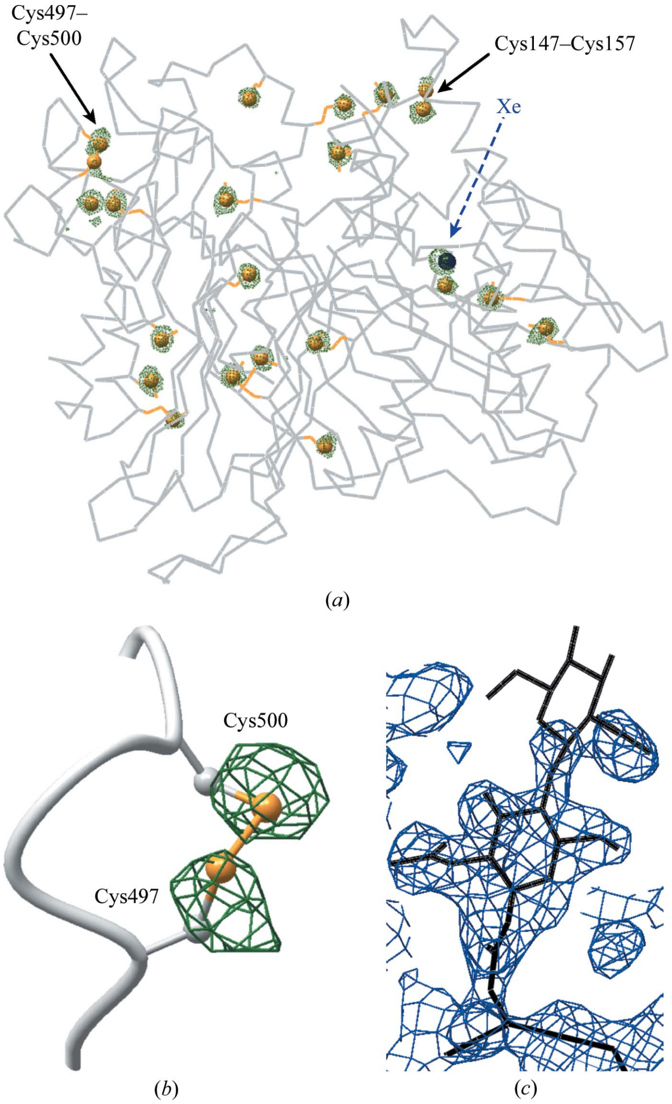


Figure 2
(*a*) Anomalous difference map contoured at 4.0σ. The map is coloured dark green. One peak represents the Xe atom, which is indicated as a blue sphere. For orientation, the C$^{\alpha}$ trace of the refined 66.3 kDa protein structure is shown with cysteine and methionine side chains in stick mode and S atoms highlighted as yellow spheres. The two disulfide bonds are marked with black arrows. (*b*) One of the two internal disulfide bridges is represented by two individual peaks in the anomalous difference map and enabled both cystine S atoms to be found. The anomalous map is contoured at the 4.0σ level. (*c*) Electron-density map after density modification. One *N*-acetylglucosamine moiety of the glycan attached to Asn115 (for example) is already clearly visible in the experimental map at a level of 1.5σ before the inclusion of any model phases and manual intervention.

In addition to the peaks representing the intrinsic S atoms of the protein in the anomalous difference map, one further

Table 2
Comparison of the phasing statistics with and without the use of the anomalous signal of the Xe atom as output by *SHARP*.

Automatic model building was performed in 100 cycles of *ARP*/*wARP* and *REFMAC*5. For comparison, the statistics from *autoSHARP* (using the Xe atom, but treating it like the S atoms) are also given.

Phasing statistics	*autoSHARP* (Xe as S)	Including the Xe atom in *SHARP*	*SHARP* without the Xe atom as a heavy-atom site
Phasing power (*SHARP*)	0.47	0.47	0.47
$R_{\rm Cullis}$	0.95	0.95	0.95
$\rm FOM_{acentric}$	0.25	0.25	0.25
$R_{\rm work}$ from *REFMAC*5 (*ARP*/*wARP* cycling) (%)	25.0	27.3	27.7

peak indicates the binding of an Xe atom at a specific site. The xenon had been caught in a large hydrophobic pocket during a xenon pressurization performed prior to data collection. In the anomalous map it is represented by a peak at the 14.8σ level, while the S atom of Cys347 nearby gives rise to a peak of only 9.3σ. However, a comparison with the anomalous densities of the S^{δ} atoms of Met175 (14.3σ) and Met171 (17.4σ; the strongest peak of all) also located in the close vicinity shows that the Xe atom does not contribute excessively to the anomalous signal. Additionally, there are no peaks of outstanding intensity in the $v = 0$ Harker section shown in Fig. 3.

For successful sulfur SAD phasing, the error in the observed intensities as described by the merging R factor $R_{\rm p.i.m.}$ (Weiss, 2001) has to be significantly smaller than the observed signal described by the anomalous R factor $R_{\rm anom}$ [Weiss *et al.*, 2001; $R_{\rm anom} = 100 \times \sum_{hkl} |I(hkl) - I(-h-k-l)| / \sum_{hkl} \langle I(hkl) \rangle$, $R_{\rm p.i.m.} = 100 \times \sum_{hkl} [1/(N-1)]^{1/2} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$]. Accordingly, Weiss and coworkers consider an $R_{\rm anom}/R_{\rm p.i.m.}$ ratio of 1.5 to be favourable. For the data presented here, the $R_{\rm anom}/R_{\rm p.i.m.}$ ratio is only 1.8%/1.7% = 1.1 and thus at first would seem to prevent sulfur SAD phasing ($R_{\rm p.i.m.}$ for all $I+$ and $I-$). However, as described in this paper, these data were successfully used to determine the structure of the 66.3 kDa protein, thus extending the previous lower limit for the $R_{\rm anom}/R_{\rm p.i.m.}$ ratio of 1.5 significantly to 1.1. In contrast, the second frequently used indicator of the amount of anomalous signal, the estimated Bijvoet intensity ratio χ, had been promising from the beginning. It can be calculated using the equation $\chi = \langle \Delta F^{\rm anom} \rangle / \langle F \rangle = (2N_{\rm A}/N_{\rm P})^{1/2} / (f''_{\rm A}/f^{\rm O}_{\rm eff})$, where $N_{\rm A}$ is the number of anomalous scatterers and $N_{\rm P}$ is the total number of protein atoms and with the atomic scattering factor $f^{\rm O}_{\rm eff} = 6.7$ of an 'average' protein atom (Hendrickson & Teeter, 1981; $\Delta F^{\rm anom} = \Delta F^{\pm}$). For the 66.3 kDa protein with 22 expected anomalously scattering S atoms of about 4400 atoms in the whole structure with a theoretical value of $f''_{\rm S} = 0.82$ at 1.900 Å, the expected Bijvoet ratio amounts to about 1.8%. Bijvoet ratios in the same range have been reported, for example, for lima bean trypsin inhibitor (1.96%) with data collected at 1.54 Å (Debreczeni, Bunkóczi, Girmann *et al.*, 2003) as well as for the protein TT0570 (1.1%), the structure of which was determined by Watanabe and coworkers using a recently developed loopless free crystal mounting as well as a Cu/Cr dual-wavelength system (Watanabe, 2006). Previously, with commonly used cryoloops, a Bijvoet ratio of only 0.6% had been predicted by Wang to be sufficient for successful SAD phasing (Wang, 1985).

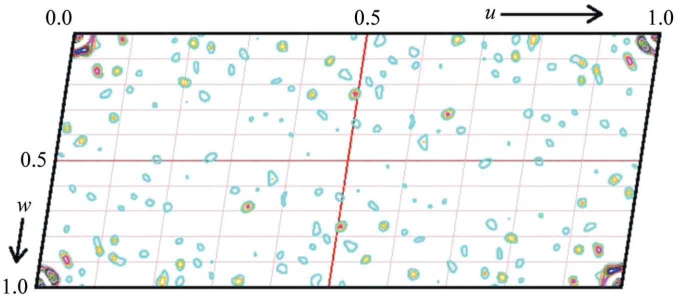


Figure 3
Anomalous difference Patterson map. Sharpened map of the Harker section $v = 0$ calculated at a resolution of 2.4 Å. Contours are in increments of 1.3σ and are coloured differently for each level.

Table 3
Occupancies and temperature factors of the heavy-atom sites.

A comparison of the occupancies of the xenon site and the S atoms as output by *SHARP* clearly reveals that the Xe atom detected as sulfur does not contribute greatly to the anomalous signal.

Heavy-atom site	Occupancy	B factor (Å^2)	Residue
1	0.75	64.9	Cys249
2	0.72	78.6	Met384
3	0.69	70.1	Cys147
4	0.68	44.1	Met140
5	0.65	33.3	Met171
6	0.61	45.8	Met551
7	0.60	34.4	Met275
8	0.59	38.0	Met575
9	0.59	54.1	Met585
10	**0.57**†	**41.8**	**Xe**
11	0.56	43.6	Met578
12	0.56	35.6	Met175
13	0.52	36.9	Met137
14	0.49	35.9	Met480
15	0.48	45.3	Met477
16	0.47	33.3	Met484
17	0.46	35.7	Cys562
18	0.42	23.2	Cys500
19	0.37	27.8	Met412
20	0.37	41.1	Cys347
21	0.36	20.2	Met549
22	0.36	26.7	Cys497

† The value of the occupancy of the Xe atom does not reflect its real occupancy, since the Xe atom was treated as an S atom during refinement of the heavy-atom sites (see text for details).

The experimental map obtained for the 66.3 kDa protein was of excellent quality. Some glyosylations were already visible in the electron-density map prior to manual intervention, *e.g.* an *N*-acetylglucosamine moiety at Asn115 as shown in Fig. 2(*c*).

3.3. Relative contribution of the Xe and the S atoms to the anomalous scattering: requirements for phasing

The Xe atom, which was caught in a hydrophobic pocket during pressurization with xenon gas, was not essential for phase determination. A comparison of the occupancies of the 22 sulfur sites with that of the single xenon site showed that the Xe atom only contributed about 5% of the overall occupancies of all anomalous scatterers and thus can be regarded as not being a prerequisite for successful phasing. In order to prove the negligible contribution of the Xe atom with the available data, phase determination was carried out explicitly omitting the xenon site in calculations using *SHARP*. A comparison of the initial electron-density maps clearly shows that there are no significant differences between the maps calculated in the absence or the presence of the Xe atom as a heavy-atom site (see Fig. 4). The phasing statistics for both procedures are summarized in Table 2. According to the program *MAPMAN* (Kleywegt & Jones, 1996), the correlation coefficient for the two maps amounts to 0.75. The quality of both maps enabled automatic model building with *ARP*/*wARP* (Perrakis *et al.*, 1999).

The minor contribution of the Xe atom to the anomalous signal reflects its quite low occupancy in the crystal, which can be explained as follows. During *autoSHARP* calculations for the determination of the 66.3 kDa protein structure, the Xe atom was assumed to display the same properties as the 21 S

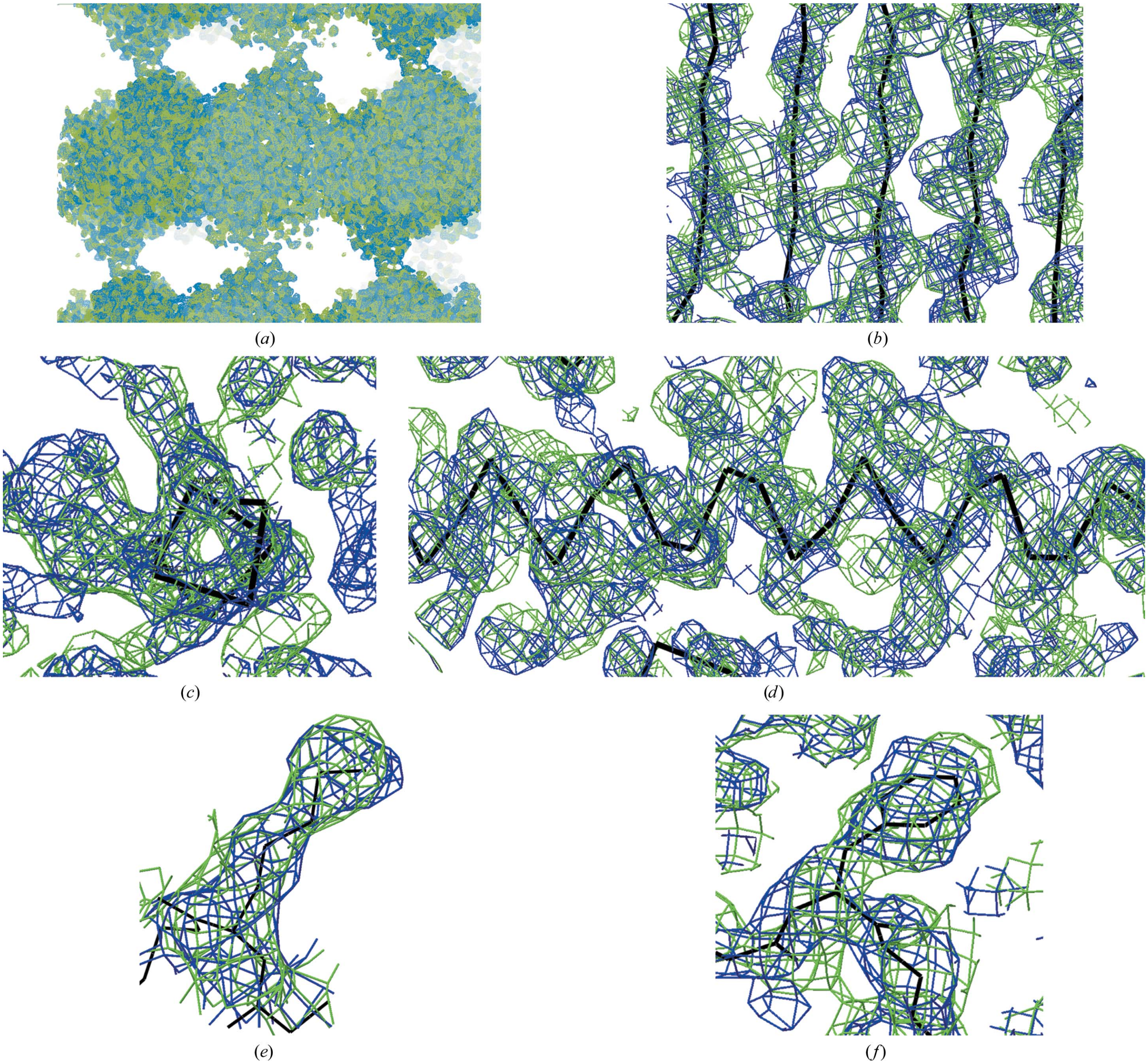

(*a*)
(*b*)
(*c*)
(*d*)
(*e*)
(*f*)

Figure 4
Comparison of the experimental electron-density maps after density modification providing *SHARP* with input parameters with heavy-atom sites including and lacking the Xe atom, respectively. The maps are coloured blue and green, respectively, and are contoured at the 1.5σ level. (*a*) The section encompassing several neighbouring molecules shows a clear solvent boundary. (*b*) The β-strands forming a β-sheet are well defined in the electron-density map. (*c*, *d*) Density for an α-helical structure is also visible, viewed from the top (*c*) and from the side (*d*) of the helix, respectively. (*e*, *f*) Close-up views of the side chains of Lys340 at the surface (*e*) and Phe371 in the hydrophobic core of the protein (*f*), respectively.

atoms used for phasing. Therefore, the occupancy of the Xe atom listed in Table 3 (0.57) was refined based on the assumption that the site was occupied by an S atom. In order to obtain the actual occupancy of the Xe atom, a correction factor has to be applied to the value listed in Table 3. This correction factor has to take into account the relationship between the anomalous signals of S and Xe atoms. At the wavelength of 1.9000 Å at which the SAD data were collected, an f'' value of about 10.29 is expected for an Xe atom, while sulfur only exhibits a value of $f'' \simeq 0.82$. Therefore, a fully occupied xenon site commonly provides a signal with an intensity 12.5 times as high as that of a fully occupied sulfur site. Thus, the correction factor for calculation of the occupancy of the xenon site amounts to 12.5 and the occupancy of 0.57 for the xenon site (Table 3) has to be divided by 12.5, resulting in an actual occupancy of only 0.046. However, the mean occupancy of the S atoms used for phasing amounts to only 0.54 (see Table 3) instead of the theoretical value of 1.0 for an intrinsic protein atom, indicating that the occupancy of all heavy-atom sites is likely to be underestimated by a factor of about 1.85. Even upon application of this further correction factor, the Xe atom seems to be bound in a maximum of 8.5% of all molecules of the 66.3 kDa protein forming the crystal. Thus, an occupancy of 0.1 has been assigned to the Xe atom in the structure (PDB code 3fbx).

3.4. Model building

Starting with an initial model from *ARP/wARP* (Perrakis *et al.*, 1999), the structure of the 66.3 kDa protein was completed manually by cycling between *Coot* and *REFMAC*5. 520 residues (Val63–Thr238 and Cys249–Pro592) were included in the final structure, which was refined to *R* factors of $R_{\rm work}$ = 15.6% and $R_{\rm free}$ = 19.8%. In the Ramachandran plot, 90.7% of the residues are located in the core region and 8.9% lie in allowed regions, while only 0.4% of the residues belong to the generously allowed region. There are no residues in the disallowed region. Stereochemical analysis with *PROCHECK* (Laskowski *et al.*, 1993) detected *cis*-peptide conformations for two prolines (Pro502 and Pro592) and for three nonproline residues: Gly76, Gly155 and Asp316. For the first N-terminal residues Leu47–Pro62, the last C-terminal residues Trp593 and Asp594 and the 11 residues of the C-terminal affinity tag, no electron density was visible in the final map. Amino acids Asn239–Ser248 were also omitted from the model owing to a lack of unambiguous density. However, there is some remaining electron density near Arg531 which might represent some of the missing residues. At the five putative glycosylation sites of the 66.3 kDa protein (Deuschl *et al.*, 2006), Asn93, Asn115, Asn236, Asn441 and Asn520, some additional density is present that indicates the presence of *N*-acetylglucosamine (NAG) moieties attached to the asparagine side chains. Of these modifications, two NAG moieties at Asn115 as well as one NAG moiety each at Asn236, Asn441 and Asn520 were clearly defined in the electron-density map and thus were included in the final model. The electron density for the second sugar moiety at Asn115 was not directly visible after density modification (Fig. 2*c*) but improved during refinement. The 66.3 kDa protein is a rigid structure with 17 β-strands, 13 α-helices and six 3_{10}-helices. The β-strands are arranged in two stacked antiparallel β-sheets, which are flanked by helices on both sides.

4. Conclusions

In this work, the lysosomal 66.3 kDa protein from mouse was crystallized in the monoclinic space group *C*2 and its structure was determined by means of sulfur/Xe SAD phasing. However, the contribution of the Xe atom to the overall scattering and therefore to the phasing power was not required for successful phase determination. The structure obtained is to our knowledge one of the largest structures that has been shown to be feasible for structure determination by sulfur SAD to date and belongs to a small group of proteins crystallized in monoclinic space groups that have been solved successfully using this method.

During phase improvement in *autoSHARP*, different values of the solvent content were systematically tested in steps of 3.0% using *SOLOMON* (Abrahams & Leslie, 1996). Valuable phase information for determining the 66.3 kDa protein structure was only provided for a solvent content of 56.4%. Thus, as previously stated by Watanabe *et al.* (2005), the solvent content and consequently density-improvement procedures such as solvent flattening and solvent flipping seem to play an essential role in sulfur SAD phasing.

Especially for the 66.3 kDa protein, the use of the anomalous signal of intrinsic S atoms was a valuable alternative to standard experimental phasing procedures for the following reasons. The expression system, a human fibrosarcoma cell line, allowed neither a high yield, which is required for extensive screening of heavy-atom derivatives, nor efficient incorporation of selenomethionine. The latter issue may also have been the reason that we could not obtain crystals in conditions containing the appropriately modified 66.3 kDa protein.

To the best of our knowledge, the only structure of similar molecular weight that has been solved by sulfur SAD phasing is that of the 69 kDa protein TT0570 from *T. thermophilus* (Watanabe *et al.*, 2005). In contrast to the procedure described here, Watanabe and coworkers did not use synchroton radiation at a wavelength of 1.9000 Å and a standard loop and mounting system, but applied longer wavelength Cr $K\alpha$ radiation and a recently developed mounting technique that reduces the absorption by the cryobuffer and cryoloop (Kitago *et al.*, 2005). The most important difference concerns the respective space group: the 66.3 kDa protein formed monoclinic crystals, whereas TT0570 crystallized in $P2_12_12$, a space group of higher symmetry. The challenge in the use of sulfur SAD phasing for monoclinic or even triclinic crystals also became obvious in a broad study of 23 different crystal forms by Mueller-Dieckmann *et al.* (2007). In this study, neither the structures of three monoclinic crystal forms nor that of a triclinic crystal form could be solved automatically by sulfur SAD, while submission to the *AutoRickshaw* pipeline

(Panjikar *et al.*, 2005) was successful for the majority of the higher symmetry examples. The determination of the 66.3 kDa protein structure is an extraordinary example of successful sulfur SAD phasing in that the protein is not only larger than most structures solved by this method so far, but was also crystallized in a low-symmetry space group (*C*2). Hence, this work is encouraging for the wider application of this experimental phasing procedure, since it uses only the anomalous signal of intrinsic protein atoms, obviating the need for crystal derivatization with heavy atoms.

The diffraction experiments were carried out on beamline BL14.2 of BESSY and Freie Universität Berlin at BESSY. We thank Georg Zocher for excellent help during data collection. This work was supported by Deutsche Forschungsgemeinschaft Grant LU 1173/1-4 to TL.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Brown, J., Esnouf, R. M., Jones, M. A., Linnell, J., Harlos, K., Hassan, A. B. & Jones, E. Y. (2002). *EMBO J.* **21**, 1054–1062.

Capasso, R., Izzo, A. A., Fezza, F., Pinto, A., Capasso, F., Mascolo, N. & Di Marzo, V. (2001). *Br. J. Pharmacol.* **134**, 945–950.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cravatt, B. F., Demarest, K., Patricelli, M. P., Bracey, M. H., Giang, D. K., Martin, B. R. & Lichtman, A. H. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 9371–9376.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.

Debreczeni, J. É., Bunkóczi, G., Girmann, B. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 393–395.

Debreczeni, J. É., Bunkóczi, G., Ma, Q., Blaser, H. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 688–696.

Debreczeni, J. É., Girmann, B., Zeeck, A., Krätzner, R. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 2125–2132.

Deuschl, F., Kollmann, K., von Figura, K. & Lubke, T. (2006). *FEBS Lett.* **580**, 5747–5752.

Djinović Carugo, K., Helliwell, J. R., Stuhrmann, H. & Weiss, M. S. (2005). *J. Synchrotron Rad.* **12**, 410–419.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Fehrenbacher, N. & Jaattela, M. (2005). *Cancer Res.* **65**, 2993–2995.

Feulner, J. A., Lu, M., Shelton, J. M., Zhang, M., Richardson, J. A. & Munford, R. S. (2004). *Infect. Immun.* **72**, 3171–3178.

Garcia, M., Platet, N., Liaudet, E., Laurent, V., Derocq, D., Brouillet, J. P. & Rochefort, H. (1996). *Stem Cells*, **14**, 642–650.

Gordon, E. J., Leonard, G. A., McSweeney, S. & Zagalsky, P. F. (2001). *Acta Cryst.* D**57**, 1230–1237.

Hansen, H. S., Moesgaard, B., Hansen, H. H. & Petersen, G. (2000). *Chem. Phys. Lipids*, **108**, 135–150.

Hasilik, A. (1992). *Experimentia*, **48**, 130–151.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

Izzo, A. A., Fezza, F., Capasso, R., Bisogno, T., Pinto, L., Iuvone, T., Esposito, G., Mascolo, N., Di Marzo, V. & Capasso, F. (2001). *Br. J. Pharmacol.* **134**, 563–570.

Jensen, A. G., Chemali, M., Chapel, A., Kieffer-Jaquinod, S., Jadot, M., Garin, J. & Journet, A. (2007). *Biochem. J.* **402**, 449–458.

Kitago, Y., Watanabe, N. & Tanaka, I. (2005). *Acta Cryst.* D**61**, 1013–1021.

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.

Kollmann, K., Mutenda, K. E., Balleininger, M., Eckermann, E., von Figura, K., Schmidt, B. & Lubke, T. (2005). *Proteomics*, **5**, 3966–3978.

Kos, J. & Lah, T. T. (1998). *Oncol. Rep.* **5**, 1349–1361.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Liu, Z.-J., Vysotski, E. S., Chen, C.-J., Rose, J. P., Lee, J. & Wang, B.-C. (2000). *Protein Sci.* **9**, 2085–2093.

Lübke, T., Lobel, P. & Sleat, D. E. (2008). *Biochim. Biophys. Acta*, doi:10.1016/j.bbamcr.2008.09.018.

Mueller-Dieckmann, C., Panjikar, S., Schmidt, A., Mueller, S., Kuper, J., Geerlof, A., Wilmanns, M., Singh, R. K., Tucker, P. A. & Weiss, M. S. (2007). *Acta Cryst.* D**63**, 366–380.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Nixon, R. A. & Cataldo, A. M. (2006). *J. Alzheimers Dis.* **9**, 277–289.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A. & Noble, M. (2004). *Acta Cryst.* D**60**, 2288–2294.

Ramagopal, U. A., Dauter, M. & Dauter, Z. (2003). *Acta Cryst.* D**59**, 1020–1027.

Roeser, D., Dickmanns, A., Gasow, K. & Rudolph, M. G. (2005). *Acta Cryst.* D**61**, 1057–1066.

Saftig, P., Hetman, M., Schmahl, W., Weber, K., Heine, L., Mossmann, H., Koster, A., Hess, B., Evers, M., von Figura, K. & Peters, C. (1995). *EMBO J.* **14**, 3599–3608.

Scriver, C. R., Sly, W. S, Childs, B., Beaudet, A. L., Kinzler, K. W. & Vogelstein, B. (2001). Editors. *The Metabolic and Molecular Bases of Inherited Disease*, 8th ed. New York: McGraw–Hill.

Sekar, K., Rajakannan, V., Velmurugan, D., Yamane, T., Thirumurugan, R., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1586–1590.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Sleat, D. E., Della Valle, M. C., Zheng, H., Moore, D. F. & Lobel, P. (2008). *J. Proteome Res.* **7**, 3010–3021.

Sleat, D. E., Jadot, M. & Lobel, P. (2007). *Proteomics Clin. Appl.* **1**, 1134–1146.

Sleat, D. E., Wang, Y., Sohar, I., Lackland, H., Li, Y., Li, H., Zheng, H. & Lobel, P. (2006). *Mol. Cell. Proteomics*, **5**, 1942–1956.

Sleat, D. E., Zheng, H., Qian, M. & Lobel, P. (2006). *Mol. Cell. Proteomics*, **5**, 686–701.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

Watanabe, N. (2006). *Acta Cryst.* D**62**, 891–896.

Watanabe, N., Kitago, Y., Tanaka, I., Wang, J., Gu, Y., Zheng, C. & Fan, H. (2005). *Acta Cryst.* D**61**, 1533–1540.

Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.

Weiss, M. S., Mander, G., Hedderich, R., Diederichs, K., Ermler, U. & Warkentin, E. (2004). *Acta Cryst.* D**60**, 686–695.

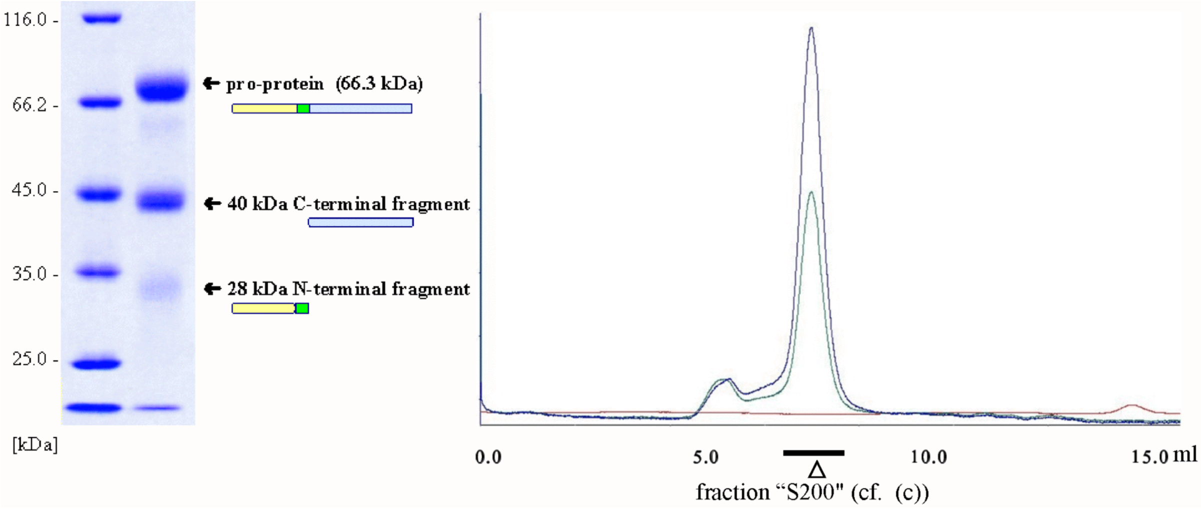
Weiss, M. S., Sicker, T., Djinovic-Carugo, K. & Hilgenfeld, R. (2001). *Acta Cryst.* D**57**, 689–695.

Yang, C. & Pflugrath, J. W. (2001). *Acta Cryst.* D**57**, 1480–1490.

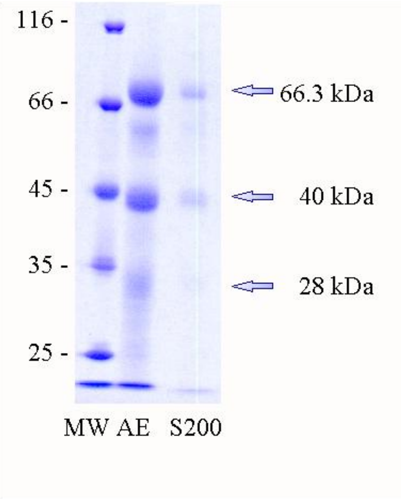
3.3. SUPPLEMENTARY MATERIAL

The data presented in the following were not included as supplementary material in the publication, but enable a more detailed insight into the structure determination procedure and the obtained results.

After the anion exchange chromatography step, the 66.3 kDa protein was purified to homogeneity, but still comprised the one chain 66.3 kDa form and the two chain form composed of the 28 kDa and 40 kDa fragment. The addition of gel filtration as the final purification step was intended to separate the three polypeptides from each other. Surprisingly, they eluted in a single peak (Fig. I-7). Nevertheless, size exclusion chromatography was required to obtain well diffracting crystals. Putative reasons for this optimization are discussed in chapter 5.



(a) (b)



(c)

Fig. I-7. Final purification step of the 66.3 kDa protein (size exclusion chromatography). The three comprised polypeptides are indicated by arrows. (a) SDS-PAGE analysis of the concentrated solution of the 66.3 kDa protein after anion exchange chromatography (AE), which was used for initial crystallization trials. (b) Chromatogram of the subsequently added gel filtration step (Superdex S200 10/300, GE Healthcare, 10 mM Tris/HCl pH 8.0). Blue and green lines represent the extinction at a wavelength of 280 nm and 260 nm, respectively. The red line indicates the conductivity. The elution volume is given in ml. (c) SDS-PAGE analysis of the concentrated solution of the 66.3 kDa protein after size exclusion chromatography (lane "S200", black bar in (b)) in comparison with the solution loaded onto the column (lane "AE"). MW = Molecular weight marker (Fermentas), the MW is indicated in kDa.

A crystal obtained from the 66.3 kDa protein that had been purified by this additional gel filtration step was used for the sulfur SAD phasing described in chapter 3. For this experiment beam stop and detector were set as close as possible in order to reduce absorption by air. The latter is a major challenge at long wavelengths since it reduces the beam intensity significantly. Moreover, radical formation causing secondary radiation damage represents a problem at long wavelengths. Radiation damage increased significantly during data collection as became obvious in the diffraction patterns (Fig. I-8), the R factor for each image as well as the overall R factor and later - during structure refinement - in negative peaks in the F_oF_c difference map for side chain carboxyl groups.

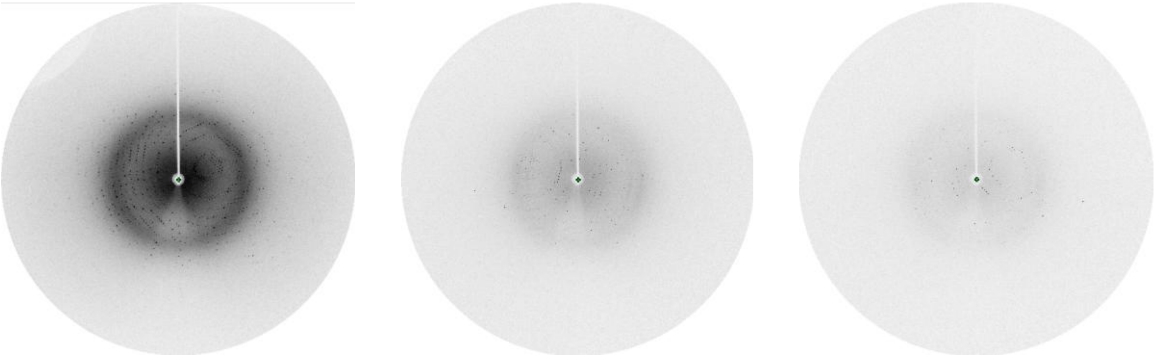
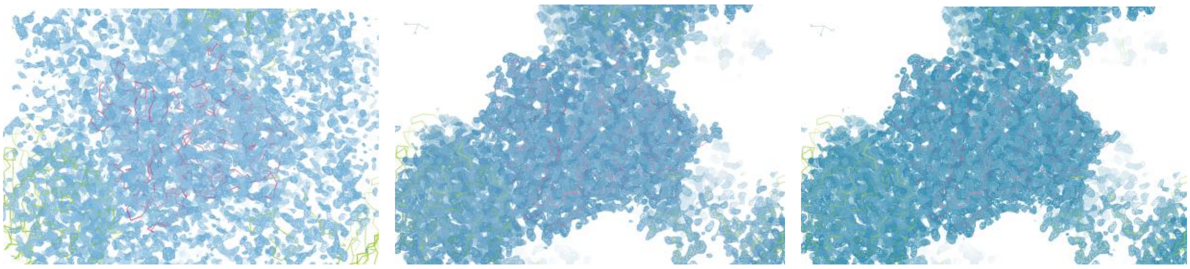


Fig. I-8. Diffraction images taken in the course of the sulfur SAD phasing experiment. From left to right: images 1, 1120, 1440 with decreasing number and intensity of the spots due to an increase in radiation damage. The shadow in the upper left corner is caused by the beam stop.

In order to reduce the effects of radiation damage such as a decreased signal to noise ratio, while still obtaining the required multiplicity of reflections, only the first 1120 images were used for data processing. In general, to reduce radiation damage, the highest dose should be used, since else than expected, the higher the radiation dose, the less is the radiation damage. However, the synchrotron beamline on which the diffraction data were collected (BL-14.2) required a minimum exposure time of 3 seconds in order to average out fluctuations in the X-ray beam. Thus, an aluminium foil was used for reduction of the X-ray beam intensity during data collection with 3.2 seconds exposure time per image. Altogether, these data show that the importance of a data collection strategy which takes into account spot intensities, resolution and radiation damage cannot be overvalued in order to find the best compromise between all these parameters.

For the successful determination of the 66.3 kDa protein structure, phase improvement by solvent flattening and histogram matching turned out to have predominantly important roles. Electron density maps before and after these procedures are represented in Fig. I-9 together with the initial structure obtained by automatic model building.



(a) (b) (c)

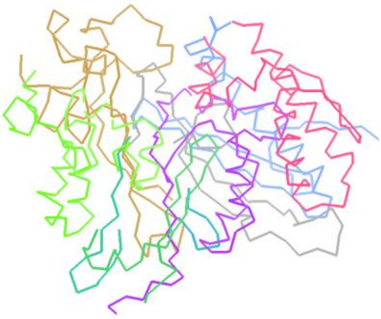
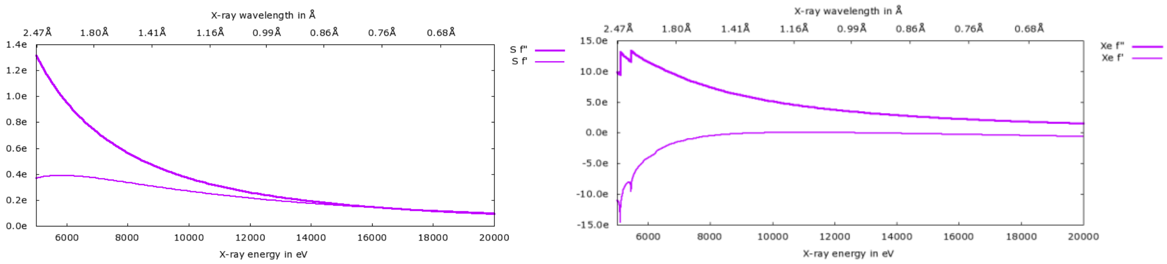


Fig. I-9. Electron density maps before (a) and after phase improvement by solvent flattening (b) and histogram matching (c). The comparison of the maps shows how the boundary between protein molecules and solvent channels became obvious during the structure determination procedure. The final structure is shown as a C_α trace and coloured in pink and green for the single molecule in the asu and its symmetry equivalents, respectively. (d) Initial model obtained by automatic model building as a C_α trace and coloured according to the comprised polypeptide chains.

(d)

Subsequent to successful phase determination, the contributions of 21 sulfur atoms and a single xenon atom of the heavy atom substructure were compared.



(a) S (b) Xe

Fig. I-10. Plots of theoretical f' and f'' values for sulfur (left, see also Fig. I-6) and xenon (right). (http://skuld.bmsc.washington.edu/scatter/AS_form.html).

The huge difference of sulfur and xenon atoms in the anomalous scattering factors f' and f'' at long wavelengths becomes obvious when comparing the respective plots (Fig. I-10). In contrast to sulfur, xenon exhibits two absorption edges. However, the corresponding wavelength cannot easily be reached at most synchrotron beamlines and thus xenon is not feasible for MAD phasing (Multiple Anomalous Dispersion). A gas for soaking crystals which is suitable for this purpose is crypton.

4. INITIAL INSIGHT INTO THE FUNCTION OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE BY MEANS OF X-RAY CRYSTALLOGRAPHY

4.1. OBJECTIVES AND AUTHORS` CONTRIBUTIONS

In a follow-up to the previously described publication, this work (submitted manuscript) reports on the structural analysis of the 66.3 kDa structure. It includes the refinement of the 2.4 Å structure obtained by sulfur SAD phasing (cf. chapter 3) to a higher resolution (1.8 Å) enabling the placing of four additional amino acid residues as well as of two structures derived from additional data sets which were collected using further crystals. The three structures were deposited in the Protein Data Bank (PDB) with the codes 3FGR, 3FGT and 3FGW. They represent different steps of the maturation process of the 66.3 kDa protein and thus were compared in detail. Searching the PDB for structural homologues revealed several proteins with significant similarity, which all belong to the same superfamily. Due to expanded analogy to these enzymes, the 66.3 kDa protein could be classified as an N-terminal nucleophile (Ntn) hydrolase as well. Based on these results, a putative hydrolytic activity on non-peptidic amide bonds and potential substrates as well as a presumable autoproteolytic mechanism of enzyme activation were derived and are outlined in this manuscript. Following biochemical experiments, in particular an activity assay on N-acylethanolamines as potential substrates, are currently performed in the laboratory of Prof. Dr. Torben Lübke (Center of Biochemistry and Molecular Cell Biology, Department Biochemistry II, GZMB, Georg-August University Göttingen).

As for the publication regarding the structure determination of the 66.3 kDa protein (chapter 3), pure protein was produced in the laboratory of Prof. Dr. Torben Lübke up to the anion exchange chromatography step. Subsequently, I performed procedures required for crystal growth such as a final purification step (gel filtration) and protein concentration, crystallization of the different maturation states of the 66.3 kDa protein, data collection and processing. My further contributions with Prof. Dr. Ralf Ficner as my supervisor and support by Dr. Achim Dickmanns (Department of Molecular Structural Biology, Institute of Microbiology and Genetics, GZMB, Georg-August University Göttingen) comprise the structure determination, refinement including PDB depositions and analysis as well as detailed comparison of the three 66.3 kDa structures with each other and with structural homologues. After classification of the 66.3 kDa protein as an N-terminal nucleophile hydrolase and based on the detection of significant similarity in particular to two bacterial acylases, the hypothesis concerning putative substrates initially was conceived by Prof. Dr. Torben Lübke and subsequently revised and surveyed and discussed in detail by all contributors.

4.2. SUBMITTED MANUSCRIPT "INITIAL INSIGHT INTO THE FUNCTION OF THE LYSOSOMAL 66.3 KDA PROTEIN FROM MOUSE BY MEANS OF X-RAY CRYSTALLOGRAPHY "

Initial insight into the function of the lysosomal 66.3 kDa protein from mouse by means of X-ray crystallography

Journal:	*Biochemistry*
Manuscript ID:	draft
Manuscript Type:	Article
Date Submitted by the Author:	
Complete List of Authors:	Ficner, Ralf; Institute of Microbiology and Genetics, Georg-August University of Goettingen, Molecular Structural Biology

scholarONE™
Manuscript Central

Initial insight into the function of the lysosomal 66.3 kDa protein from mouse by means of X-ray crystallography

Kristina Lakomek[1], *Achim Dickmanns*[1], *Matthias Kettwig*[2], *Ralf Ficner*[1]*, *Torben Luebke*[2]

[1] Department of Molecular Structural Biology, Institute of Microbiology and Genetics,
GZMB, Georg-August University Goettingen, Justus-von-Liebig-Weg 11, D-37077
Goettingen, Germany

[2] Center of Biochemistry and Molecular Cell Biology, Department Biochemistry II, Georg-
August University Goettingen, Heinrich-Dueker-Weg 12, D-37073 Goettingen, Germany

rficner@gwdg.de

* corresponding author:

tel: +49 551 3914071

fax: +49 551 3914082.

e-mail-address: rficner@gwdg.de

ABSTRACT

The lysosomal 66.3 kDa protein from mouse is a soluble, mannose 6-phosphate containing protein of so far unknown function. It is synthesized as a glycosylated 75 kDa precursor that undergoes limited proteolysis leading to a 28 kDa N-terminal fragment and a 40 kDa C-terminal fragment. In order to gain insight into its function, the glycosylated 66.3 kDa protein from mouse was crystallized, and its structure was solved by means of sulphur SAD phasing. Here, we report three different crystal structures of the 66.3 kDa protein representing different states of the maturation process. These structures demonstrate that the fragments of the proteolytic cleavage process remain associated. The crystal structures reveal a significant similarity of the 66.3 kDa protein to several bacterial hydrolases. The core αββα sandwich fold and a cysteine residue at the N-terminus of the 40 kDa fragment (C249) classify the 66.3 kDa protein as a member of the structurally defined N-terminal nucleophile (Ntn) hydrolase superfamily. The 66.3 kDa protein closely resembles conjugated bile acid hydrolase (CBAH) and penicillin V acylase (PVA) suggesting a hydrolytic activity on non-peptide amide bonds. The similarity to these bacterial proteins also implies an autocatalytic maturation of the lysosomal 66.3 kDa protein. Upon cleavage between S248 and C249, a deep pocket becomes solvent accessible which harbors the putative active site of the 66.3 kDa protein.

KEYWORDS

66.3 kDa protein, lysosomal degradation, N-terminal nucleophile hydrolase, N-acylethanolamines, structure-function relationship, crystal structure

INTRODUCTION

In order to spatially separate the vast number of divergent reactions carried out by intracellular enzymes, eukaryotic cells are compartmentalized into several membrane-bound organelles. Among these organelles, the lysosomal compartment contains more than 50 hydrolases required for degradation of macromolecules or even whole organelles entering the lysosome by endocytotic or autophagic pathways (*1, 2*); reviewed in (*3*).

This degradation process and thus the hydrolases involved are essential for the cell as reflected by the manifestation of severe diseases caused by the accumulation of undigested substrates in the lysosome due to the lack of hydrolytic enzyme activities. The associated pathogenic phenotypes are collectively referred to as “lysosomal storage diseases” (reviewed in (*4*)). However, the lysosomal compartment does not only serve as a digestive compartment but also plays a key role in many other cellular processes like modulation of peptide hormones and bioactive lipids, tissue homeostasis, inflammation (*5-8*) as well as neuroprotection (*9*). Furthermore, lysosomes are involved in the pathogenesis of Alzheimer disease (*10*), autoimmune diseases and in the initiation and progression of cancer (*11*).

Recently, several proteome studies of the lysosomal compartment have identified a considerable set of novel lysosomal proteins. Most of these sub-proteomic studies took advantage of a specific carbohydrate modification, the mannose 6-phosphate residue (M6P), of newly synthesized soluble lysosomal proteins (*1, 2, 12-16*) as reviewed in (*3*). *In vivo*, M6P-containing proteins are recognized by mannose 6-phosphate receptors (MPRs) at the trans-Golgi network (TGN) and transported to endosomes, where the receptor-ligand complex dissociates due to the acidic pH. Finally, the M6P-containing proteins are delivered to lysosomes, while the MPRs return to the TGN. In most lysosomal sub-proteome analyses, M6P-containing proteins were purified by affinity chromatography on immobilized MPRs and subsequently analyzed by mass spectrometry-based techniques.

One novel protein that was identified in tissues derived from mouse, rat and human was referred to as hypothetical 66.3 kDa protein (*14-16*).

Subsequently, the murine 66.3 kDa protein (*17*) and its human ortholog p76 (*18*) were characterized in more detail regarding their lysosomal localization, processing and glycosylation status. The maturation of the orthologs from mouse and human includes both, limited proteolysis and the usage of all five and six potential N-glycosylation sites, respectively. The murine 66.3 kDa protein is synthesized as glycosylated preproprotein of about 75 kDa in apparent molecular mass. After the co-translational removal of the N-terminal signal peptide, the remaining proprotein is sorted to the lysosomal compartment and matures into a 28 kDa N-terminal fragment and a 40 kDa C-terminal fragment (*17*) which might be further processed into a 25 kDa N- and a 15 kDa C-terminal fragment (unpublished data T.L.). A similar processing was described for the human ortholog p76 resulting in a 32 kDa N-terminal fragment and a 45 kDa C-terminal fragment (*18*). The same authors suggested an additional maturation step for the 40 kDa fragment from mouse into a C-terminal 27 kDa fragment. Such a limited proteolysis in the endosomal / lysosomal compartment is a common hallmark of lysosomal hydrolases and a prerequisite to their hydrolytic activation (*19*).

The 66.3 kDa protein is conserved among vertebrates and shows homology to the Lamina ancestor precursor of *Drosophila melanogaster* (*20*) (29 % identity for 416 aligned residues), the ribonuclease P protein subunit p30 of *Entamoeba histolytica* (*21*) (30 % for 349 aligned residues from C249 to I505), phospholipase B from *Dictyostelium discoideum* (*22*) (39% identity for 518 aligned residues) as well as to the highly glycosylated integral membrane protein p67 from *Trypanosoma brucei* (33% identity for 473 aligned residues) (*17, 18, 23, 24*). The trypanosomal protein p67 has recently been demonstrated to be essential for maintenance of normal lysosomal structure and physiology in bloodstream-stage cells (*25*).

In contrast, no homologuous proteins of the 66.3 kDa protein have been found in yeast and in prokaryotes.

However, since neither bioinformatics analysis nor the detailed characterization of the mouse lysosomal 66.3 kDa protein and its human ortholog p76 have provided any hint regarding the activity and the physiological function we determined the three-dimensional structure of the mouse 66.3 kDa protein. This work demonstrates that the 66.3 kDa protein belongs to the superfamily of N-terminal nucleophile (Ntn) hydrolases. Despite the lack of a significant sequence similarity there is a close resemblance to several bacterial hydrolases regarding the protein fold and residues forming the catalytic centre. Furthermore, a detailed comparison of the three crystal structures of the 66.3 kDa protein reported here with the homologuous structures provide initial insight into its catalytic activity as well as suggest a mechanism of the enzyme`s activation by an autocatalytic proteolysis.

EXPERIMENTAL METHODS

Data collection and structure determination

The glycosylated 66.3 kDa protein from mouse was produced by overexpression in the human fibrosarcoma cell line HT1080 and purified as described (*17*) except for some minor modifications that are summarized below. The protein was crystallized under acidic conditions and the structure was solved at 2.40 Å by means of sulphur SAD (S-SAD) phasing using long wavelength radiation (*25*). The three data sets described in the following had been collected prior to the S-SAD experiment. At a shorter wavelength (0.8141 Å) a data set was collected from the same crystal that was used for the S-SAD (data set “xe1h”, PDB-ID 3FGR) on the BESSY beamline BL-14.2 which was equipped with an SX165 detector (Rayonics LLC, Illinois, USA), and processed with HKL2000 (HKL Research, Inc., Charlottesville, VA). Two additional data sets had been collected previously from a native crystal (data set "native", PDB-ID 3FGT) and from a crystal soaked with potassium iodide (data set "KI", PDB-ID 3FGW). While the protein batch used for the S-SAD phasing was subjected to size exclusion chromatography, the final gel filtration step had not been added to the previously established purification protocol in the cases of the crystals used to collect the data sets „native" and „KI". The crystal, on which the native data set was collected on, was grown under the previously described conditions, whereas the crystal for the KI data set was obtained under slightly different conditions. Instead of 10 mM Tris/HCl pH 8.0 (*25*), the concentrated protein was dissolved in a buffer system of 70 mM NaCl and 5 mM NaH_2PO_4/Na_2HPO_4 pH 7.4. Furthermore, the crystallization drop was composed of 0.7 µl of protein solution (23 mg/ml) and reservoir each (12 % (w/v) PEG 4000, 100 mM NaAc/HAc pH 4.6, 100 mM NH_4Ac). Thus, the final salt concentration was slightly reduced by about 19 % and Tris was exchanged by (di)hydrogen phosphate.

The data sets "native" and "KI" were collected on the DESY beamline X13 (DESY, Hamburg, Germany) which was equipped with a marccd165 detector (Marresearch GmbH, Norderstedt, Germany) and on the BESSY beamline BL-14.1 (BESSY, Berlin, Germany) on a marmosaic225 detector (Marresearch GmbH, Norderstedt, Germany), respectively. The images of both data sets were integrated with XDS (*26*) and scaled using SCALA of the CCP4 program suite (*27*). The iodide soaked crystal diffracted only to 3.2 Å. and severely suffered from radiation damage. However, a 97 % complete data set with a reasonable R factor of $R_{p.i.m.}$ = 7.3 % could be obtained. The three structures derived from the different data sets were solved by means of Molecular Replacement with MOLREP (*28*) using the 2.4 Å structure of the 66.3 kDa protein as a search model (*25*) and manually completed by cycling between REFMAC5 from the CCP4 program suite and COOT (*29*). Data collection and refinement statistics for the three structures are summarized in Tab. 1.

Structure analysis

Four structures of the 66.3 kDa protein have been refined and deposited with the Protein Databank. The structure 3FBX has been solved by SAD and is published elsewhere [25]. This work describes structures of the cleaved forms 3FGR (xe1h) and 3FGT (native) as well as of the "uncleaved form" 3FGW (KI).

The final 1.8 Å structure of crystal form I (PDB-ID 3FGR) includes 524 amino acid residues. While V63-T238 and G245-S248 belong to the polypeptide chain A, C249-P592 form the continuous chain B. Additionally, five N-glycans are included in the final structure. Two N-acetylglucosamine (NAG) moieties are linked to N115 and N441 each, while only one NAG moiety each could be placed at N93, N236 and N520. One xenon atom that had been caught in a hydrophobic pocket during a soak in a xenon gas chamber (*25*) and one sodium ion as well as two acetate anions from the crystallization buffer and eleven glycerol molecules from the cryo protecting solution are included in the solvent model. SIOCS (version 2007/07 alpha_test 0.1; Heisen & Sheldrick, in preparation) was used for prediction of the amide / imidazole orientations of asparagine, glutamine and histidine side chains. The final structure was refined to R factors of R_{work} = 15.2 % and R_{free} = 18.2 % with a FOM (figure of merit) of 0.90. The stereochemical analysis of the refined structure with PROCHECK (*30*) detected two proline residues (P502 and P592) as well as one aspartate residue (D316) to exhibit a *cis* peptide conformation and six residues with torsion angles outside the expected Ramachandran regions (M275, S306, N394, R401, Y431 and H577).

In contrast to the structure 3FGR, the native 2.4 Å structure 3FGT comprises three more residues at the N-terminus (D60-P62) and one additional residue in the intermediate region of the sequence, namely N239, but lacks four amino acids at the C-terminus of chain A (G245-S248). Chain B contains the same residues as in 3FGR resulting in altogether 524 amino acids in 3FGT (D60-N239, C249-P592). In the structure 3FGT, four NAG moieties are attached to the residues N115 (2 NAGs), N236 (1 NAG) and N441 (1 NAG), respectively. Thus, compared to the structure 3FGR, the structure 3FGT lacks the NAG residues at N93 and N520 and one NAG of the glycan linked to N115. Three acetate anions as well as five glycerol, one triethylene glycol and two tetraethylene glycol molecules are included in the solvent model of the structure 3FGT.

The structure derived from the KI derivative crystal (PDB-ID 3FGW) includes the residues V63-N239 and G245-P592. It includes the same residues as 3FGR with the following exceptions. Additionally, N239 could be placed, but G245-S248 are connected to C249 in the structure 3FGW. The structure 3FGW contains five NAG moieties and one mannose (MAN) moiety (1 NAG each at N93, N236 and N441 as well as 2 NAGs and 1 MAN at N115). Furthermore, the solvent model of 3FGW comprises four glycerol molecules and nine iodide anions.

Superpositions for the determination of root mean square deviations (r.m.s.d.s) between two structures as well as for graphical comparison were performed with the program SUPERPOSE of the CCP4 program suite using the superposition of specified atoms if possible (for 3FGR, 3FGT and 3FGW) and secondary structure matching for less related structures (e.g. lysosomal AGA). For superposition with the about 330 amino acids containing enzymes PVA and CBAH only chain B of the 66.3 kDa protein (344 aa, 3FGR) was used, while the whole molecule served as the reference for the larger structures of cephalosporin acylase and penicillin G acylase (557 residues). Calculations of the electrostatic surface potential were performed with DELPHI 4.1 (*31*).

Figure preparation

Figures 1a, 2, 3 and 5-7 as well as the supplementary figures S2 and S3 were prepared with PyMOL (*32*), whereas Figures 4 and S4 were prepared with CCP4 MOLECULAR GRAPHICS (*33*). Fig. S5 was prepared with CHEMSKETCH (*34*).

RESULTS AND DISCUSSION

Structure determination

The glycosylated lysosomal 66.3 kDa protein from mouse was produced by overexpression in the human fibrosarcoma cell line HT1080 and purified as described (*17*). Since the 66.3 kDa protein and its proteolytic 28 kDa and 40 kDa fragments could not be separated, the mixture containing all three polypeptide chains was used for crystallization. Two different crystal forms were obtained under acidic conditions close to the physiological pH of the lysosomal compartment. Crystals of form I belong to space group C2 with cell constants a = 148.7 Å, b = 89.6 Å, c =64.8 Å and β = 98.7° and contain one molecule in the asymmetric unit. The crystals of form II exhibit the same space group (C2) with one molecule in the asymmetric unit, but differ in cell parameters (a = 147.1 Å, b = 88.6 Å, c = 73.5 Å and β = 110.9°). The crystal form II was obtained under slightly different conditions concerning the composition of both, the protein and the reservoir solution as described in materials and methods. The 2.4 Å structure 3FBX of the 66.3 kDa protein, which includes the residues 63-238 and 249-592, was previously obtained by means of sulphur SAD phasing (*25*) and revealed the 28 kDa N-terminal and the 40 kDa C-terminal fragments of the cleaved 66.3 kDa protein still to form one globular entity. Now, this crystal structure has been further refined to a resolution of 1.80 Å using another data set (PDB-ID 3FGR).

In the course of solving the crystallographic phase problem, additional data sets were collected which turned out to be of interest as they represent different states of the maturation process of the 66.3 kDa protein. Diffraction data from a native crystal of crystal form I at a resolution of 2.4 Å (3FGT) and from a non-isomorphous, potassium iodide soaked crystal (crystal form II) which diffracted only to 3.2 Å (3FGW), were analyzed in detail. The crystal structures described here were solved by means of Molecular Replacement using the initial structure of the 66.3 kDa protein (3FBX). The crystal packing of crystal forms I and II is quite different regarding the deviations in the length of the c axis (Δ 8.7 Å) as well as the β angle

(Δ 12.2°). While the protein monomers are arranged in a head-to-tail like manner in crystal form I (3FGR, 3FGT), two symmetry equivalent molecules form contacts head-to-head with each other in crystal form II (3FGW).

The three crystal structures were refined to final R factors of R_{work} = 15.2 % and R_{free} = 18.3 % (3FGR), R_{work} = 16.5 % and R_{free} = 21.1 % (3FGT) and R_{work} = 21.6 % and R_{free} = 28.0 % (3FGW). Data collection and refinement statistics are summarized in Table 1.

The final structure 3FGR with the highest resolution of the three described structures, contains 180 residues of the N-terminal 28 kDa fragment and 344 residues of the C-terminal 40 kDa fragment (V63-T238 and G245-S248 in chain A, C249-P592 in chain B) (Tab. 1; Figs. 1, 2, S1). The N-terminal amino acids L47-P62, N239-L244 as well as the last C-terminal residues (W593, D594 and the eleven residues of the C-terminal affinity tag) are missing due to the lack of unambiguously interpretable electron density. However, the structure 3FGR comprises the residues G245-S248 which could not be built in the initial structure (3FBX). Besides the loop residues 73-75, only T238 exhibits an average temperature factor of B > 45 Å^2. The average temperature factor of 24.1 Å^2 for the amino acids of chain A and even only 19.5 Å^2 for that of chain B indicates an overall well defined conformation of the 66.3 kDa protein structure.

Non-interpreted electron density was found at the sulfhydryl group of C249, which is the N-terminal cysteine of the 40 kDa fragment. This sulfhydryl group appears to be partially oxidized (Fig. 3) which could be a consequence of the fact that the 66.3 kDa protein was purified and crystallized in the absence of a reducing agent. The side chain of the N-terminal cysteine is involved in the octahedral coordination of a cation which is additionally bound by the side chains of S246, E328, T330 and Y379 as well as by the main chain carbonyl group of D315. The nature of this metal ion is so far not known. Since sodium acetate was present in the crystallization buffer and due to the absence of a peak in the anomalous electron density maps calculated with diffraction data sets collected at a wavelength of 0.8 Å, 1.7 Å and 1.9 Å (data not shown) as well as the results of fluorescence scans (carried out at BESSY BL14.1, data not shown), it seems likely that a Na^+ cation is bound to the protein. This is further supported by the octahedral coordination and metal-ligand atom distances of 2.73 - 3.14 Å (*35*).

The structure 3FGW refined at 3.2 Å resolution includes the residues V63-N239 as well as G245-P592 in chain A. Thus, it comprises the same residues as the structure 3FGR and additionally N239. However, in contrast to the high resolution structure, the amino acids G245-S248 are directly connected to C249 (Tab. 1; Fig. S1). The protein residues exhibit an average temperature factor of B = 41.3 $Å^2$. The N-terminal amino acids L47-P62, T240-L244 as well as the last C-terminal residues (W593, D594 and the eleven residues of the C-terminal affinity tag) are missing due to the lack of unambiguously interpretable electron density. In contrast to the structure 3FGR, the sulfhydryl group of C249 is not oxidized but it is involved in the coordination of a cation as well, which is also bound by the side chains of D316, E328, T330, Y379 and by the main chain carbonyl groups of G314 and D315 (coordination sphere ≤ 3.9 Å). As outlined for the structure 3FGR, the nature of the metal cation is not known, but it is assumed to be a Na^+ ion.

The structure 3FGT which was refined at a resolution of 2.4 Å, contains the residues D60-N239 of the N-terminal 28 kDa fragment and the residues C249-P592 of the C-terminal 40 kDa fragment assigned to chain A and B, respectively (Tab. 1; Fig. S1). The protein atoms exhibit an average temperature factor of B = 27.3 $Å^2$. The N-terminal amino acids L47-P59, T240-S248 as well as the last C-terminal residues (W593, D594 and the eleven residues of the C-terminal affinity tag) are missing due to the lack of unambiguously interpretable electron density. The structure 3FGT additionally includes three more residues at the N-terminus (D60-P62) and N239 but lacks the amino acids G245-S248 of the intermediate region. As observed in the high resolution structure, the N-terminal cysteine 249 of chain B of the structure 3FGT seems to be partially oxidized. Likewise, the same residues as in 3FGR except for the main chain carbonyl of G314 substituting S246 are involved in the coordination of the putative Na^+ ion which is bound by the side chains of E328, T330 and Y379 as well as by the main chain carbonyl group of D315 (coordination sphere ≤ 3.9 Å).

Differences between the three structures concerning the glycan moieties and the solvent model are described in the methods section.

Overall structure

Both structures obtained from crystal form I (3FGR and 3FGT) contain the cleaved form of the 66.3 kDa protein, which comprises two polypeptide chains corresponding to the 28 kDa and 40 kDa proteolytic fragments (Fig. 2). If not stated otherwise, the structure 3FGR of crystal form I is described in detail below, since it has been refined at the highest resolution.

The compact globular structure shows two closely associated polypeptide chains (Figs. 1, 2) forming 37 hydrogen bonds as well as two salt bridges (K280-E127, R283-D107; 3FGT). The existence of the 28 kDa and 40 kDa chains as one entity is in accordance with the observation that both fragments as well as the uncleaved 66.3 kDa protein elute in a single absorption peak from the affinity column, anion exchange column and gel filtration column during protein purification, respectively. The gel filtration peak corresponds to an apparent molecular weight of about 140 kDa indicating the existence of the 66.3 kDa protein as a stable dimer in solution. Contact areas between symmetry equivalent molecules in the crystals were analyzed with PISA (*36*). In accordance with the results from the gel filtration, the complexation significance score calculated for the 66.3 kDa protein suggests the existence of a stable homodimer.

The N-terminal 28 kDa fragment consists of six α-helices (α1-α6) and four β-strands (ß1-ß4). The 40 kDa C-terminal fragment contains 13 β-strands (ß5-ß17), seven α-helices (α7–

α13) as well as six 3/10-helices (η1 –η6). Both fragments together form an αββα fold. The core is dominated by two highly twisted β-sheets. The six-stranded β-sheet (β-sheet I) is packed tightly against an extended eleven-stranded β-sheet (β-sheet II) (Figs. 1a-b). The α- and 3/10-helices form two layers (α-layer I and II) flanking the central β-sheets on both sides, engulfing the both sides like a horseshoe, leaving one side of the β-sheet solvent accessible. Most strands of the stacked β-sheets forming the central core derive from the 40 kDa fragment (β5-β17). They are slightly tilted against each other with β-strands β5, β6, β14-β17 forming β-sheet I with the topology β14-β5-β6-β15-β16-β17 and β7-β13 in combination with β1-β4 of the 28 kDa fragment building β-sheet II with the topology β2-β1-β3-β4-β7-β8-β9-β10-β11-β12-β13 (Fig. 1). All β-strands are oriented in an anti-parallel fashion except for a break at β7 which is oriented parallel to the preceding β4. The β-strands β1 and β2 partially protrude from the globular structure. Stabilization is achieved by some additional hydrophobic interactions which are mainly formed between the α-helices α4 and α9 and β-strands β4 and β7. Additionally, two intramolecular disulfide bridges are formed between C147 and C157 of the N- as well as between C497 and C500 of the C-terminal fragment (Fig. 2). In contrast, intermolecular disulfide bonds are not observed which is in accordance with the electrophoretic separation of the fragments under non-reducing conditions (*17*).

The crystal structure contains seven N-acetylglucosamine moieties (NAG) in total which are part of five N-glycans at the asparagine residues N93 (1 NAG), N115 (2 NAGs), N236 (1NAG), N441 (2 NAGs) and N520 (1 NAG) (Fig. 2) and are well defined in the electron density map. The glycosylation sites are evenly distributed on the surface of the molecule. The three N-glycosylation sites of the 40 kDa fragment surround a prominent cavity – the putative substrate binding pocket - in close proximity, while the remaining two sites are localized on the opposing side of the protein molecule (Fig. 2).

Differences between the three structures of the 66.3 kDa protein

Superposition of the three refined structures of the 66.3 kDa protein reveals several significant differences in the conformation of four loops of the 28 kDa fragment. These loops connect β1 and β2, β2 and β3, β4 and α1, and α-helices α1 and α2, respectively, and they are involved in intermolecular crystal contacts with symmetry equivalent protein molecules. The overall r.m.s. deviations between the structures 3FGT and 3FGW compared to 3FGR amount to 0.36 Å and 0.41 Å for 520 common Cα atoms (V63-T238 and C249-P592), respectively. The most important difference concerns the peptide bond connecting residues S248 and C249. While in 3FGR and 3FGT there is no covalent bond between S248 and C249, continuous electron density was observed between these residues in 3FGW indicating the uncleaved form of the protein (Fig. 3). Upon cleavage, the conformation of S248 and C249 changes significantly. The incision also causes a rearrangement of S248 leading to an extensive hydrogen bonding network which includes a salt bridge formed between the terminal carboxyl group of S248 and the side chain of R531 (3FGR).

In the uncleaved structure (3FGW) C249 falls into the generously allowed region of the Ramachandran plot and exhibits a *cis* configuration, while after cleavage it is located in the core region of the Ramachandran plot corresponding to β-strand conformation. In analogy to other auto-proteolytically cleaved enzymes (*37*), this strong distortion most likely helps in providing the potential required for the proteolytic cleavage (see below). The cleavage is additionally accompanied by slight changes of the torsion angles of the adjacent residues S248 and S250 which also are within the allowed β region before and in the core region of the Ramachandran plot after the cleavage (Fig. 3).

The proximate residues T240 – L244 appear to be flexible or absent in all three structures, while the adjacent amino acids G245 - S248 exhibit an ordered conformation in the structures 3FGR and 3FGW, but are disordered or absent in 3FGT (Fig. 4). Interestingly, this loop adopts quite different conformations in 3FGR and 3FGW (Figs. 4, 5). In 3FGR the residues

G245 – S248 are oriented perpendicular to the first β-strand of the 40 kDa fragment (β5), whereas they extend this β-strand in 3FGW, even though the β-strand secondary structure is significantly distorted. In the structure 3FGT, residues G245 to S248 are not defined in the electron density map. Due to the lack or disorder of these residues, a large pocket with a highly negative surface potential becomes solvent accessible (Figs. 4, 5, S2). This cavity emerged to have a putative important role for the function of the 66.3 kDa protein (see below).

Structurally related proteins

In order to obtain insight into the function of the lysosomal 66.3 kDa protein, the Protein Data Bank (PDB) was searched for structurally related proteins with known function. The retrieval using the program DALI (*38*) revealed significant similarities to cephalosporin acylase (CA) (*39*) (Fig. 6), two different kinds of penicillin acylase (penicillin acylase G (PGA) (*40*) and V (PVA) (*41*)), as well as the conjugated bile acid hydrolase (CBAH) (*42*). For these four bacterial proteins the number of the structurally equivalent residues is in the range from 222 (PVA) to 360 (CA) with regard to 520 amino acids of the 66.3 kDa protein. The r.m.s. deviations for the positions of aligned Cα atoms amount to 3.0 Å (PVA) - 3.6 Å (CA). Furthermore, some less similarity was found to inosine monophosphate (IMP) cyclohydrolase (IMPC) (*43*) and proteasome subunits (*44, 45*) (for details see Tab. S1). Interestingly, only a few of the aligned residues are conserved between the 66.3 kDa protein and the structurally related proteins. Merely 6% (PVA, CBAH) to 14% (IMPC) of the structurally equivalent amino acids are identical. Superpositions of the 66.3 kDa protein and these four acylases are shown in Figs. 6 and S3. All structures exhibit a highly similar central overall fold with the highest degree of similarity concerning the formation and orientation of the β-sheet core, while the arrangement of the surrounding α-helices differs.

Although most of these enzymes lack significant sequence similarity among each other, they belong to the superfamily of Ntn hydrolases which is defined by a common fold. The characteristic structural motif is a four-layered αββα sandwich (*46, 47*) (Fig. 1). Based on the crystal structure, the 66.3 kDa protein can be assigned unambiguously to this superfamily.

The PDB contains the crystal structure of another lysosomal Ntn hydrolase, namely that of aspartylglucosaminidase (AGA) (*48*). However, this enzyme has not been revealed by DALI. Using secondary structure matching for the C-terminal fragment only allowed the alignment of 80 residues with an r.m.s.d. of 4.1 Å.

Putative active site

Based on structural homology, the lysosomal 66.3 kDa protein belongs to the superfamily of Ntn hydrolases. Except for IMPC which only shares the common fold but is actually not an Ntn hydrolase in terms of function, all Ntn hydrolases known so far are activated by autocatalytic cleavage. The N-terminal residue generated at the cleavage site represents the canonical catalytic residue attacking the carbonyl carbon of a non-peptide substrate amide bond in a nucleophilic manner. The catalytically essential nucleophile is either threonine, serine or cysteine (such as serine 170 of CA, serine 1β of PGA, threonine 206 of lysosomal AGA and cysteine 2 of CBAH and cysteine 1 of PVA). While the hydroxyl oxygen or the sulphur atom of these N-terminal residues acts as a nucleophile, its α-amino group serves as general base. Based on the superpositions of the 66.3 kDa protein with known Ntn hydrolases (Figs. 6, 7, S4), we suggest C249 at the N-terminus of the 40 kDa fragment to represent the conserved nucleophilic residue. C249 becomes solvent-accessible only after the proteolytic cleavage between S248 and C249 as can be seen by comparison of the structures 3FGW and 3FGR (Figs. 3, 4).

In addition, other known active site residues of Ntn hydrolases are conserved like an asparagine and an arginine residue (Fig. 7). These residues corresponding to N432 and R463 of the 66.3 kDa protein have been shown to be essential in other Ntn hydrolases, e.g. for the catalytic activity of PGA (N241 and R263) (*49, 50*). The Oδ atom of the asparagine residue is hydrogen-bonded to the backbone nitrogen of the N-terminal nucleophilic amino acid in all four Ntn hydrolase structures closely related to the 66.3 kDa protein, while the interaction partner of the side chain nitrogen atom depends on the presence or absence of a ligand. When no substrate or product is bound to the enzyme, the Nδ of the asparagine forms a hydrogen bond with both a side chain nitrogen atom of the conserved arginine (corresponding to R463 of the 66.3 kDa protein) (*87*) and a backbone carbonyl oxygen of a residue located nearby (superposing with T330 of the 66.3 kDa protein) as shown in Fig. 7. Upon ligand binding, the fixation of the asparagine side chain by the guanidinium group of the conserved arginine remains intact, whereas the backbone hydrogen bonding partner is replaced by a functional group of the ligand such as the carboxyl group of the reaction product glutarate bound by CA (*51*).

The important role of the asparagine N432 is additionally stressed by the conservation of two N-terminally adjacent residues among the orthologs of the 66.3 kDa protein, namely S430 and a ubiquitous aromatic amino acid (Y431). The carbonyl oxygen of Y431 forms water mediated hydrogen bonds with a glycerol molecule from the cryo protecting buffer and with the side chain of R463 which is positioned directly above the Y431-N432 peptide bond and is involved in an extended hydrogen bonding network to the catalytically active nucleophilic C249 via N432.

Another residue conserved in the active site of Ntn hydrolases is either a histidine or an arginine corresponding to H266 of the 66.3 kDa protein (Fig. 7). A histidine occupies this position in some Ntn hydrolases which exhibit an N-terminal cysteine like the 66.3 kDa protein such as glutamine phosphoribosylpyrophosphate (PRPP) amidotransferase (*52*) and glucosamine 6-phosphate synthase (*53, 54*). This positively charged side chain most likely enhances the nucleophilic character of the catalytic N-terminal amino acid by decreasing the pKa value of the N-terminal nucleophilic residue. Thus, its conservation is most likely based rather on the catalytic mechanism than on substrate specificity. Due to the acidic lysosomal environment H266 is protonated and therefore able to take over the role of the arginine.

The backbone nitrogen of T330 and Nδ2 of N432 most likely form the oxyanion hole in the 66.3 kDa protein. A third residue appears to be involved as well, namely W269. Like the structural equivalents, the backbone nitrogen of W269 forms a hydrogen bond with the N-terminal nucleophile. The corresponding residues Qβ23 of PGA and H192 of CA form a second hydrogen bond to the N-terminal amino group or Oδ of the conserved active site asparagine, respectively, via their side chains, and mutation of H192 to serine completely abolished autoproteolysis showing this residue to have an important role not only for the catalytic turnover of a substrate, but also for the activation of CA. W269 is not able to form equivalent interactions. Based on this difference we suggest W269 to be important for the catalytic activity of the 66.3 kDa protein but not essential for its autoproteolytic activation.

Further polar side chains in proximity to the catalytic center suitable for interactions with a putative substrate molecule are delivered by S225, T238, N274 and T378 of the 66.3 kDa protein.

Thus, all active site residues as well as some hydrogen bonding patterns of the Ntn hydrolases CBAH, CA, PVA and PGA are conserved in the 66.3 kDa protein (Fig. 7) suggesting that the same reaction mechanism is applied to hydrolyse a non-peptide amide bond. In contrast, several other amino acids involved in substrate binding do not have functional equivalents. This lack of sequence conservation concerning the binding site is not surprising and has been observed for almost all members of the Ntn hydrolase superfamily (*47*). It reflects the wide variety of substrate molecules despite the similar active site structure.

Putative substrates

The members of the Ntn hydrolase superfamily differ significantly in substrate specificity and in the respective substrate binding pocket. So far, the substrate(s) of the 66.3 kDa protein remain(s) unknown. The structural classification of the 66.3 kDa protein as an Ntn hydrolase and in particular the high similarity to members of the choloylglycine family (CBAH and PVA) suggest an enzymatic function similar to that of lysosomal members of this family such as acid ceramidase (AC) and the NAE - hydrolyzing acid amidase (NAAA). As outlined in the following the 66.3 kDa protein probably is involved in the degradation of specific N-acylethanolamines (NAEs) leading to 2-aminoethanol (ethanolamine) and the corresponding free fatty acids.

NAEs represent a class of tissue hormones (mediators) that are synthesized in a variety of organisms and tissues (*55*), reviewed in (*8, 56, 57*). In mammalia, NAEs normally occur in trace amounts in virtually all kinds of cells, but under pathological conditions tissue NAE levels increase significantly (*8, 58-60*). Tissues under inflammatory conditions and human tumor tissues in general were shown to contain substantially higher amounts of phospholipids, N-acylphosphatidylethanolamines and NAEs. The ethanolamines of long-chain fatty acids arise from their corresponding N-acylphosphatidylethanolamines by the action of phospholipase D (*61*) and act as bioactive molecules. N-palmitoylethanolamine exhibits anti-inflammatory (*62-64*) and neuroprotective (*65*) activity. Immunosuppressive (*66*) and analgesic (*9*) functions have been determined for some NAEs as well. Thus, their spread has to be strictly regulated.

One of the enzymes involved in their degradation is NAAA which is located in lysosomes and belongs to the choloylglycine hydrolase family (*67*), reviewed in (*68, 69*). To date, two other lysosomal members of this family are known, namely aspartylglucosaminidase (AGA) and acid ceramidase AC (*69, 70*). AGA and AC have been well characterized and shown to be defect in aspartylglucosaminuria (*71*) and Farber disease (*72*), respectively. While AGA hydrolases the N-glycosidic bond between oligosaccharides and asparagine, AC acts on the amide bond of ceramides releasing a shingosine moiety and the corresponding fatty acid.

NAAA exhibits NAE-hydrolyzing activity at acidic pH with N-palmitoyl-EA as the best substrate (100 %) and 27 % activity towards N-myristoyl-EA (14:0) but only residual activity (2-4 %) on N-lauroyl- (12:0), N-stearoyl- (18:0) and N-arachidonoyl-EA (20:4). A second NAE degrading enyme specific for NAEs of a different set of chain lengths, the membrane-bound fatty acid amide hydrolase (FAAH), had already been identified several years ago (*73, 74*). In contrast to NAAA, FAAH preferentially hydrolyses the poly-unsaturated endocannabinoid N-arachidonoyl-EA (20:4) (anandamide) and N-linoleoyl-EA (18:2) at neutral pH, while it displays only weak activity on saturated NAEs like N-palmitoyl-EA (16:0) (*75-77*). In contrast to NAAA, FAAH does not belong to the Ntn hydrolase superfamily, but to the amidase signature family. While NAAA is a soluble lysosomal protein, FAAH is a membrane protein of the ER and / or Golgi compartment.

However, the enzyme(s) degrading NAEs like N-lauroyl-EA (12:0), N-stearoyl- (C18:0), N-oleoyl-EA (18:1), N-linoleoyl- (C18:2), N-γ-linolenoyl- (C18:3), N-eicosamonoenoyl- (C20:1), N-homo-γ-linolenoyl- (C20:3) as well as some longer fatty acid EAs (C22:1, C22:6) are unknown. Hence, the 66.3 kDa protein could be involved in the hydrolysis of one or several of these compounds.

The active site residues typical of Ntn hydrolases as well as some amino acids which border the putative substrate binding pocket are conserved between the murine 66.3 kDa protein and p67 from *T. brucei*. The knock down of p67 by RNAi results in abnormal lysosome morphology and finally in high mortality of the bloodstream-stage cells of the African trypanosomes in the mammalian host indicating p67 to play a key role in the structural maintenance of the lysosomal structure. Based on the high degree of similarity between the orthologs from mouse and trypanosomes we suggest the 66.3 kDa protein to fulfill this important, more general function for the integrity of lysosomes as well.

Activation by auto-proteolytic removal of the linker peptide

Activation of Ntn hydrolases requires an auto-proteolytic cleavage resulting in the removal of several amino acids or even a whole polypeptide chain N-terminal of the nucleophilic residue. CA which exhibits the most significant structural similarity to the 66.3 kDa protein, is activated by a multi-step maturation process leading to a two chain form of the protein (*78*). During this maturation, two proteolytic cleavages cause the release of a spacer peptide from the protein, which makes the substrate binding pocket solvent accessible. The lysosomal 66.3 kDa protein bears such a linker peptide most likely comprising amino acids K237 to S248 which connects the 28 kDa fragment and the 40 kDa fragment prior to maturation (Figs. 2, 3, 4). The linker region is highly flexible (Fig. 2), especially the region T240 to L244 which is not defined in any of the determined crystal structures.

Most known Ntn hydrolases (*39, 79*) as well as inteins (*80*) contain a glycine residue adjacent to the nucleophilic amino acid on the N-terminal side. However, in the 66.3 kDa protein, a serine residue (S248) is located N-terminally to C249. A similar exception is found in another lysosomal Ntn hydrolase, human AGA containing an aspartate (D182) at the equivalent position (*81*), as well as in plant asparaginases (*82*). However, in the 66.3 kDa protein, a glycine residue is located two amino acids apart from the catalytic C249 with a serine residue in between. N-terminal of G247 and S248 another glycine-serine pair (G245, S246) probably increases the high flexibility of the linker peptide. In the structure 3FGW, the linker residue range from G245 to S248 which is still covalently bound to C249, seems to exhibit a strongly distorted conformation with the scissile peptide bond between S248 and C249 in *cis* conformation. Upon the first proteolytic cleavage (Fig. S5), the strained conformation is most likely released as becomes obvious in the structure 3FGR (Fig. 3), in which all peptide bonds of the defined part of the linker exhibit *trans* conformation. Although these explanations are still preliminary, especially due to the low resolution of the uncleaved 66.3 kDa protein structure (3FGW), the hypothesis is supported by similar observations for the processes during the autoproteolytic removal of a linker peptide in lysosomal AGA.

Upon the first cleavage between S248 and C249, the C-terminal residues of the 28 kDa fragment probably move to the protein surface as implied by the lack of appropriate density in the structure 3FGT, that presumably represents the last out of the maturation states observed in the crystal structures.

A second autocatalytic cleavage releasing a spacer peptide has been reported for CA (*78, 83*). The glutamic acid residue 159 is required for this cleavage at the N-terminus of the spacer peptide between G160 and D161, which is abolished in the mutant proteins E159Q and E159M. The superposition of CA and the 66.3 kDa protein shows the side chain carboxyl groups of E159 and E153, respectively, to be located similarly, even though they belong to non-equivalent β-strands (Fig. S6). However, a residue feasible to form the oxyanion hole in the 66.3 kDa protein could not be identified. Thus, we suggest the C-terminus of the 28 kDa fragment to be trimmed from residue S248 by lysosomal hydrolases rather than by a second autocatalytic step. Upon cleavage between S248 and C249, the C-terminal residues of the 28 kDa fragment could protrude from the protein and would be accessible for proteases which are quite abundant in the lysosomal compartment. The N-glycan attached to N236 of the 66.3 kDa protein which was shown to be included in the mature 28 kDa fragment (*17*) should protect it against further C-terminal degradation. The enzymatic maturation mechanism would also explain the presence of at least two residues C-terminal of the glycosylated N236 in all structures (Fig. 3), since the crystallized protein had not reached the lysosomal compartment but was secreted by exocytosis. Furthermore, such a mechanism would be in accordance with the lysosomal Ntn hydrolase AGA (*37, 48, 71*). So far, the C-terminal residue of the 28 kDa fragment has not been defined. However, the exact length of the linker might not have any effect on the acylase activity – at least in CA from different *Pseudomonas* species, for which natively occuring variations from 8 to 11 amino acids have been investigated (*83-87*). Alternatively, the C-terminus of the 28 kDa might not even be trimmed at all but move to the protein surface due to a high degree of flexibility. This conformational change already could give substrates sufficient access to the active site for turnover as demonstrated for CA (*83*).

CONCLUSIONS

Three structures of the lysosomal 66.3 kDa protein from mouse were solved by Molecular Replacement and refined to a resolution ranging from 1.8 Å to 3.2 Å (PDB-ID 3FGR, 3FGT, 3FGW). They provide initial insight into its so far unknown function and shed light on its maturation which comprises an autocatalytic cleavage and most likely also a C-terminal processing by other lysosomal hydrolases. Each of the three structures represents a different state of the auto-proteolytic process which gives rise to a 28 kDa N- and a 40 kDa C-terminal fragment.

The major difference between the three structures concerns a linker peptide of about ten amino acids N-terminal of C249. In the uncleaved protein (3FGW), this linker peptide is still covalently connected to C249 and occupies a large cavity. During maturation, the peptide backbone is incised between S248 and C249 (as observed in the structure 3FGR). Subsequently, the flexible linker region moves to the surface of the protein, and a deep pocket becomes accessible (confer structure 3FGT) for the binding of putative substrates.

The structures of the 66.3 kDa protein revealed significant structural similarities but no sequence homology to several bacterial acylases which belong to the N-terminal nucleophile (Ntn) hydrolase superfamily and act on non-peptide amide bonds. Based on this structural homology including both the overall fold and the active site residues the 66.3 kDa protein could be assigned to the superfamily of Ntn hydrolases.

Among the structurally similar bacterial acylases there are two members of the choloylglycine hydrolase family. Representatives of this family have been identified in eukaryotes where at least two members of this family localize to the lysosomal compartment like the 66.3 kDa protein - namely acid ceramidase and NAAA. These structurally related enzymes are involved in the degradation of the amidated lipids ceramides and N-acylethanolamines (NAEs), respectively. NAEs of various chain lengths have been shown to exhibit neuroprotective, anti-inflammatory and immunosuppressive effects as well as to accumulate in e.g. degenerating tissues or tumor tissues. The lysosomal compartment plays a major role in the regulation of the NAE level in the cell, but the degradation of the entire set of the various NAEs cannot be explained completely by the action of the enzymes identified so far. In this context, some of these compounds appear to represent plausible substrates of the 66.3 kDa protein. Certainly, this hypothesis has to be confirmed by further biochemical studies. Currently, a gene trap knockout mouse is under construction and might help to evaluate the physiological function of the 66.3 kDa protein. In case that definite NAEs represent the substrates of the 66.3 kDa protein it is questionable whether a knockout mouse would exhibit a classical lysosomal storage phenotype since NAEs only occur in trace amounts and might be also degraded by other enzymes than the 66.3 kDa protein due to functional redundancy. So far, no disease has been described where NAEs accumulate. However, the comparative analysis of NAE levels in tissues derived from knockout and wild-type mice might help to further narrow down the spectrum of putative substrates.

ACKNOWLEDGEMENTS

We thank Uwe Müller, Jörg Schulz and Georg Zocher from BESSY, Berlin, Germany as well as Paul Tucker from EMBL at DESY, Hamburg, Germany for excellent help during data collection and Florian Deuschl and Katrin Kollmann, University of Goettingen, Germany, as well as Jens Meiler, Vanderbilt University, Nashville, TN, USA for helpful discussions.

SUPPORTING INFORMATION AVAILABLE

Fig. S1. Schematic representation of the amino acid residue ranges comprised by the structures 3FGR, 3FGT and 3FGW.

Fig. S2. Superposition of the 66.3 kDa protein with penicillin V acylase, conjugated bile acid hydrolase and penicillin G acylase.

Fig. S3. Superposition of linker residues and ligands of the 66.3 kDa protein, cephalosporin acylase (CA) and conjugated bile acid hydrolase (CBAH).

Fig. S4. Surface representation of the substrate binding pocket of the 66.3 kDa protein according to its hydrophilic / hydrophobic character.

Fig. S5. Presumable mechanism of the auto-proteolytic cleavage between S248 and C249 during the maturation process of the 66.3 kDa protein.

Tab. S1. Extended list of structures with a similar fold as the 66.3 kDa protein revealed using the program DALI.

FIGURE CAPTIONS

Fig. 1. Overall structure of the 66.3 kDa protein from mouse. (a) The residues are rainbow-coloured according to their position in the polypeptide chain from the N-terminus (blue) to the C-terminus (orange) and represented in cartoon mode with smoothed loops. (b) In the topology diagram, light blue circles represent helices below or above the anti-parallel β-sheets, on either side of them, blue circles show helices which sandwich the sheets, and light blue stars display helical structures in loops above or below the sandwich between both β-sheets. N(A), C(A), N(B) and C(B) mark the N- and C-termini of the 28 kDa and 40 kDa fragment, respectively.

Fig. 2. Cartoon model of the 66.3 kDa protein (3FGR) viewed along the β sheets (at the top) and from the top (after a turn by 90°) (at the bottom). The 28 kDa and 40 kDa fragment are coloured in orange and blue, respectively. The last four C-terminal residues of the 28 kDa fragment (G245-S248) as well as the two intermolecular disulfide bonds are highlighted in ball and stick mode and coloured in orange and brown, respectively. The five glycans and the asparagine residues, at which they are attached, are shown as thick black lines.

Fig. 3. Comparison of the region of the 66.3 kDa protein, which differs significantly between the one chain and two chain variants. The structures are shown in the putative order of the maturation process starting at the top. The most significant differences concern the residue range N239-C249. This range and additionally the adjacent residues K237, T238 and S250 as well as the side chains of E507 and R531 and a glycerol molecule are shown in stick mode with the surrounding electron density of the final $2F_oF_c$ map at a contour level of 1.0 σ (carbon, oxygen and nitrogen atoms in green, red and blue, respectively). The bound Na^+ ion

and the coordinating bonds are shown as a blue sphere and black dashed lines, respectively. Hydrogen bonds involving the side chain of R531 are represented by orange dashed lines, while the green dashed lines indicate the missing residues of the linker region. For orientation, L231-N236 and A251–K254 are shown in cartoon mode.

Fig. 4. Comparison of the solvent accessibility of the substrate binding pocket in the three structures. The residues P60/V63-T238 of the N-terminal and C249-P592 of the C-terminal fragment are shown as orange and blue surfaces. The residues N239-S248 are shown in stick mode (same colour code as in Fig. 3), whereas the coordinated metal ion is represented by a black sphere.

Fig. 5. Electrostatic surface potential of the 66.3 kDa protein. While the overall structures represent the structure 3FGR, the close views show the central region of the structure 3FGW for comparison. In the two images on the left and right side, respectively, the structures are rotated by 90°. The residue ranges V63-T238 and C249-P592 are shown as surfaces and coloured according to their electrostatic potential with positive and negative charges in blue and red, respectively. The residues N239-S248 which occupy significantly different positions in the structures 3FGR and 3FGW are shown in ball and stick mode and coloured in black. The bound Na^+ ion is represented as a yellow sphere.

Fig. 6. Superposition of the 66.3 kDa protein with cephalosporin acylase. The structures of the 66.3 kDa protein (3FGR) and cephalosporin acylase (1OQZ) are shown in cartoon mode and coloured in blue and orange, respectively.

Fig. 7. Superposition of the conserved active site residues of the 66.3 kDa protein and the four most related N-terminal nucleophile hydrolases. The conserved N-terminal nucleophile is shown completely, while of the ubiquitous asparagine and arginine residues as well as of the residue in the lower right corner only the side chains are represented, since the main chain atoms are not directly involved in catalysis. In contrast, concerning the other three residues only the main chain atoms are depicted due to their participation in the catalytic reaction and a lack of sequence conservation. The residues are coloured by atom. Nitrogen, oxygen and sulphur atoms are shown in blue, red and light orange, respectively, for all structures, whereas the carbon atoms are represented distinctly for the various structures as follows: 66.3 kDa protein in grey, cephalosporin acylase (1OQZ) in pink, penicillin V acylase (3PVA) in yellow, conjugated bile acid hydrolase (2BJF) in green, penicillin G acylase (1K5S) in orange.

Fig. S1. Schematic representation of the amino acid residue ranges comprised by the structures 3FGR, 3FGT and 3FGW. The residues of the N-terminal 28 kDa fragment, the linker region and the C-terminal 40 kDa fragment, which are included in each structure, are represented as boxes coloured in yellow, light grey and blue, respectively. The first and the last residue of each region are given in bold letters. The dotted lines represent missing residues of the intermediate region.

Fig. S2. Superposition of the 66.3 kDa protein with penicillin V acylase (PVA), conjugated bile acid hydrolase (CBAH) and penicillin G acylase (PGA). The structures are represented in cartoon mode and coloured in blue (66.3 kDa protein), pink (PVA), red (CBAH) and green (PGA), respectively (see Tab. 2).

Fig. S3. Superposition of linker residues and ligands of the 66.3 kDa protein, cephalosporin acylase (CA) and conjugated bile acid hydrolase (CBAH). The active site residues of the 66.3 kDa protein (3FGR) are represented according to Fig. 7 with the carbon atoms coloured in light grey. The linker residues N239 as well as G245-S248 of the structures 3FGR and 3FGW are shown as black and blue stick model, respectively. They fit well with the linker regions and ligands of the aligned structures of CA and CBAH, which are coloured as follows: glutarate in yellow, 7-β-(4-carboxybutanamido)-cephalosporanic acid in light orange (1JVZ) (*51*), D161-G169 of CA in dark orange (*39*), taurine and deoxycholate in red (*42*).

Fig. S4. Surface representation of the substrate binding pocket of the 66.3 kDa protein according to its hydrophilic / hydrophobic character. The residues V63-T238 as well as C249-P592 of the structure 3FGR are shown in surface representation. Hydrophilic amino acids and glycans are coloured in yellow, whereas hydrophobic residues are shown in grey. The linker residues G245-S248 (3FGR) are shown in stick mode, the coordinated Na^+ ion is represented as a blue sphere.

Fig. S5. Putative mechanism of the auto-proteolytic cleavage between S248 and C249 during the maturation process of the 66.3 kDa protein. Residues of and adjacent to the scissile peptide bond are labeled in blue, while residues of which side chain and backbone atoms are involved in the represented interactions, are labeled in black and grey, respectively. The first nucleophilic attack at the carbonyl carbon of S248 by the sulfhydryl group of C249 and the subsequent formation of the oxyanion are indicated by orange arrows. Possible attacks following this transition state are represented by green and blue arrows depending on whether the oxygen atom is part of the serine side chain or of a bound water molecule.

ACCESSION NUMBERS

The coordinates and structure factors have been deposited in the Protein Structure Databank with the accession numbers 3FGR, 3FGT and 3FGW for the 1.8 Å structure, the 2.4 Å native and the 3.2 Å iodide soaked structure of the lysosomal 66.3 kDa from mouse, respectively.

TABLES

Tab. 1. Summary of crystallographic data.

PDB-ID	**3FGR**	**3FGW**	**3FGT**
data set	xe1h	KI	native
wavelength (Å)	0.91841	1.80000	0.80150
number of images	305	280	406
oscillation steps (°)	0.5	0.5	0.4
space group	C 1 2 1	C 1 2 1	C 1 2 1
cell [Å, °]	148.74 89.56 64.81 β 98.68	147.05 88.62 73.58 β 110.90	145.57 88.22 63.27 β 98.10
resolution range[a] (Å)	50.00-1.70 (1.76-1.70)	46.00-3.24 (3.42-3.24)	30.00-2.40 (2.53-2.40)
completeness (%)	99.5 (96.2)	97.0 (98.1)	99.8 (100.0)
redundancy	3.2 (2.6)	3.0 (2.9)	3.4 (3.4)
unique reflections (rejections)	91,683 (164)	14,178 (418)	31,031 (3,487)
R_{sym}* or $R_{p.i.m.}$[#] (%)	3.3 (41.9)*	7.3 (13.1)#	6.1 (29.4)#
I/sigma	32.1 (2.4)	8.4 (4.4)	9.5 (3.5)
X-ray source	BL-14.2	BL-14.1	X13
Refinement statistics			
amino acids in asu (chain)	524: V63-T238 (A) G245-S248 (A) C249-P592 (B)	525: V63-N239 (A) G245-P592 (A)	524: P60-N239 (A) C249-P592 (B)
molecules in asu	1	1	1
resolution (Å)	29.26-1.80	44.32-3.24	29.49-2.40
R_{work}[e]	15.2	20.4	16.6
R_{free}[f]	18.2	27.5	20.7
number of non-H atoms protein water solvent	 4396 576 78	 4265 20 34	 4275 299 90
rmsd[g] bonds (Å) angles (°)	0.015 1.533	0.009 1.235	0.012 1.493
average B factors	24.3	42.2	28.1

Tab. 2. Comparison of the 66.3 kDa protein with Ntn hydrolases of known structure (DALI). Only hits with a Z-score ≥ 7 and with an assigned cellular function are listed here. A complete list of all revealed similar structures can be found in Tab. S1.

protein	**Abbre-viation**	**PDB-ID***	**Z-score**	**rmsd [Å]**	**L_{ali}**	**N_{res}**	**% ID**
cephalosporin acylase	CA	1oqz	17.0	3.6	360	684	11
penicillin V acylase	PVA	3pva	16.2	3.0	222	334	6
conjugated bile acid (=choloylglycine) hydrolase	CBAH	2bjf	16.2	3.1	224	328	6
penicillin G acylase	PGA	1k5s	15.4	3.4	244	557	11
IMP cyclohydrolase	IMPC	2ntm	8.4	3.2	165	202	14
20 S proteasome	-	1ryp	8.3	3.1	161	205	7

* For redundant proteins, the PDB-ID and the corresponding values are given only for the best hit. Z-score = value for comparison. Hits with Z-scores ≤ 2 are spurious. rmsd = root mean square deviation between the aligned residues, L_{ali}: number of structurally equivalent residues, N_{res}: number of amino acids in the protein, % ID: percentage of identical amino acids over all structurally equivalent residues.

Tab. S1. Extended list of structures with a similar fold as the 66.3 kDa protein revealed using the program DALI.

protein	**PDB-ID***	**Z-score**	**rmsd [Å]**	**L_{ali}**	**N_{res}**	**% ID**
Cephalosporin acylase* (CA)	1oqz	17.0	3.6	360	684	11
Penicillin V acylase (PVA)	2pva	16.2	3.0	222	334	6
Conjugated bile acid (=choloylglycine) hydrolase (CABH)	2bjf	16.2	3.1	224	328	6
Penicillin G acylase (PGA)	1k5s	15.4	3.4	244	557	11
IMP cyclohydrolase (IMPC)	2ntm	8.4	3.2	165	202	14
20 S proteasome	1ryp	8.3	3.1	161	205	7
conserved protein[#]	1kuu	8.2	3.2	161	202	14
Proteasome component Y7	1g0u	8.2	3.0	157	196	8
Proteasome α subunit	1j2q	7.8	3.2	154	202	11
Proteasome α-type subunit 1	2h6j	7.3	4.1	173	242	16
HSLV protease	1g3k	7.2	3.0	141	173	9
Proteasome component C7-α	1z7q	7.2	5.6	167	243	11
ATP-dependent HSL protease ATP-binding subunit	1ofh	7.2	3.1	141	173	9
ATP-dependent HSLU protease ATP-binding subunit	1g3i	7.0	3.4	146	173	9
Protein YPL144W	2z5c	6.4	3.3	139	189	13
UNP Q5LQD5_SILPO (hypothetical protein)	2imh	6.1	4.1	157	226	17
Glutamine PRPP amidotransferase	1gph	2.8	4.3	129	465	10
Horse plasma gelsolin	1d0n	2.8	4.1	109	729	7
Antithrombin III	1att	2.6	7.8	86	420	5

*: cephalosporin acylase = glutarylamidase = glutaryl acylase = glutaryl-7-aminocephalosporanic acid acylase, PRPP = phosphoribosylpyrophosphate, [#]: conserved protein of unknown function (structural genomics)

REFERENCES

(1) Sleat, D. E., Zheng, H., and Lobel, P. (2007) The human urine mannose 6-phosphate glycoproteome. *Biochim Biophys Acta 1774*, 368-372.

(2) Sleat, D. E., Della Valle, M. C., Zheng, H., Moore, D. F., and Lobel, P. (2008) The mannose 6-phosphate glycoprotein proteome. *J Proteome Res 7*, 3010-3021.

(3) Lübke, T., Lobel, P., and Sleat, D. E. (2008) Proteomics of the lysosome. *Biochim Biophys Acta., in press.*

(4) Scriver, C. R., Beaudet, Sly, W.S, A.L., Childs, B., Kinzler, K.W., Vogelstein, B. (eds) (2001) *The Metabolic & Molecular Bases of Inherited Disease*, Vol. III, 8th ed., McGraw-Hill, New York.

(5) Capasso, R., Izzo, A. A., Fezza, F., Pinto, A., Capasso, F., Mascolo, N., and Di Marzo, V. (2001) Inhibitory effect of palmitoylethanolamide on gastrointestinal motility in mice. *Br J Pharmacol 134*, 945-950.

(6) Izzo, A. A., Fezza, F., Capasso, R., Bisogno, T., Pinto, L., Iuvone, T., Esposito, G., Mascolo, N., Di Marzo, V., and Capasso, F. (2001) Cannabinoid CB1-receptor mediated regulation of gastrointestinal motility in mice in a model of intestinal inflammation. *Br J Pharmacol 134*, 563-570.

(7) Feulner, J. A., Lu, M., Shelton, J. M., Zhang, M., Richardson, J. A., and Munford, R. S. (2004) Identification of acyloxyacyl hydrolase, a lipopolysaccharide-detoxifying enzyme, in the murine urinary tract. *Infect Immun 72*, 3171-3178.

(8) Hansen, H. S., Moesgaard, B., Hansen, H. H., and Petersen, G. (2000) N-Acylethanolamines and precursor phospholipids - relation to cell injury. *Chem Phys Lipids 108*, 135-150.

(9) Cravatt, B. F., Demarest, K., Patricelli, M. P., Bracey, M. H., Giang, D. K., Martin, B. R., and Lichtman, A. H. (2001) Supersensitivity to anandamide and enhanced endogenous cannabinoid signaling in mice lacking fatty acid amide hydrolase. *Proc Natl Acad Sci U S A 98*, 9371-9376.

(10) Nixon, R. A., and Cataldo, A. M. (2006) Lysosomal system pathways: genes to neurodegeneration in Alzheimer's disease. *J Alzheimers Dis 9*, 277-289.

(11) Fehrenbacher, N., and Jaattela, M. (2005) Lysosomes as targets for cancer therapy. *Cancer Res 65*, 2993-2995.

(12) Journet, A., Chapel, A., Kieffer, S., Louwagie, M., Luche, S., and Garin, J. (2000) Towards a human repertoire of monocytic lysosomal proteins. *Electrophoresis 21*, 3411-3419.

(13) Journet, A., Chapel, A., Kieffer, S., Roux, F., and Garin, J. (2002) Proteomic analysis of human lysosomes: application to monocytic and breast cancer cells. *Proteomics 2*, 1026-1040.

(14) Kollmann, K., Mutenda, K. E., Balleininger, M., Eckermann, E., von Figura, K., Schmidt, B., and Lubke, T. (2005) Identification of novel lysosomal matrix proteins by proteome analysis. *Proteomics 5*, 3966-3978.

(15) Sleat, D. E., Wang, Y., Sohar, I., Lackland, H., Li, Y., Li, H., Zheng, H., and Lobel, P. (2006) Identification and validation of mannose 6-phosphate glycoproteins in human plasma reveal a wide range of lysosomal and non-lysosomal proteins. *Mol Cell Proteomics 5*, 1942-1956.

(16) Sleat, D. E., Zheng, H., Qian, M., and Lobel, P. (2006) Identification of sites of mannose 6-phosphorylation on lysosomal proteins. *Mol Cell Proteomics 5*, 686-701.

(17) Deuschl, F., Kollmann, K., von Figura, K., and Lubke, T. (2006) Molecular characterization of the hypothetical 66.3 kDa protein in mouse: lysosomal

targeting, glycosylation, processing and tissue distribution. *FEBS Lett 580*, 5747-5752.

(18) Jensen, A. G., Chemali, M., Chapel, A., Kieffer-Jaquinod, S., Jadot, M., Garin, J., and Journet, A. (2007) Biochemical characterization and lysosomal localization of the mannose-6-phosphate protein p76 (hypothetical protein LOC196463). *Biochem J 402*, 449-458.

(19) Hasilik, A. (1992) The early and late processing of lysosomal enzymes: proteolysis and compartmentation. *Experientia 48(2)*, 130-151.

(20) Perez, S. E., and Steller, H. (1996) Molecular and genetic analyses of lama, an evolutionarily conserved gene expressed in the precursors of the Drosophila first optic ganglion. *Mech Dev 59*, 11-27.

(21) Loftus, B., Anderson, I., Davies, R., Alsmark, U. C., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R. P., Mann, B. J., Nozaki, T., Suh, B., Pop, M., Duchene, M., Ackers, J., Tannich, E., Leippe, M., Hofer, M., Bruchhaus, I., Willhoeft, U., Bhattacharya, A., Chillingworth, T., Churcher, C., Hance, Z., Harris, B., Harris, D., Jagels, K., Moule, S., Mungall, K., Ormond, D., Squares, R., Whitehead, S., Quail, M. A., Rabbinowitsch, E., Norbertczak, H., Price, C., Wang, Z., Guillen, N., Gilchrist, C., Stroup, S. E., Bhattacharya, S., Lohia, A., Foster, P. G., Sicheritz-Ponten, T., Weber, C., Singh, U., Mukherjee, C., El-Sayed, N. M., Petri, W. A., Jr., Clark, C. G., Embley, T. M., Barrell, B., Fraser, C. M., and Hall, N. (2005) The genome of the protist parasite Entamoeba histolytica. *Nature 433*, 865-868.

(22) Morgan, C. P., Insall, R., Haynes, L., and Cockcroft, S. (2004) Identification of phospholipase B from Dictyostelium discoideum reveals a new lipase family present in mammals, flies and nematodes, but not yeast. *Biochem J 382*, 441-449.

(23) Kelley, R. J., Alexander, D. L., Cowan, C., Balber, A. E., and Bangs, J. D. (1999) Molecular cloning of p67, a lysosomal membrane glycoprotein from Trypanosoma brucei. *Mol Biochem Parasitol 98*, 17-28.

(24) Alexander, D. L., Schwartz, K. J., Balber, A. E., and Bangs, J. D. (2002) Developmentally regulated trafficking of the lysosomal membrane protein p67 in Trypanosoma brucei. *J Cell Sci 115*, 3253-3263.

(25) Peck, R. F., Shiflett, A. M., Schwartz, K. J., McCann, A., Hajduk, S. L., and Bangs, J. D. (2008) The LAMP-like protein p67 plays an essential role in the lysosome of African trypanosomes. *Mol Microbiol 68*, 933-946.

(26) Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst. 26*, 795-800.

(27) (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr 50*, 760-763.

(28) Vagin, A. A., Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Cryst. 30*, 1022-1025.

(29) Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr 60*, 2126-32.

(30) Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993) Main-chain bond lengths and bond angles in protein structures. *J Mol Biol 231*, 1049-1067.

(31) Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem 23*, 128-137.

(32) DeLano, W. L. (2008).

(33) Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A., and Noble, M. (2004) Developments in the CCP4 molecular-graphics project. *Acta Crystallogr D Biol Crystallogr 60*, 2288-2294.

(34) Advanced Chemistry Development, I., Toronto, ON, Canada. (2007) ACD/ChemSketch Freeware, version 11.00.

(35) Harding, M. M. (2002) Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallogr D Biol Crystallogr 58*, 872-4.

(36) Krissinel, E., and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol 372*, 774-797.

(37) Saarela, J., Oinonen, C., Jalanko, A., Rouvinen, J., and Peltonen, L. (2004) Autoproteolytic activation of human aspartylglucosaminidase. *Biochem J 378*, 363-371.

(38) Holm, L., and Sander, C. (1996) Alignment of three-dimensional protein structures: network server for database searching. *Methods Enzymol 266*, 653-662.

(39) Kim, J. K., Yang, I. S., Rhee, S., Dauter, Z., Lee, Y. S., Park, S. S., and Kim, K. H. (2003) Crystal structures of glutaryl 7-aminocephalosporanic acid acylase: insight into autoproteolytic activation. *Biochemistry 42*, 4084-4093.

(40) Duggleby, H. J., Tolley, S. P., Hill, C. P., Dodson, E. J., Dodson, G., and Moody, P. C. (1995) Penicillin acylase has a single-amino-acid catalytic centre. *Nature 373*, 264-268.

(41) Suresh, C. G., Pundle, A. V., SivaRaman, H., Rao, K. N., Brannigan, J. A., McVey, C. E., Verma, C. S., Dauter, Z., Dodson, E. J., and Dodson, G. G. (1999) Penicillin V acylase crystal structure reveals new Ntn-hydrolase family members. *Nat Struct Biol 6*, 414-416.

(42) Rossocha, M., Schultz-Heienbrok, R., von Moeller, H., Coleman, J. P., and Saenger, W. (2005) Conjugated bile acid hydrolase is a tetrameric N-terminal thiol hydrolase with specific recognition of its cholyl but not of its tauryl product. *Biochemistry 44*, 5739-5748.

(43) Kang, Y. N., Tran, A., White, R. H., and Ealick, S. E. (2007) A novel function for the N-terminal nucleophile hydrolase fold demonstrated by the structure of an archaeal inosine monophosphate cyclohydrolase. *Biochemistry 46*, 5050-5062.

(44) Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H. D., and Huber, R. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature 386*, 463-471.

(45) Hines, J., Groll, M., Fahnestock, M., and Crews, C. M. (2008) Proteasome inhibition by fellutamide B induces nerve growth factor synthesis. *Chem Biol 15*, 501-512.

(46) Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C., Smith, J. L., Tomchick, D. R., and Murzin, A. G. (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature 378*, 416-419.

(47) Oinonen, C., and Rouvinen, J. (2000) Structural comparison of Ntn-hydrolases. *Protein Sci 9*, 2329-2337.

(48) Oinonen, C., Tikkanen, R., Rouvinen, J., and Peltonen, L. (1995) Three-dimensional structure of human lysosomal aspartylglucosaminidase. *Nat Struct Biol 2*, 1102-1108.

(49) McVey, C. E., Walsh, M. A., Dodson, G. G., Wilson, K. S., and Brannigan, J. A. (2001) Crystal structures of penicillin acylase enzyme-substrate complexes: structural insights into the catalytic mechanism. *J Mol Biol 313*, 139-150.

(50) Prabhune, A. A., and Sivaraman, H. (1990) Evidence for involvement of arginyl residue at the catalytic site of penicillin acylase from Escherichia coli. *Biochem Biophys Res Commun 173*, 317-322.

(51) Kim, Y., and Hol, W. G. (2001) Structure of cephalosporin acylase in complex with glutaryl-7-aminocephalosporanic acid and glutarate: insight into the basis of its substrate specificity. *Chem Biol 8*, 1253-1264.

(52) Chen, S., Tomchick, D. R., Wolle, D., Hu, P., Smith, J. L., Switzer, R. L., and Zalkin, H. (1997) Mechanism of the synergistic end-product regulation of Bacillus subtilis glutamine phosphoribosylpyrophosphate amidotransferase by nucleotides. *Biochemistry 36*, 10718-10726.

(53) Isupov, M. N., Obmolova, G., Butterworth, S., Badet-Denisot, M. A., Badet, B., Polikarpov, I., Littlechild, J. A., and Teplyakov, A. (1996) Substrate binding is required for assembly of the active conformation of the catalytic site in Ntn amidotransferases: evidence from the 1.8 A crystal structure of the glutaminase domain of glucosamine 6-phosphate synthase. *Structure 4*, 801-810.

(54) Teplyakov, A., Obmolova, G., Badet, B., and Badet-Denisot, M. A. (2001) Channeling of ammonia in glucosamine-6-phosphate synthase. *J Mol Biol 313*, 1093-1102.

(55) Schmid, H. H., Schmid, P. C., and Natarajan, V. (1990) N-acylated glycerophospholipids and their derivatives. *Prog Lipid Res 29*, 1-43.

(56) Schmid, H. H., and Berdyshev, E. V. (2002) Cannabinoid receptor-inactive N-acylethanolamines and other fatty acid amides: metabolism and function. *Prostaglandins Leukot Essent Fatty Acids 66*, 363-376.

(57) Sugiura, T., Kobayashi, Y., Oka, S., and Waku, K. (2002) Biosynthesis and degradation of anandamide and 2-arachidonoylglycerol and their possible physiological significance. *Prostaglandins Leukot Essent Fatty Acids 66*, 173-192.

(58) Epps, D. E., Schmid, P. C., Natarajan, V., and Schmid, H. H. (1979) N-Acylethanolamine accumulation in infarcted myocardium. *Biochem Biophys Res Commun 90*, 628-633.

(59) Schmid, P. C., Krebsbach, R. J., Perry, S. R., Dettmer, T. M., Maasson, J. L., and Schmid, H. H. (1995) Occurrence and postmortem generation of anandamide and other long-chain N-acylethanolamines in mammalian brain. *FEBS Lett 375*, 117-120.

(60) Kondo, S., Sugiura, T., Kodaka, T., Kudo, N., Waku, K., and Tokumura, A. (1998) Accumulation of various N-acylethanolamines including N-arachidonoylethanolamine (anandamide) in cadmium chloride-administered rat testis. *Arch Biochem Biophys 354*, 303-310.

(61) Okamoto, Y., Morishita, J., Tsuboi, K., Tonai, T., and Ueda, N. (2004) Molecular characterization of a phospholipase D generating anandamide and its congeners. *J Biol Chem 279*, 5298-5305.

(62) Facci, L., Dal Toso, R., Romanello, S., Buriani, A., Skaper, S. D., and Leon, A. (1995) Mast cells express a peripheral cannabinoid receptor with differential sensitivity to anandamide and palmitoylethanolamide. *Proc Natl Acad Sci U S A 92*, 3376-3380.

(63) Mazzari, S., Canella, R., Petrelli, L., Marcolongo, G., and Leon, A. (1996) N-(2-hydroxyethyl)hexadecanamide is orally active in reducing edema formation and inflammatory hyperalgesia by down-modulating mast cell activation. *Eur J Pharmacol 300*, 227-236.

(64) Berdyshev, E., Boichot, E., Corbel, M., Germain, N., and Lagente, V. (1998) Effects of cannabinoid receptor ligands on LPS-induced pulmonary inflammation in mice. *Life Sci 63*, PL125-129.

(65) Skaper, S. D., Facci, L., Romanello, S., and Leon, A. (1996) Mast cell activation causes delayed neurodegeneration in mixed hippocampal cultures via the nitric oxide pathway. *J Neurochem 66*, 1157-1166.

(66) Berdyshev, E. V., Boichot, E., Germain, N., Allain, N., Anger, J. P., and Lagente, V. (1997) Influence of fatty acid ethanolamides and delta9-tetrahydrocannabinol on cytokine and arachidonate release by mononuclear cells. *Eur J Pharmacol 330*, 231-240.

(67) Schmid, P. C., Zuzarte-Augustin, M. L., and Schmid, H. H. (1985) Properties of rat liver N-acylethanolamine amidohydrolase. *J Biol Chem 260*, 14145-14149.

(68) Ueda, N., Puffenbarger, R. A., Yamamoto, S., and Deutsch, D. G. (2000) The fatty acid amide hydrolase (FAAH). *Chem Phys Lipids 108*, 107-121.

(69) Tsuboi, K., Sun, Y. X., Okamoto, Y., Araki, N., Tonai, T., and Ueda, N. (2005) Molecular characterization of N-acylethanolamine-hydrolyzing acid amidase, a novel member of the choloylglycine hydrolase family with structural and functional similarity to acid ceramidase. *J Biol Chem 280*, 11082-11092.

(70) Tsuboi, K., Takezaki, N., and Ueda, N. (2007) The N-acylethanolamine-hydrolyzing acid amidase (NAAA). *Chem Biodivers 4*, 1914-1925.

(71) Ikonen, E., Baumann, M., Gron, K., Syvanen, A. C., Enomaa, N., Halila, R., Aula, P., and Peltonen, L. (1991) Aspartylglucosaminuria: cDNA encoding human aspartylglucosaminidase and the missense mutation causing the disease. *Embo J 10*, 51-58.

(72) Sugita, M., Dulaney, J. T., and Moser, H. W. (1972) Ceramidase deficiency in Farber's disease (lipogranulomatosis). *Science 178*, 1100-1102.

(73) Bachur, N. R., and Udenfriend, S. (1966) Microsomal synthesis of fatty acid amides. *J Biol Chem 241*, 1308-1313.

(74) Bracey, M. H., Hanson, M. A., Masuda, K. R., Stevens, R. C., and Cravatt, B. F. (2002) Structural adaptations in a membrane enzyme that terminates endocannabinoid signaling. *Science 298*, 1793-1796.

(75) Ueda, N., and Yamamoto, S. (2000) Anandamide amidohydrolase (fatty acid amide hydrolase). *Prostaglandins Other Lipid Mediat 61*, 19-28.

(76) Ueda, N. (2002) Endocannabinoid hydrolases. *Prostaglandins Other Lipid Mediat 68-69*, 521-534.

(77) Bisogno, T., De Petrocellis, L., and Di Marzo, V. (2002) Fatty acid amide hydrolase, an enzyme with many bioactive substrates. Possible therapeutic implications. *Curr Pharm Des 8*, 533-547.

(78) Kim, J. K., Yang, I. S., Shin, H. J., Cho, K. J., Ryu, E. K., Kim, S. H., Park, S. S., and Kim, K. H. (2006) Insight into autoproteolytic activation from the structure of cephalosporin acylase: a protein with two proteolytic chemistries. *Proc Natl Acad Sci U S A 103*, 1732-1737.

(79) Li, Y., Chen, J., Jiang, W., Mao, X., Zhao, G., and Wang, E. (1999) In vivo post-translational processing and subunit reconstitution of cephalosporin acylase from Pseudomonas sp. 130. *Eur J Biochem 262*, 713-719.

(80) Perler, F. B., Olsen, G. J., and Adam, E. (1997) Compilation and analysis of intein sequences. *Nucleic Acids Res 25*, 1087-1093.

(81) Xu, Q., Buckley, D., Guan, C., and Guo, H. C. (1999) Structural insights into the mechanism of intramolecular proteolysis. *Cell 98*, 651-661.

(82) Michalska, K., Bujacz, G., and Jaskolski, M. (2006) Crystal structure of plant asparaginase. *J Mol Biol 360*, 105-116.

(83) Kim, Y., Kim, S., Earnest, T. N., and Hol, W. G. (2001) Precursor structure of cephalosporin acylase. Insights into autoproteolytic activation in a new N-terminal hydrolase family. *J Biol Chem 277*, 2823-2829.

(84) **Kim, S., and Kim, Y. (2001) Active site residues of cephalosporin acylase are critical not only for enzymatic catalysis but also for post-translational modification. *J Biol Chem 276*, 48376-48381.**

(85) **Sykes, R. B., Cimarusti, C. M., Bonner, D. P., Bush, K., Floyd, D. M., Georgopapadakou, N. H., Koster, W. M., Liu, W. C., Parker, W. L., Principe, P. A., Rathnum, M. L., Slusarchyk, W. A., Trejo, W. H., and Wells, J. S. (1981) Monocyclic beta-lactam antibiotics produced by bacteria. *Nature 291*, 489-491.**

(86) **Ishii, Y., Saito, Y., Fujimura, T., Isogai, T., Kojo, H., Yamashita, M., Niwa, M., Kohsaka, M. (1994) A novel 7-β-(4-carboxybutanamido)-cephalosporanic acid acylase isolated from Pseudomonas strain C427 and its high-level production in Escherichia coli *Journal of Fermentation and Bioengineering 77*, 591-597.**

(87) **Kim, Y., Yoon, K., Khang, Y., Turley, S., and Hol, W. G. (2000) The 2.0 A crystal structure of cephalosporin acylase. *Structure 8*, 1059-1068.**

FIGURES

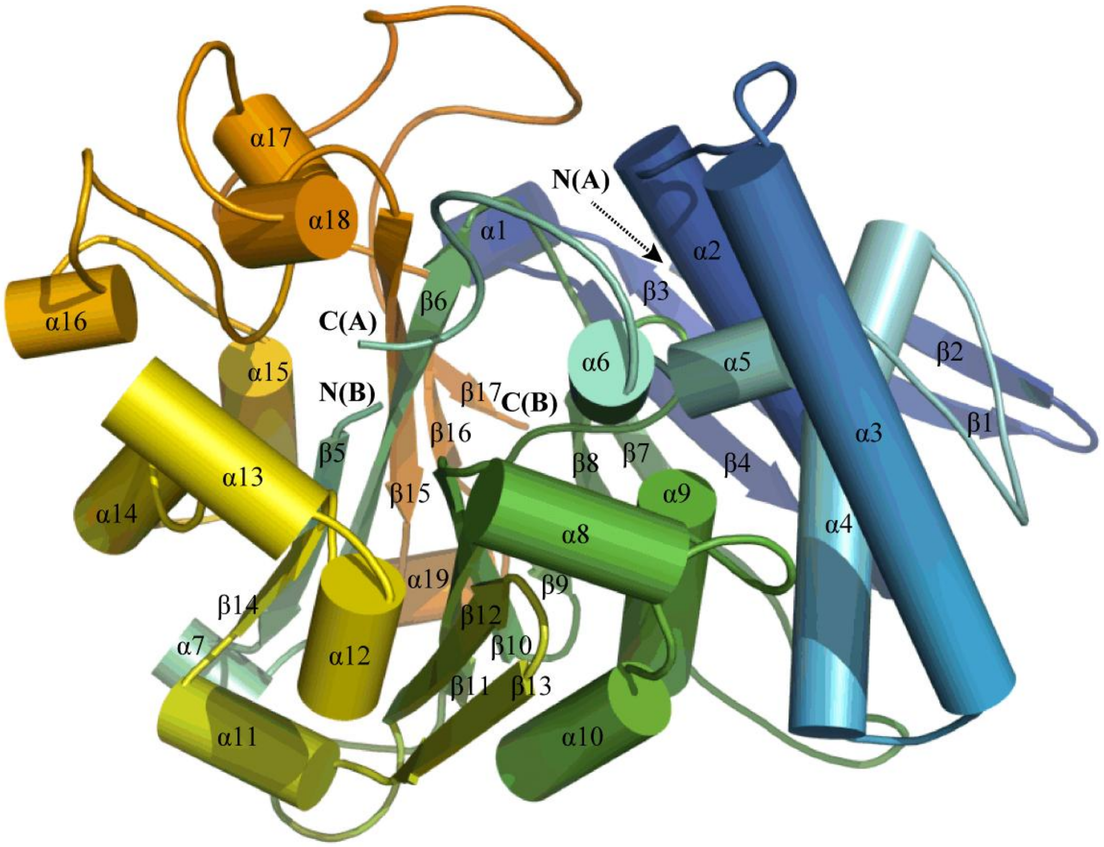


Fig. 1a.

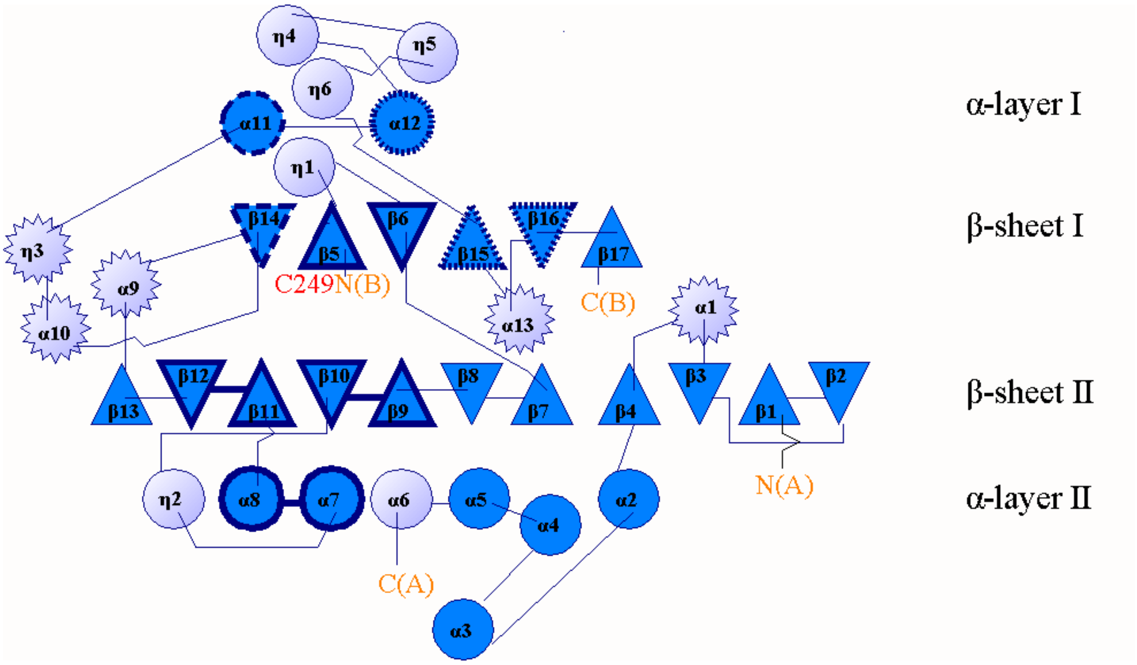


Fig. 1b.

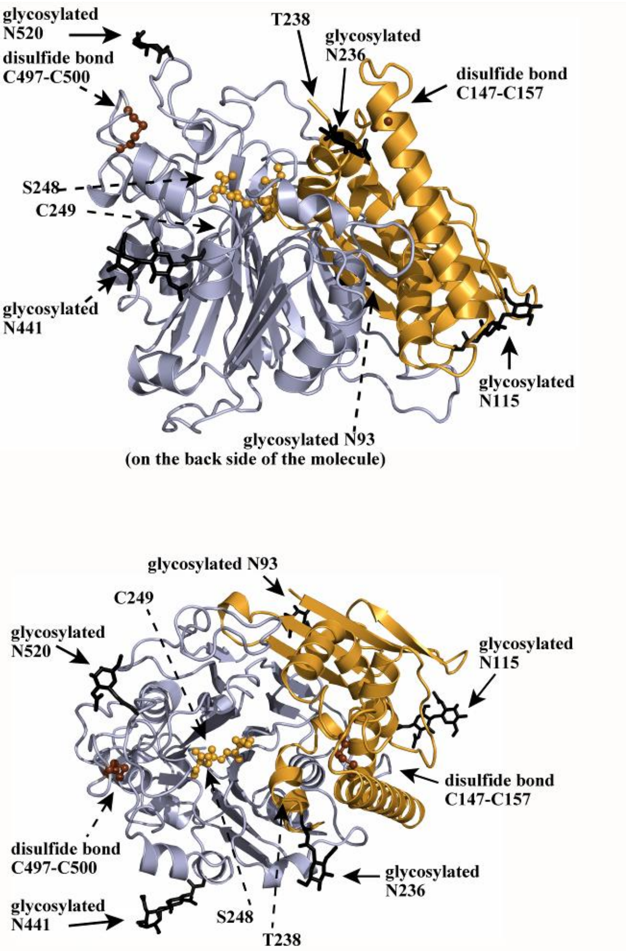

glycosylated N520
disulfide bond C497-C500
T238
glycosylated N236
disulfide bond C147-C157
S248
C249
glycosylated N441
glycosylated N115
glycosylated N93 (on the back side of the molecule)
glycosylated N93
C249
glycosylated N520
glycosylated N115
disulfide bond C147-C157
disulfide bond C497-C500
glycosylated N236
glycosylated N441
S248
T238

Fig. 2.

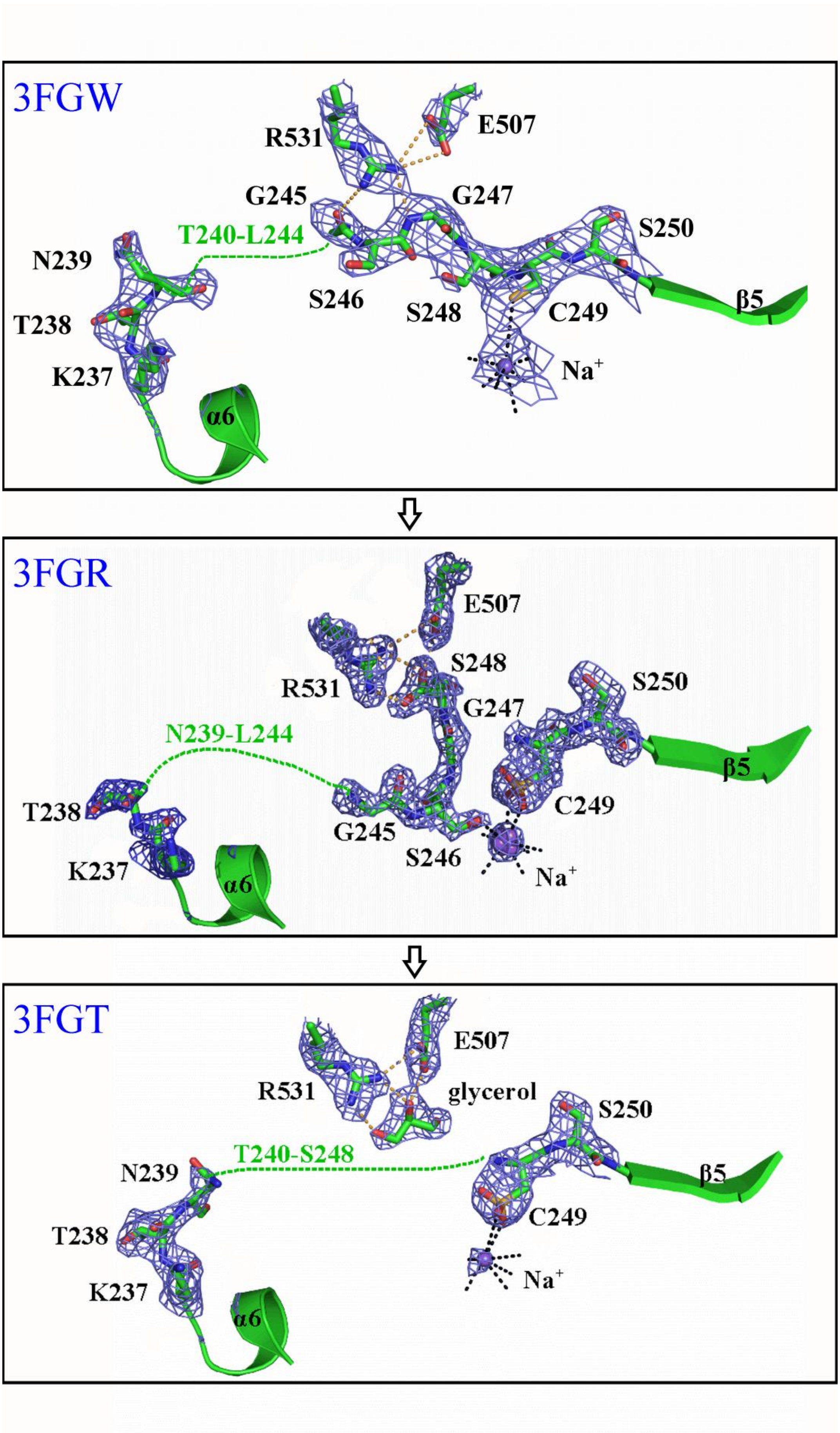

3FGW
R531
E507
G245
G247
T240-L244
S250
N239
S246
S248
C249
β5
T238
Na⁺
K237
α6
3FGR
E507
S248
R531
G247
N239-L244
S250
β5
C249
T238
G245
S246
K237
α6
Na⁺
3FGT
E507
R531
glycerol
S250
T240-S248
N239
β5
C249
T238
Na⁺
K237
α6

Fig. 3.

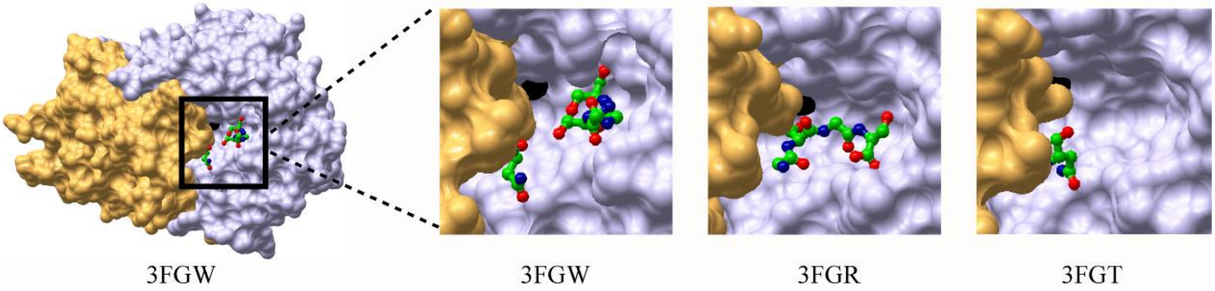


Fig. 4.

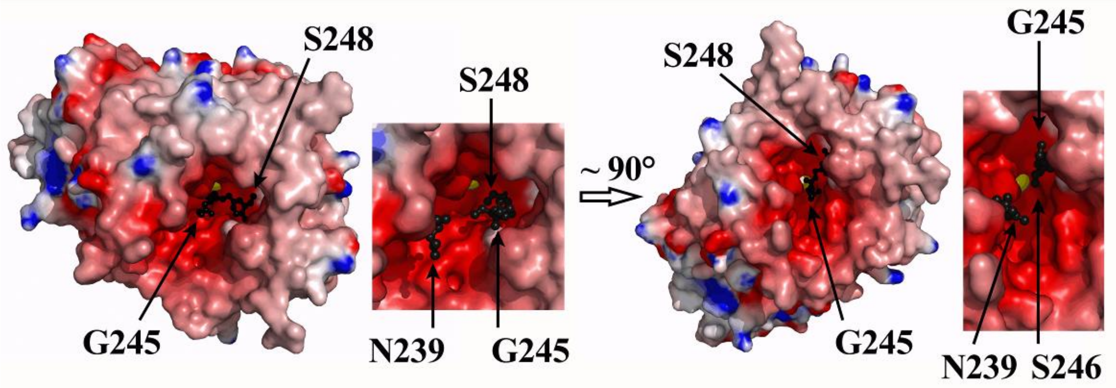


Fig. 5.

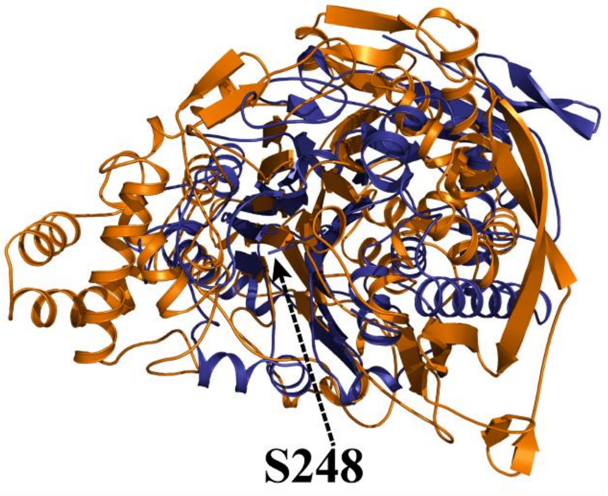


Fig. 6.

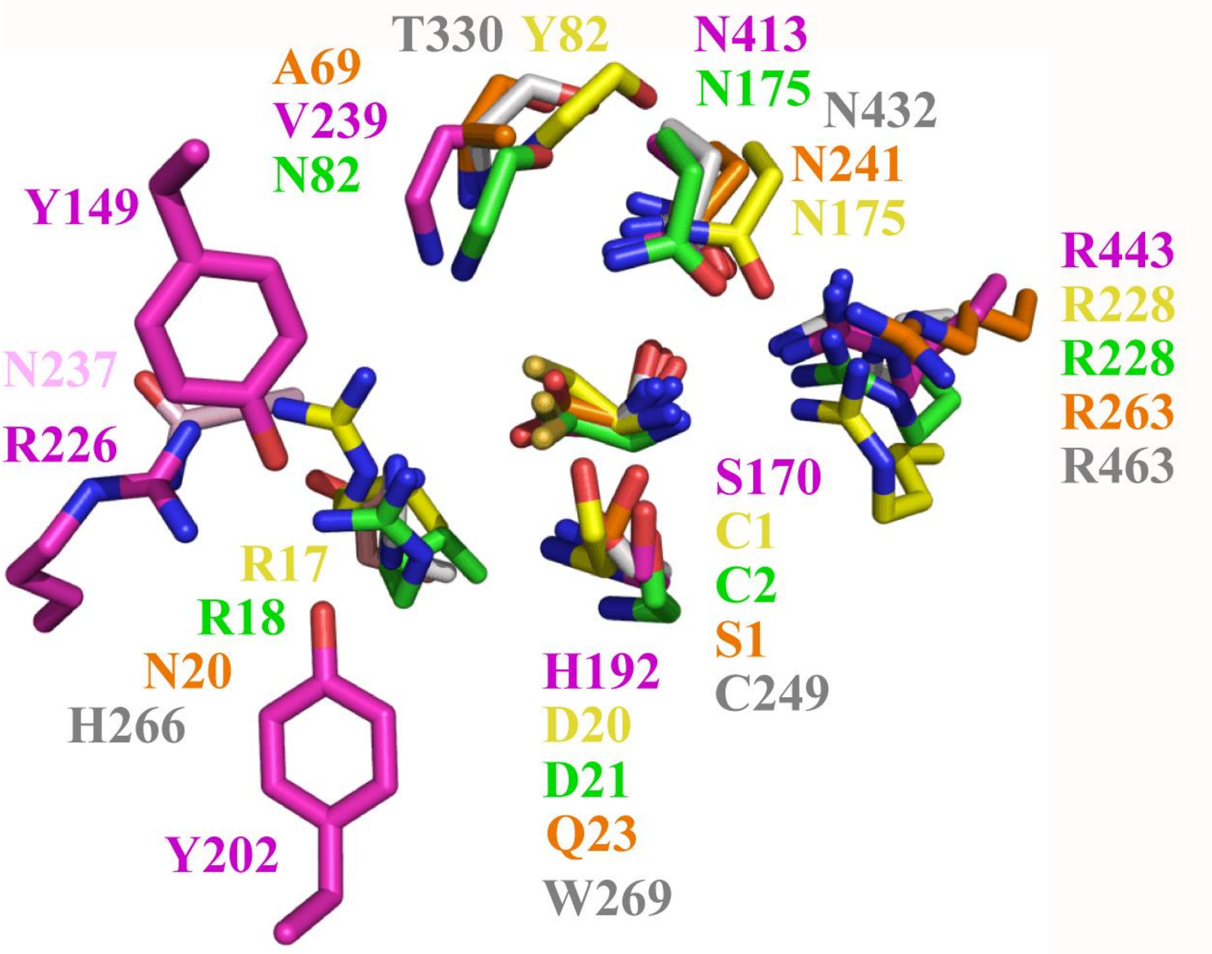

T330
Y82
N413
N175
A69
V239
N82
N432
N241
N175
Y149
R443
R228
R228
R263
R463
N237
R226
S170
C1
C2
S1
C249
R17
R18
N20
H266
H192
D20
D21
Q23
W269
Y202

Fig. 7.

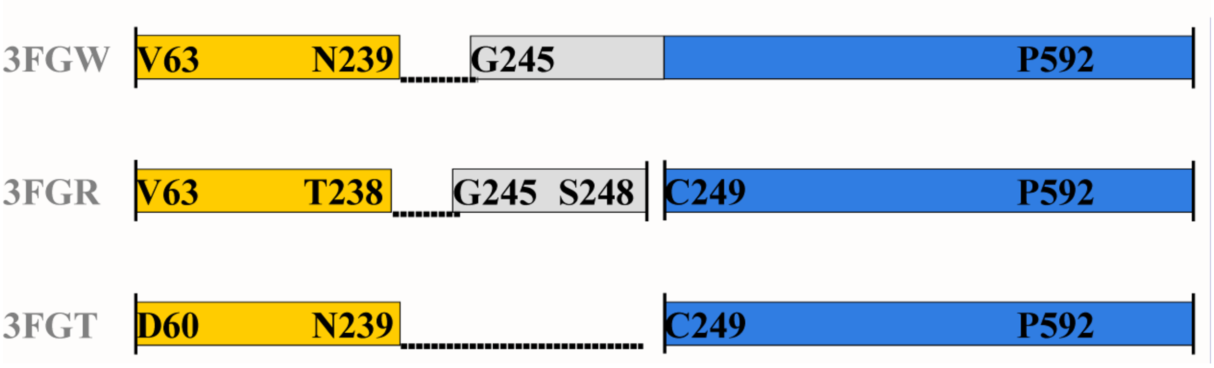

3FGW
V63
N239
G245
P592
3FGR
V63
T238
G245
S248
C249
P592
3FGT
D60
N239
C249
P592

Fig. S1.

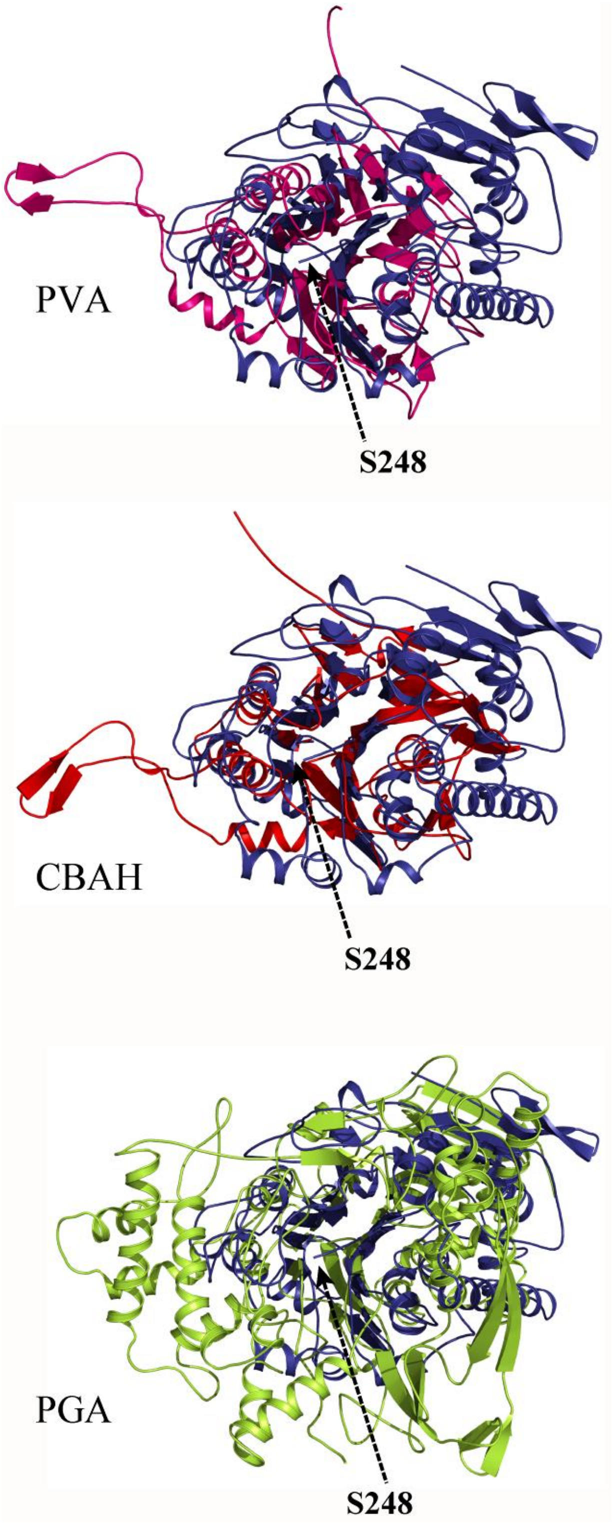


Fig. S2.

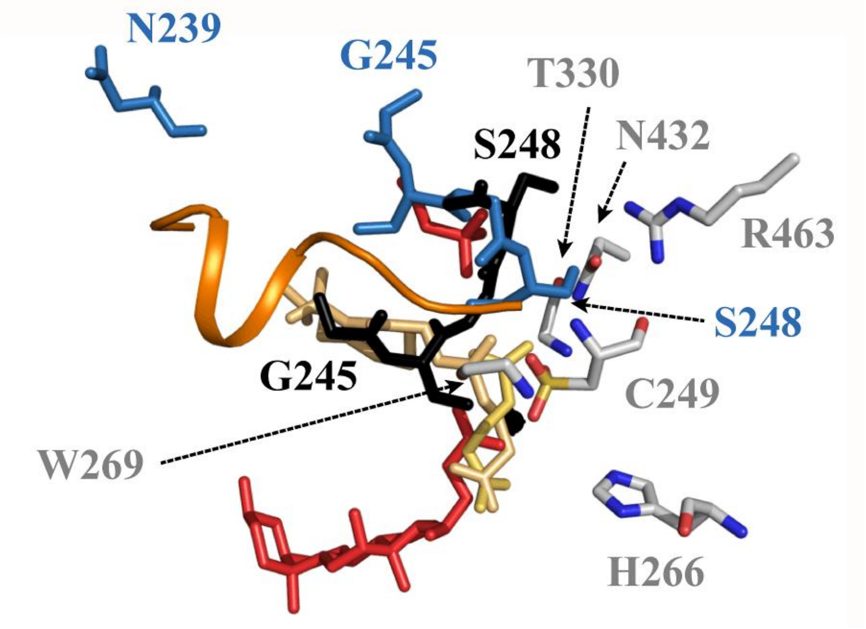


Fig. S3.

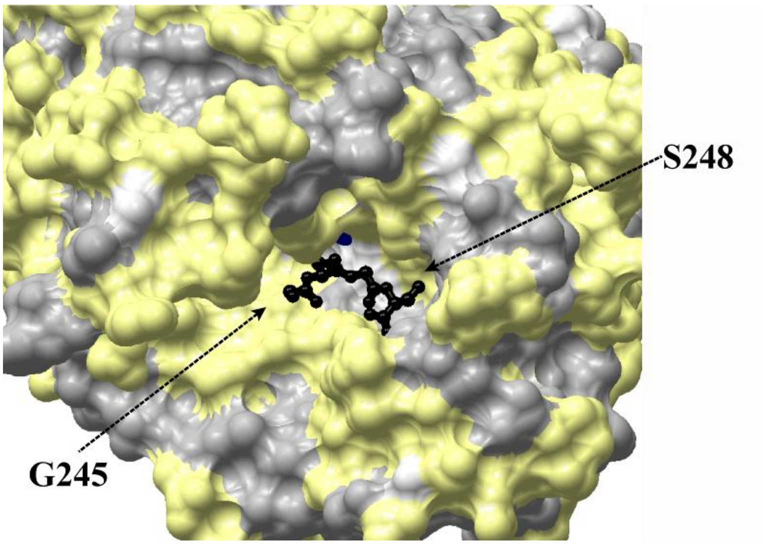


Fig. S4.

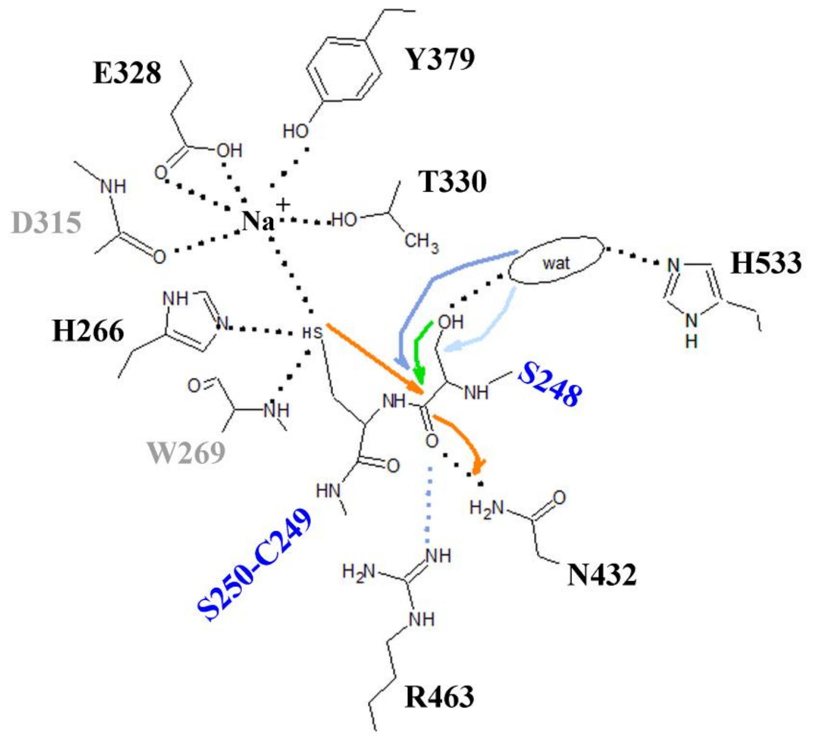

E328
Y379
HO
OH
O
NH
D315
Na+
T330
HO
CH3
O
wat
H533
N
N
H
H266
NH
N
HS
OH
S248
NH
O=
NH
W269
NH
O
O
HN
H2N
O
S250-C249
H2N
NH
NH
N432
R463

Fig. S5.

5. DISCUSSION

The main issues of the structural analysis have been discussed in the submitted manuscript. Some complementary details of already described procedures are outlined and discussed in the following. Most of the web servers for protein structure analysis mentioned in the general introduction (I.1.4.) have been used in the process of characterization of the 66.3 kDa protein structure. However, only a few of them provided more insight than revealed by DALI (Holm & Sander, 1996) and manual database and literature search. Some information supporting obtained data as well as additional results are discussed below.

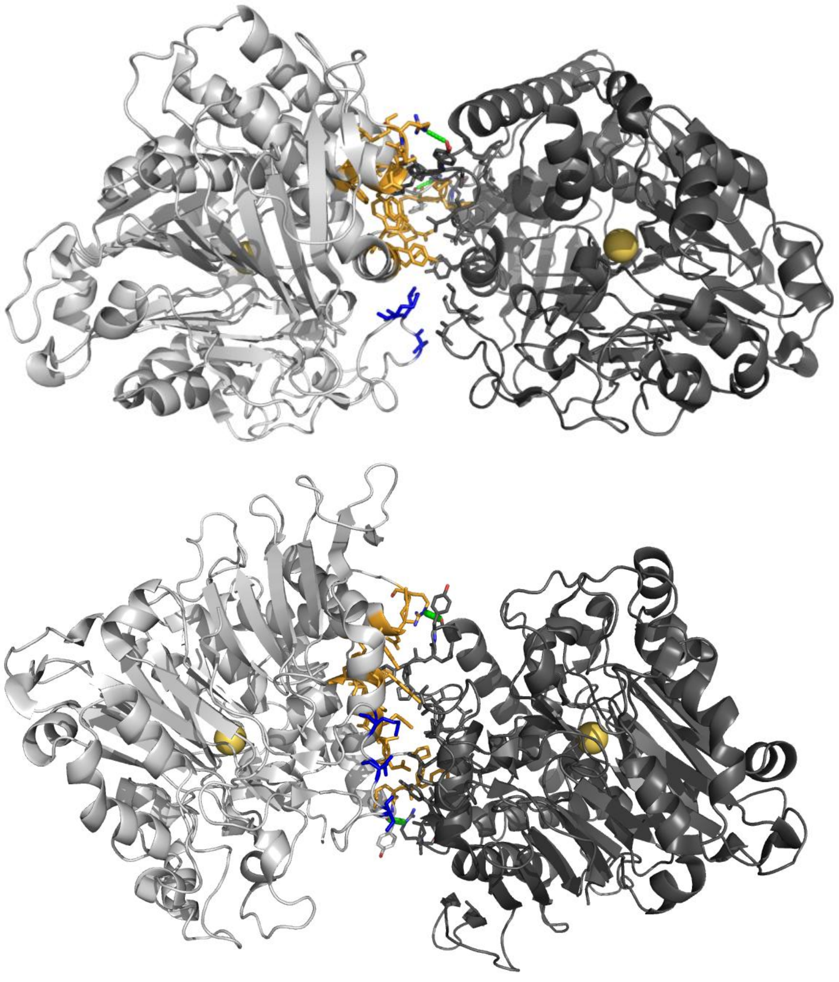


Fig. I-11. Interface between the two molecules of the 66.3 kDa assumed to form a stable dimer according to results from the PISA server. The molecules are coloured in dark and light grey, respectively, and shown in cartoon mode except for the interfacing residues which are represented as sticks. In the molecule on the left side, amino acid residues involved in the formation of the interface are highlighted in orange and blue according to their location within the 28 kDa and 40 kDa fragment, respectively. Hydrogen bonds are indicated by green lines. For orientation, the metal ion bound at the active site is shown as a yellow sphere.

The program PISA (Krissinel et al., 2007) mentioned in the submitted manuscript assigns a complexation significance score (CSS) to each predicted oligomer ranging from 0 to 1 as interface relevance to complexation increases. For the 66.3 kDa protein, the CSS was estimated to lie in the range between 0.7 (3FGT) and 0.8 (3FGW) suggesting the existence of a stable dimer. The dimer interface amounts to about 800-900 Å^2 which is equivalent to about 4 % of the solvent-accessible surface (Fig. I-11).

The interface involves the same residues of both structures with the following exceptions. Hydrogen bonds are formed between the backbone carbonyl oxygen of Tyr152 and Arg204 in 3FGW, while Asn145 and Tyr136 as well as Lys215 and the backbone carbonyl oxygen of Pro149 participate in intermolecular hydrogen bonds in 3FGT. Except for Ser574, Leu576 and Met578 all 22 residues of the interface are located in the N-terminal region (mainly in the residue ranges Tyr136-Val154, Glu198-Arg204 and Lys215-Phe219).

Searching the PDB with PISA for similar interfaces revealed the Ntn hydrolases which had also been found with DALI (cf. Tab. 2 of chapter 4). Most of the similar interfaces comprise a more extended area of up to 3400 Å^2, whereas the dimer interface of cephalosporin acylase (CA) is comparable to that of the 66.3 kDa protein. Although not finally shown we suggest the 66.3 kDa protein to form a stable dimer as indicated by two independent methods, namely by means of size exclusion chromatography (unpublished data, Prof. Lübke, personal communication) and analysis of crystal contacts.

Attention to several active site residues was drawn by the results from searching the catalytic site atlas which is available at the European Bioinformatics Institute (EMBL-EBI) (Porter et al., 2004) and confirmed expectations on the basis of the data from DALI.

In order to detect putative ways by which substrate molecules could enter the active site, the program MOLE was used (Petrek et al., 2006; Petrek et al., 2007). According to the obtained results there are three different sets of amino acids that form a tunnel leading to the putative catalytic key residue Cys249. Six amino acids are comprised in the border of all predicted tunnels, namely Glu229 of the 28 kDa fragment and Cys249, Thr330, Asn432, Arg531 and His533 of the C-terminal fragment. They have already been discussed in the course of the comparison with other Ntn hydrolases. Half of the residues are conserved among the LamL2 homologues (laminin-like proteins) which the 66.3 kDa protein is assigned to (Cys249, Thr330, Asn432). While two tunnels include mainly residues from the 40 kDa fragment, the third tunnel comprises residues of both fragments in an almost equal ratio. The conserved Asp

230 is located at the entrance of the tunnels and forms a tight hydrogen bond with a glycerol molecule, thus most likely playing an important role in preliminary substrate recognition.

These tunnels become only available during the maturation process which involves limited proteolysis. The cleavage between Ser248 and Cys249 most likely occurs in an autocatalytic process resulting in a flexible linker peptide as discussed in detail in chapter 3. In contrast, a putative second cleavage might be required at the N-terminus of this linker region in order to achieve complete access to the substrate binding pocket. However, so far neither it is known whether such an incision is necessary at all nor the exact cleavage site or the underlying mechanism have been determined.

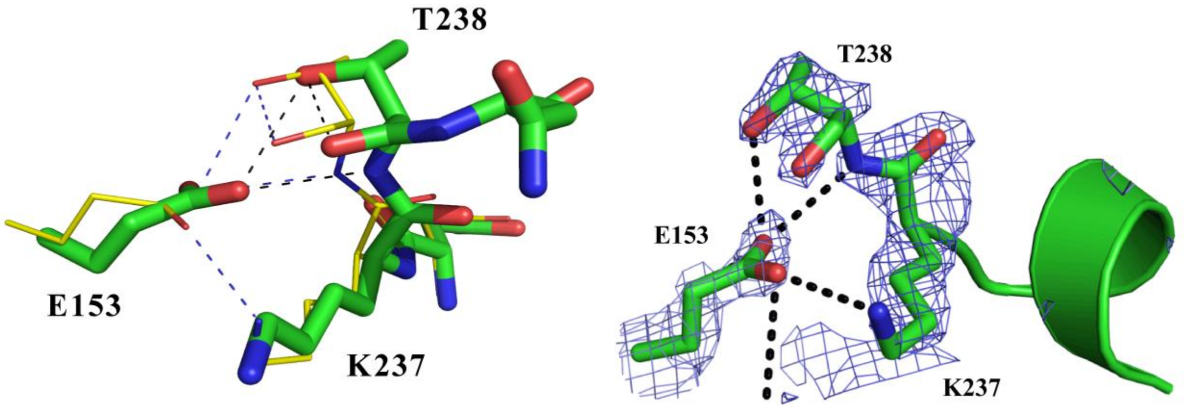


Fig. I-12. Arrangement of the amino acid residues around the N-terminus of the presumable linker peptide N239-S248. The residues N236-N239 as well as the side chain of E153, which might be involved in a second autocatalytic cleavage step between N236 and K237, are shown in stick mode coloured by atom (oxygen in red, nitrogen in blue, carbon in green and yellow, respectively). (a) Comparison of the region between the structures 3FGR (carbon atoms in yellow) and 3FGT (carbon atoms in green). Hydrogen bonds are indicated by blue (3FGR) and black (3FGT) dashed lines, respectively. (b) Surrounding electron density for the structure 3FGR. The $2F_oF_c$ map is contoured at the 1.0 σ level.

A second autocatalytic cleavage releasing a spacer peptide has been reported for CA (Kim et al., 2001; Kim et al., 2006). The glutamic acid residue 159 is required for this cleavage at the N-terminus of the spacer peptide between Gly160 and Asp161, which is abolished in the mutant proteins E159Q and E159M. The superposition of CA and the 66.3 kDa protein shows the side chain carboxyl groups of Glu159 and Glu153, respectively, to be located similarly, even though they belong to non-equivalent β-strands (Fig. I-12). Glu159 forms one part of the structural requirement for the autocatalytic cleavage mechanism of CA. However, a residue feasible to form the oxyanion hole in the 66.3 kDa protein could not be identified. Thus, we suggest the C-terminus of the 28 kDa fragment to be trimmed from residue Ser248 by lysosomal hydrolases rather than by a second autocatalytic step.

Upon cleavage between Ser248 and Cys249, the C-terminal residues of the 28 kDa fragment could protrude from the protein and would be accessible for proteases which are quite abundant in the lysosomal compartment. The N-glycan attached to Asn236 of the 66.3 kDa protein which was shown to be included in the mature 28 kDa fragment (Deuschl et al., 2006), should protect it against further C-terminal degradation. The enzymatic maturation mechanism would also explain the presence of at least two residues C-terminal of the glycosylated Asn236, in all structures (Fig. 3 in chapter 4), since the crystallized protein had not reached the lysosomal compartment, but was secreted by exocytosis. Furthermore, such a mechanism would be in accordance with the lysosomal Ntn hydrolase aspartylglucosaminidase (AGA) (Ikonen et al., 1991; Saarela et al., 2004; Oinonen et al., 1995). As mentioned above, the C-terminal residue of the 28 kDa fragment has not been defined yet. However, the exact length of the linker might not have any effect on the acylase activity – as suggested by studies on CA from different *Pseudomonas* species, for which natively occurring variations from 8 to 11 amino acids have been investigated (Sykes et al., 1981; Ishii et al., 1994; Kim et al., 2000; Kim et al., 2001). Alternatively, the C-terminus of the 28 kDa might not even be trimmed at all but move to the protein surface due to a high degree of flexibility. This conformational change already could give substrates sufficient access to the active site for turnover as demonstrated for CA (Kim et al., 2001).

In murine tissues, the 40 kDa fragment of the 66.3 kDa protein is further processed to a 25 kDa and a 15 kDa fragment with the peptide bond between Arg513 and Ser514 as the cleavage site (Deuschl et al., 2006). However, all structures of the 66.3 kDa protein contain the uncleaved 40 kDa fragment. Thus, this cleavage site has so far not been confirmed by structural analysis limiting the certainty of the following hypothesis about a putative mechanism of the cleavage between Arg513 and Ser514.

Based on the absence of a protease cleavage site according to the analysis of the amino acid sequence using the ExPASy PeptideCutter tool (Gasteiger et al., 2005) in concert with the challenging access of the scissile bond for such an enzyme we suggest this incision to be of an autocatalytic manner. The arrangement of the amino acids around Ser514 is shown in Fig. I-13. In analogy to the cleavage between Cys249 and Ser248, such an autoproteolytic cleavage most likely is catalyzed by the Oγ atom of Ser514, the nucleophilicity of which might be enhanced by the side chain of the conserved histidine His533. In the model 3FGT, the imidazole ring of this histidine is involved in a hydrogen bonding network with a water and a glycerol molecule in the substrate binding pocket. The first hydrogen bond is formed directly, whereas the latter is mediated by a water molecule and the side chain of Arg531. Due to these interactions and since the cleavage between Ser514 and Arg513 has not occured in any of the three crystal structures the incision might be substrate-assisted. Accordingly, it took place only upon conformational rearrangement caused by the binding of a substrate at the active site around Cys249. The oxyanion generated by the attack of Ser514 on the carbonyl carbon of Arg513 could be stabilized by the backbone amide groups of Asp515 and Leu516.

This proteolytic cleavage might generate a second N-terminal nucleophilic residue, but whether Ser514 has a catalytic function is not known so far. A pocket is found close to the putative second N-terminal nucleophilic amino acid, which is probably blocked by a second linker peptide located N-terminal of Arg513. In order to get more insight into a putative functional role specifically associated with each of these additionally arising fragments, the PDB was separately searched with DALI for related structures using either the 25 kDa or the 15 kDa fragment. While the 25 kDa fragment contains significant similarity to the same Ntn hydrolases revealed previously (Tab. 2 in chapter 4), no structure with a similar fold was found by a DALI search with the 15 kDa fragment. Due to the absence of any structural similarity, the 15 kDa fragment might have a unique role in the regulation or substrate specificity of the activity of the 66.3 kDa protein. All three fragments of the mature 66.3 kDa protein seem to be involved in formation of the substrate binding site. Alternatively, the 15 kDa fragment might be degraded upon binding of regulatory compounds to the active site initiating the turnover of the 66.3 kDa protein by lysosomal proteases. Further molecular and biochemical analysis is required to shed light onto additional maturation procedures of the 66.3 kDa protein.

According to the results obtained so far, N-acylethanolamines of specific chain lengths seem to be the most plausible candidates serving as substrates of the 66.3 kDa protein. In particular, they fit well in the recently developing model of a significant involvement of the lysosomal compartment into secretory pathways in addition to degradation processes (Fig. I-14).

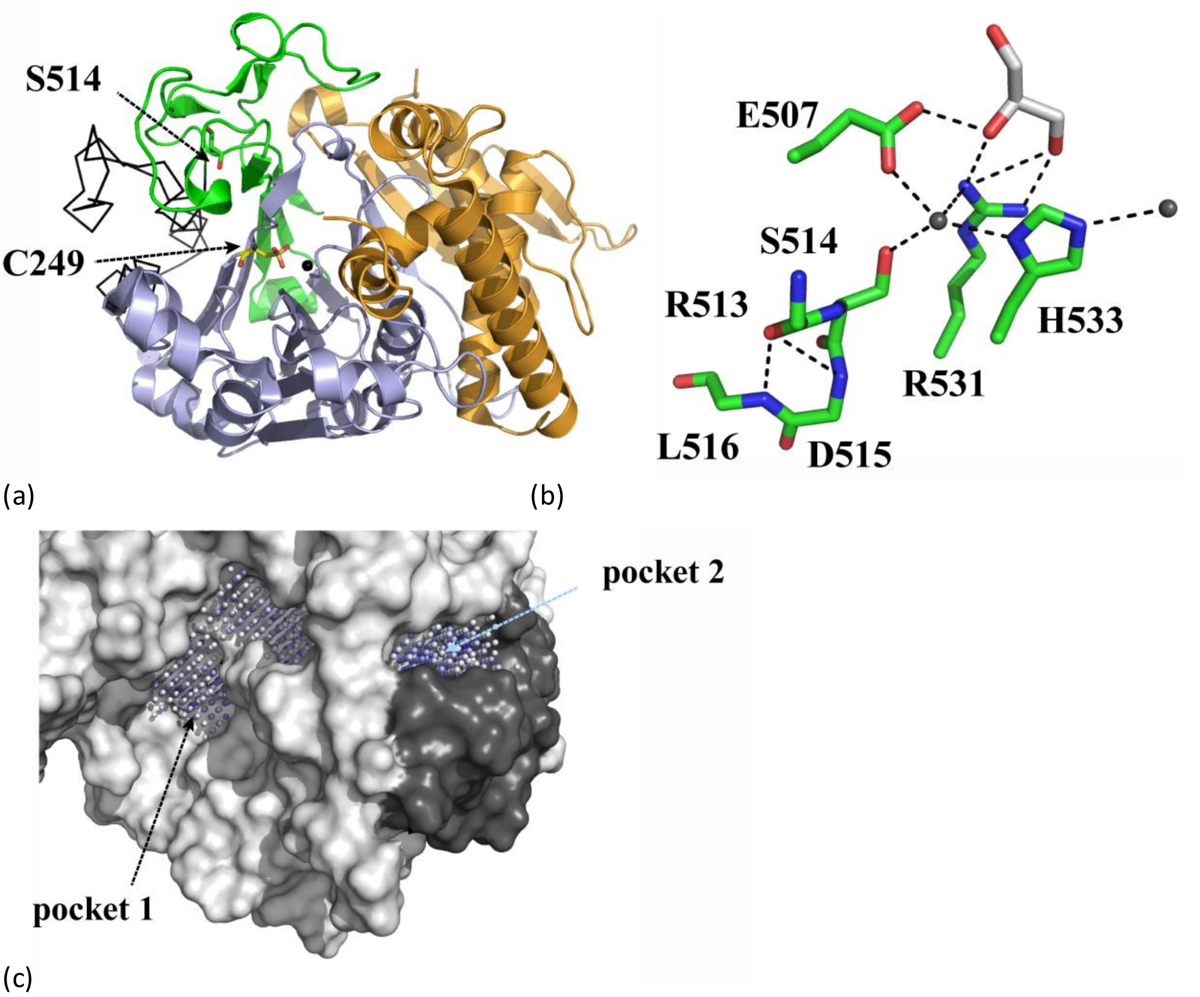


Fig. I-13. The 66.3 kDa protein after putative modelled further maturation procedures involving a proteolytic cleavage between Arg513 and Ser514. (a) Cartoon model of the 66.3 kDa protein consisting of three fragments of about 28 kDa, 25 kDa and 15 kDa in apparent molecular weight. The 28 kDa fragment is coloured in orange according to Fig. 2 of chapter 3. The colour code of the 40 kDa fragment is split. While the putative 25 kDa fragment is shown in light blue, the presumable 15 kDa fragment is shown in green with the N-terminal Ser514 in stick mode. The polypeptide chain which might be removed after the cleavage between Arg513 and Ser514 is shown as a black ribbon. For orientation, the nucleophilic Cys249 and a Na^+ ion bound at the putative active site are shown as sticks and a black sphere, respectively. (b) Enlarged view of the region around the scissile peptide bond between Arg513 and Ser514 including residues with a putative functional role. (c) Surface representation of the 66.3 kDa protein consisting of the 28 kDa and the 40 kDa protein (3FGT). The residue range, which might be removed in the very late steps of the maturation process due to the proteolysis within the 40 kDa fragment is indicated in dark grey. The region (Asn461-Arg513) has been chosen based on calculations of the theoretical molecular weight in concert with structural analysis. The substrate binding pocket arising around Cys249 (pocket 1) as well as a solvent-accessible pocket close to the residues which might be removed after the cleavage between Arg513 and Ser514 are indicated by small blue spheres according to the program POCKETPICKER (Weisel et al., 2007).

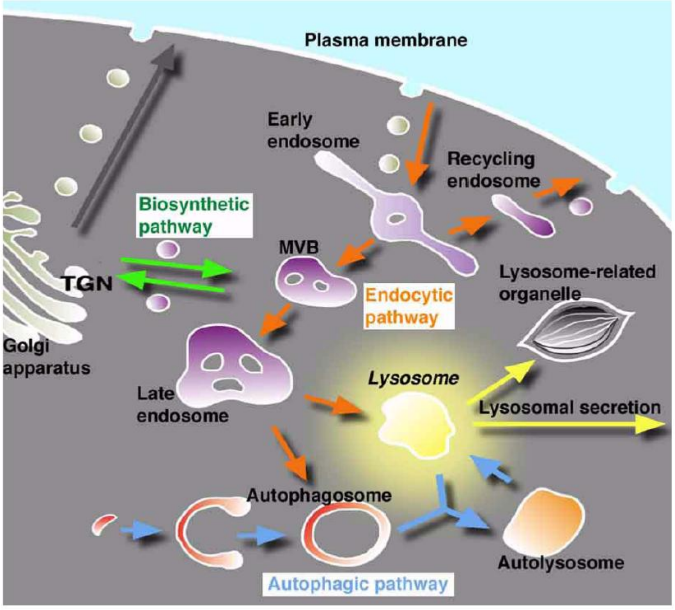


Fig. I-14. Pathways involved in the lysosomal network (Yoshimori, 2002).

Alternatively, other non-peptidic bonds seem to be suitable substrates of the 66.3 kDa protein. They occur only in few natural compounds such as lipid-modified proteins, sphingosines and acetylated lysine residues, and the 66.3 kDa protein might be involved in their degradation. While enzymes responsible for the degradation of farnesylated and geranylated proteins or peptides arising from lipid-modified proteins have been identified, an activity for the demyristoylation of proteins within lysosomes is only speculative at present (reviewed in Lu et al., 2006). If the 66.3 kDa protein had a function in this degradation pathway, its dysfunction would be assumed to cause a severe phenotype typical of lysosomal storage diseases due to an accumulation of the non-degraded material. However, in recent studies on 66.3 kDa protein knockout mice such a phenotype with swollen lysosomes has not been observed (unpublished data, laboratory of Prof. Dr. Lübke, Diploma thesis of Ina Hohensee, 2008).

Acetylated lysine residues are beyond others found in the basic charged N-terminal region of histones (Strahl & Allis, 2000; Zhang & Reinberg, 2001; Berger, 2002). These proteins have important roles in the organization of the DNA structure in eukaryotic cells. They are extensively post-translationally modified in an elaborated system, which is crucial for the regulation of gene expression (Jenuwein & Allis, 2001). Commonly, the larger the number of acetylated residues in histones is, the higher is the transcriptional activity, and decreased levels of acetylation are associated with repression of gene expression (Grunstein, 1997; Wade et al., 1997; Peterson, 2002). Dysfunction of the deacylating enzymes has been shown to be involved in cancer development (Grignani et al. 1998, Lin et al. 1998, Minucci et al. 2001). In contrast to these enzymes, the 66.3 kDa protein might remove the acetyl moiety from the proteins irreversibly in the course of degradation.

6. CONCLUSION AND FUTURE PERSPECTIVES

In this work, the lysosomal 66.3 kDa protein from mouse was crystallized in the monoclinic space group C2 and its structure was determined by means of sulfur / Xe SAD phasing. However, the contribution of the xenon atom to the overall scattering and therefore to the phasing power is not required for successful phase determination. The obtained structure to our knowledge belongs to the largest structures that have been feasible for structure determination by sulfur SAD so far and to a small group of proteins crystallized in monoclinic space groups and successfully solved by this method.

Especially for the 66.3 kDa protein the use of the anomalous signal of intrinsic sulfur atoms was a valuable alternative to standard experimental phasing procedures for the following reasons. The expression system, a human fibrosarcoma cell line, did neither allow for a high yield, which is required for extensive screening of heavy atom derivatives, nor for an efficient incorporation of selenomethionine. The latter issue might also have been the reason, that we could only recently obtain a limited number of crystals in conditions containing the appropriately modified 66.3 kDa protein.

To the best of our knowledge, the only structure of similar molecular weight that has been solved by sulfur SAD phasing is that of the 69 kDa protein TT0570 from *Thermus thermophilus* (Watanabe et al., 2005). In contrast to the procedure described here, Watanabe and coworkers did not use synchroton radiation at a wavelength of 1.9000 Å and a standard loop and mounting system, but applied a longer wavelength, Cr Kα radiation, and in addition a recently developed mounting technique that reduces the absorption by the cryo buffer and cryo loop (Kitago et al., 2005). The most important difference concerns the respective space group: the 66.3 kDa protein formed monoclinic crystals, whereas TT0570 has been crystallized in $P2_12_12$, a space group of higher symmetry. The challenge in the usage of sulfur SAD phasing for monoclinic or even triclinic crystals has also become obvious in a broad study of 23 different crystal forms by Mueller-Dieckmann et al. (Mueller-Dieckmann et al., 2007). In this study, neither the structures of three monoclinic nor that of a triclinic crystal form could be solved automatically by sulfur SAD, while the submission to the AutoRickshaw (Panjikar et al., 2005) pipeline was successful for the majority of the higher-symmetry examples. The determination of the 66.3 kDa protein structure is an extraordinary example of successful sulfur SAD phasing in that the protein is not only larger than most structures solved by this method so far but also crystallized in a low-symmetry space group (C2). Hence, this work is encouraging to apply this experimental phasing procedure more widely, since it uses only the anomalous signal of intrinsic protein atoms obviating crystal derivation with heavy atoms.

Currently, the processed X-ray data are used at the EMBL, Hamburg, Germany in order to improve the program pipeline AUTORICKSHAW (Panjikar et al., 2005) which is aimed to solve crystal structures automatically at the beamline (Santosh Panjikar and Manfred Weiss, personal communication).

Based on the structure determination, the updated working model of the activity of the 66.3 kDa protein involves NAEs as a substrate. In order to test this hypothesis, currently activity assays are performed in the laboratory of Prof. Dr. Torben Lübke.

Docking studies might be an alternative strategy to test the protein structure *in silico* for putative ligands that bind in the large pocket and fit into the active site arrangement, since these calculations allow for a broader screening. However, first trials using the test version of the free virtual screening server DOCKBLASTER (v1.0β, http://blaster.docking.org) did not reveal convincing hits, although it has been recently updated and now also includes a natural compounds database. The negative outcome might be due to the fact that exclusively commercially available compounds are used for screening and thus NAEs for example are excluded. More sophisticated approaches with stand-alone docking programs as well as calculations which consider high energy states of potential ligands as they might exist in transition states could help to make even more use of the obtained structural information. Obtained substances of course would have to be verified in subsequent activity assays and finally might result in the determination of the physiological function of the 66.3 kDa protein.

When appropriate substrates have been found, the next step in the course of the 66.3 kDa protein characterization will include studies of the binding affinity to the enzyme by means of isothermal calorimetry (ITC) or fluorimetry depending among others on the specific substrate.

In addition, subsequently or even in parallel co-crystallization experiments using the identified substrate compound and the mature form of the 66.3 kDa protein, in which the linker peptide has been removed or is at least flexible due to an incision N-terminal of Cys249, will be performed. In this context, new initial screenings might be required.

Moreover, in order to clarify the question about the exact C-terminus of the 28 kDa fragment, mutational studies have to be carried out that could be supported by crystallographic studies as well.

7. REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. (2004). Molecular Biology of the Cell. 4. ed., New York: *Garland Science*.

Berger, S. L. (2002). *Curr. Opin. Genet. Dev.* 12, 142-148.

Brown, J., Esnouf, R. M., Jones, M. A., Linnell, J., Harlos, K., Hassan, A. B., Jones, E. Y. (2002). *EMBO J.* 21, 1054-1062.

Capasso, R., Izzo, A. A., Fezza, F., Pinto, A., Capasso, F., Mascolo, N., Di Marzo, V. (2001). *Br. J. Pharmacol. 134*, 945-50.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D. (2003). *Nucl. Acids Res.* 31 (13), 3497-3500.

Ciechanover A. (2005). *Nat.Rev. Mol. Cell Biol.* 6, 79–86.

Cravatt, B. F., Demarest, K., Patricelli, M. P., Bracey, M. H., Giang, D. K., Martin, B. R., Lichtman, A. H. (2001). *Proc. Natl. Acad. Sci. U S A* 98, 9371-6.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. ,Sheldrick, G. M. (1999). *J. Mol. Biol.* 289, 83-92.

Debreczeni, J. E., Bunkoczi, G., Girmann, B. ,Sheldrick, G. M. (2003). *Acta Cryst. D*59, 393-395.

Debreczeni, J. E., Bunkoczi, G., Ma, Q., Blaser, H., Sheldrick, G. M. (2003). *Acta Cryst. D*59, 688-696.

Debreczeni, J. E., Girmann, B., Zeeck, A., Kratzner, R., Sheldrick, G. M. (2003). *Acta Cryst. D*59, 2125-2132.

De Duve, C. (1969). The lysosome in retrospect. Lysosomes in Biology and Pathology. D. J. D. a. F. H. Amsterdam, North-Holland Publishing Company. 1, 3-40.

De Duve, C. (1983). *Eur. J. Biochem.* 137(3), 391-7.

De Duve, C., Wattiaux, R. (1966). *Annu. Rev. Physiol.* 28, 435-92.

Deuschl, F., Kollmann, K., von Figura, K., Lübke, T. (2006). *FEBS Lett.* 580, 5747-52.

Djinovic-Carugo, K., Helliwell, J. R., Stuhrmann, H., Weiss, M. S. (2005). *J. Synchrotron Radiat.* 12, 410-419.

Fehrenbacher, N., Jaattela, M. (2005). *Cancer Re.s* 65, 2993-5.

Feulner, J. A., Lu, M., Shelton, J. M., Zhang, M., Richardson, J. A., Munford, R. S. (2004). *Infect. Immun. 72*, 3171-8.

Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy Server. In: John, M., Walker. (Eds). The proteomics protocols handbook. *Humana Press*, p. 571-607.

Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A., Henrick, K. (2005). *Proteins* 58, 190-199.

Gordon, E. J., Leonard, G. A., McSweeney, S., Zagalsky, P. F. (2001). *Acta Cryst. D* 57, 1230-1237.

Gouet, P., Courcelle, E., Stuart, D.I., Metoz, F. (1999). *Bioinformatics* 15, 305-8.

Grignani, F., De Matteis, S., Nervi, C., Tomassoni, L., Gelmetti, V., Cioce, M., Fanelli, M., Ruthardt, M., Ferrara, F.F., Zamir, I., Seiser, C., Grignani, F., Lazar, M. A., Minucci, S., Pelicci, P. G. (1998). *Nature* 391, 815-8.

Grunstein, M. (1997). *Nature* 389, 349-52.

Haas, A. (2007). *Traffic* 8(4), 311-30.

Hansen, H. S., Moesgaard, B., Hansen, H. H., Petersen, G. (2000). *Chem. Phys. Lipids 108*, 135-50.

Hasilik, A. (1992). *Experientia* 48(2), 130-51.

Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., Raushel, F.M. (2007) *Nature* 448, 775-779.

Holm, L., Sander, C. (1996). *Methods Enzymol.* 266, 653-62.

Ikonen, E., Baumann, M., Gron, K., Syvanen, A. C., Enomaa, N., Halila, R., Aula, P., Peltonen, L. (1991). *Embo J.* 10, 51-8.

Ishii, Y., Saito, Y., Fujimura, T., Isogai, T., Kojo, H., Yamashita, M., Niwa, M., Kohsaka, M. (1994). J. Ferment. and Bioeng. 77, 591-597.

Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A., Kolchanov, N.A. (2005). *Nucl. Acids Res.* 33, D183-D187.

Izzo, A. A., Fezza, F., Capasso, R., Bisogno, T., Pinto, L., Iuvone, T., Esposito, G., Mascolo, N., Di Marzo, V., Capasso, F. (2001). *Br. J. Pharmacol. 134*, 563-70.

Jensen, A. G., Chemali, M., Chapel, A., Kieffer-Jaquinod, S., Jadot, M., Garin, J., Journet, A. (2007). *Biochem. J.* 402, 449-58.

Jenuwein, T., Allis, C. D. (2001). *Science* 293(5532), 1074-80.

Kim, Y., Kim, S., Earnest, T. N., Hol, W. G. (2001). *J. Biol. Chem.* 277, 2823-9.

Kim, J. K., Yang, I. S., Shin, H. J., Cho, K. J., Ryu, E. K., Kim, S. H., Park, S. S., Kim, K. H. (2006). *Proc. Natl. Acad. Sci. U S A* 103, 1732-7.

Kim, Y., Yoon, K., Khang, Y., Turley, S., Hol, W. G. (2000). *Structure* 8, 1059-68.

Kollmann, K., Mutenda, K. E., Balleininger, M., Eckermann, E., von Figura, K., Schmidt, B., Lübke, T. (2005). *Proteomics* 5, 3966-78.

Kolodny, R., Koehl, P., Levitt, M. (2005). *J. Mol. Biol.* 346, 1173-1188.

Kornfeld, S., Mellmann, I. (1989). *Annu. Rev. Cell. Biol.* 5, 483-525.

Kos, J., Lah, T.T. (1998). *Oncol. Rep.* 5(6), 1349-61.

Krissinel, E., Henrick, K. (2004). *Acta Cryst. D* 60, 2256-2268.

Krissinel, E., Henrick, K. (2007). *J. Mol. Biol.* 372, 774-97.

Laskowski, R.A., Watson, J.D., Thornton, J.M. (2005). *Nucl. Acids Res.* 33, W89-W93.

Lin, R. J., Nagy, L., Inoue, S., Shao, W., Miller, W.H. Jr., Evans, R. M. (1998). *Nature* 391, 811-4.

Liu, Z. J., Vysotski, E. S., Chen, C. J., Rose, J. P., Lee, J., Wang, B. C. (2000). *Protein Science* 9, 2085-2093.

Lu, J.Y., Hofmann, S.L. (2006). *J. of Lipid Res.* 47, 1352-1357.

Lübke, T., Lobel, P., Sleat, D. E. (2009). *Biochim. Biophys. Acta,* in press.

Marti-Renom, M.A., Rossi, A., Al-Shahrour, F., Davis, F.P., Pieper, U., Dopazo, J, Sali A. (2007). *BMC Bioinformatics* 8(Suppl 4), S4.

Minucci, S., Nervi, C., Lo Coco, F., Pelicci, P. G. (2001). *Oncogene* 20(24), 3110-5.

Nixon, R. A., Cataldo, A. M. (2006). *J. Alzheimers Dis.* 9, 277-89.

Ohkuma, S., Poole, B. (1978). *Proc. Natl. Acad. Sci. U S A* 75(7), 3327-31.

Oinonen, C., Tikkanen, R., Rouvinen, J., Peltonen, L. (1995). *Nat. Struct. Biol.* 2, 1102-8.

Pal, D., Eisenberg, D. (2005). *Structure* 13, 121-130.

Panjikar, S., Parthasarathy, V., Lamzin, V.S., Weiss, M.S., Tucker, P.A. (2005). *Acta Cryst.D*61, 449-457.

Pazos, F., Sternberg, M.J.E. (2004). *Proc. Natl. Acad. Sci. USA* 101,14754-14759.

Peterson, C. L. (2002). *Mol. Cell* 9(5), 921-2.

Petrek, M., Otyepka, M., Banas, P., Kosinova, P., Koca, J. & Damborsky, J. (2006). *BMC Bioinformatics* 7, 316.

Petrek, M., Kosinova, P., Koca, J., Otyepka, M. (2007). *Structure* 15, 1357-63.

Porter, C.T., Bartlett, G.J., Thornton, J.M. (2004). *Nucl. Acids Res.* 32:D129-D133.

Ramagopal, U. A., Dauter, M., Dauter, Z. (2003). *Acta Cryst. D*59, 1020-1027.

Redfern, O.C., Dessailly, B., Orengo, C.A. (2008). *Curr. Opinion in Struct. Biol.* 18 (3), 394-402.

Redfern, O.C., Harrison, A., Dallman, T., Pearl, F.M., Orengo, C.A. (2007). *PLoS Comput. Biol.* 3, e232.

Roeser, D., Dickmanns, A., Gasow, K., Rudolph, M. G. (2005). *Acta Cryst.* D61, 1057-1066.

Saarela, J., Oinonen, C., Jalanko, A., Rouvinen, J., Peltonen, L. (2004). *Biochem. J.* 378, 363-71.

Saftig, P., Hetman, M., Schmahl, W., Weber, K., Heine, L., Mossmann, H., Köster, A., Hess, B., Evers, M., von Figura, K. (1995). *EMBO J.* 14(15), 3599-608.

Scriver, C. R., Beaudet, Sly, W.S, A.L., Childs, B., Kinzler, K.W., Vogelstein, B. (eds.) (2001). The Metabolic & Molecular Bases of Inherited Disease, Vol. III, 8th ed., McGraw-Hill, New York.

Sekar, K., Rajakannan, V., Velmurugan, D., Yamane, T., Thirumurugan, R., Dauter, M., Dauter, Z. (2004). *Acta Cryst.* D60, 1586-1590.

Shindyalov, I.N., Bourne, P.E. (1998). *Protein. Eng.* 11, 739-747.

Sleat, D. E., Wang, Y., Sohar, I., Lackland, H., Li, Y., Li, H., Zheng, H., Lobel, P. (2006). *Mol. Cell Proteomics* 5, 1942-56.

Sleat, D. E., Zheng, H., Qian, M., Lobel, P. (2006). *Mol. Cell Proteomics* 5, 686-701.

Sleat, D. E., Zheng, H., Lobel, P. (2007). *Biochim Biophys Acta 1774*, 368-72.

Sleat, D. E., Della Valle, M. C., Zheng, H., Moore, D. F., Lobel, P. (2008). *J. Proteome Res. 7*, 3010-21.

Song, L., Kalyanaraman, C., Fedorov, A.A., Fedorov, E.V., Glasner, M. E., Brown, S., Imker, H.J., Babbitt, P.C., Almo, S. C., Jacobson, M. P., Gerlt, J.A. (2007). *Nat. Chem. Biol.* 3, 486-491.

Strahl, B. D., Allis, C. D. (2000). *Nature* 403, 41-45.

Sykes, R. B., Cimarusti, C. M., Bonner, D. P., Bush, K., Floyd, D. M., Georgopapadakou, N. H., Koster, W. M., Liu, W. C., Parker, W. L., Principe, P. A., Rathnum, M. L., Slusarchyk, W. A., Trejo, W. H., Wells, J. S. (1981). *Nature* 291, 489-91.

Wade, P. A., Pruss , D., Wolffe, A. P. (1997). *Trends Biochem. Sci.* 22(4), 128-32.

Watanabe, N., Kitago, Y., Tanaka, I., Wang, J., Gu, Y., Zheng, C., Fan, H. (2005). *Acta Cryst. D*61, 1533-1540.

Watson, J.D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R.A., Thornton, J.M. (2007). *J. Mol. Biol.* 367, 1511-1522.

Weisel, M., Proschak, E., Schneider, G. (2007). *Chem. Cent. J.* 1:7.

Weiss, M. S., Sicker, T., Djinovic-Carugo, K., Hilgenfeld, R. (2001). *Acta Cryst. D*57, 689-695.

Weiss, M. S., Mander, G., Hedderich, R., Diederichs, K., Ermler, U., Warkentin, E. (2004). *Acta Cryst. D*60, 686–695.

Yang, C., Pflugrath, J. W. (2001). *Acta Cryst.* D57, 1480-1490.

Ye, Y., Godzik, A. (2003). *Bioinformatics* 19 (Suppl 2), ii246-ii255.

Yoshimori, T. (2002). *Cell structure and function* 27 (6), 401-402.

Zhang, Y., Reinberg, D. (2001). *Genes Dev.* 15, 2343-2360.

PART II

STRUCTURAL CHARACTERIZATION OF THE DNA REPAIR ENZYME MTH0212 FROM THE THERMOPHILIC ARCHAEON METHANOTHERMOBACTER THERMOAUTOTROPHICUS ALONE AS WELL AS IN COMPLEX WITH DIFFERENT SUBSTRATE DNAS

8. INTRODUCTION

The second major project of this PhD work was the determination of crystal structures of the enzyme Mth0212 alone as well as in complex with different substrate DNAs. Mth0212 plays a crucial role in DNA damage repair and exhibits a unique combination of catalytic activities including a 2`-deoxyuridine endonuclease activity which has not been detected in any other protein so far. To make the reader familiar with the biological context of Mth0212, in the following the present state of research in the field of DNA damage and repair is briefly summarized. Special emphasis will be put on the functional role of Mth0212 and related enzymes.

8.1. DAMAGES IN DNA

Deoxyribonucleic acid (DNA) ubiquitously bears the genetic information in all organisms known so far, in retroviruses its function is taken over by ribonucleic acid (RNA). The encoded information is crucial for survival of the respective cell since it is converted into protein or non-coding RNA molecules (e.g. ribosomal RNA, transfer RNA and micro RNA) via transcription and translation processes. In healthy cells, DNA is composed of four different kinds of deoxyribonucleotide monophosphates (dNMPs) consisting of a base, a sugar moiety and a phosphate group. The four commonly occurring bases are adenine and guanine as well as cytosine and thymine bases grouped as purine and pyrimidine bases, respectively. As a part of regulatory mechanisms the base and sugar moieties of each dNMP can be modified by specific enzymes. In contrast to these well-directed modifications, nucleotides can also undergo accidental changes referred to as DNA damage. Unrepaired DNA damage can lead to inaccurate RNA sequences as well as amino acid substitutions in proteins, which subsequently might result in severe cellular dysfunctions such as abnormal cellular regulatory mechanisms and enzyme inhibition. Finally, DNA damage might result in pathogenic phenotypes or even in apoptosis.

8.1.1. Causes of DNA damage

DNA damage is caused by environmental, chemical or enzymatic mechanisms and occurs quite frequently. This especially applies to thermophilic bacteria due to the pronounced base damaging effects of the extreme environmental temperature.

8.1.2. Kinds of DNA damage

The dominant form of DNA damage concerns the bases. In each cell, thousands of bases of the genomic DNA are damaged per day, thereof several hundred spontaneous hydrolytic deaminations of cytosines resulting in uridine residues (Bernstein, 1991; Ames et al., 1993; Lindahl, 1993; Fondufe-Mittendorf et al., 2002). Particularly in single-stranded DNA and therefore in actively transcribed genes and in replication forks deaminations occur frequently (Ames et al., 1993).

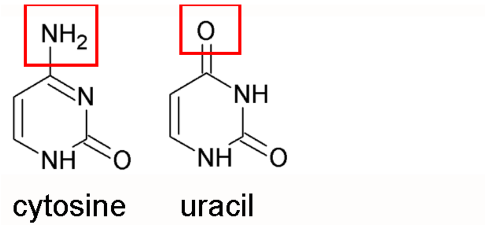


Fig. II- 1. Uracil as a result of cytosine deamination.

8.1.3. 2`-Deoxyuridine residues arising in DNA

The presence of uracil within DNA as a result of the spontaneous deamination of cytosine (Fig. II-1) is pre-mutagenic since it can lead to a CG → AT transition mutation. Besides, a uridine residue within DNA can arise from its wrong incorporation during the replication process due to a low level of the dTTP pool in the cell (Tye et al., 1977; Wist et al., 1978). The substitution of a uridine residue by the correct cytosine is critical for the maintenance of the genomic information, and therefore enzymes acting in the repair of uridines play an important role in all organisms.

8.2. DNA DAMAGE REPAIR

8.2.2. The base excision repair (BER) pathway

Many different DNA repair mechanisms for the recognition and removal of damaged DNA bases have been identified. Most frequently observed is the rapid damage-scanning mechanism which probes for both conformational deviations and local deformability of the DNA base stack that leads to enzyme-induced conformational changes at susceptible lesions and thus results in interactions with specific damaged bases.

Commonly, uridine residues are excised from DNA in two steps catalyzed by the successive action of a uracil DNA glycosylase (UDG) cleaving the N-glycosidic bond (Lindahl et al., 1977) and an AP endonuclease acting on the phosphodiester bond on the 5`-side of the generated apyrimidinic / apurinic site (AP site). Both enzymes belong to the base excision repair pathway (BER) (Fig. II-2) and are functionally substituted by Mth0212 as described below (8.2.4.).

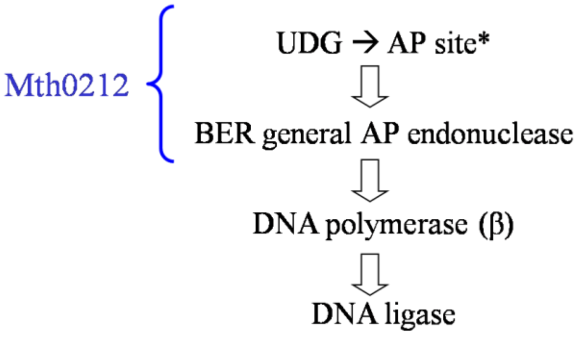


Fig. II-2. The base excision repair pathway (BER). For clarity, in black only the minimally required steps in human are shown. UDG = uracil DNA glycosylase, * AP site = apurinic / apyrimidinic site. Mth0212 takes over the function of both, UDG and a BER general AP endonuclease - but catalyzed in a single step - in the archaeon *M. thermoautotrophicus* which lacks a general UDG as outlined below.

8.2.3. Uracil DNA glycosylases generating apurinic / apyrimidinic (AP) sites

Recently, Mth0212 has been demonstrated to take over the role of a UDG in *Methanothermobacter thermoautotrophicus* (Schomacher et al., 2009). Thus, UDGs are shortly introduced in the following. These enzymes are conserved from bacteria to human (Aravind and Koonin, 2000). So far, only a few organisms are known to lack a homologue of the UDG superfamily - among them some larval stages of *Drosophila melanogaster* (Bekesi et al., 2007) and *M. thermoautotrophicus* (Georg et al., 2006; Schomacher et al., 2009). Singularly mono-functional UDGs have been identified so far. The enzyme class of bi-functional DNA glycosylases comprises only cytosine, thymidine, adenine as well as guanine DNA glycosylases (CDGs, TDGs, ADGs, GDGs) (reviewed in Lindahl et al., 1977). They specifically remove one of the four bases by the cleavage of the N-glycosidic bond and subsequently also the generated AP site. Pseudouridine containing DNA (ψ-DNA) cannot be processed by either mono- or bifunctional glycosylases including UDG (Lindahl et al., 1977), since both enzyme classes act on the glycosidic bond.

UDGs hydrolytically act on both single- (ss) and double-stranded (ds) DNA (reviewed in Lindahl et al., 1977; Pearl, 2000). The structure of human UDG in complex with dsDNA is represented in Fig. II-3.

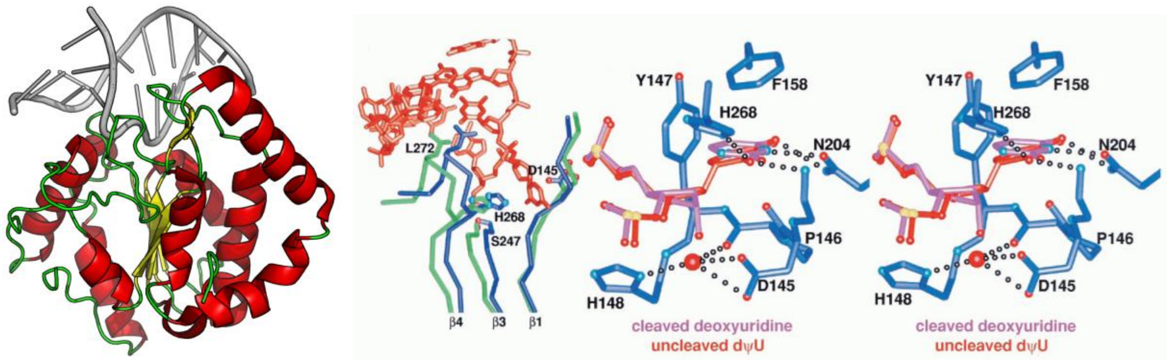


Fig. II-3. Uracil DNA glycosylase (UDG). Cartoon representation (PDB code 1EMH) coloured by secondary structure elements and close up onto the uracil binding pocket (Parikh et al., 2000).

8.2.4. AP endonucleases

Mth0212 serves as an AP endonuclease (Georg et al., 2006). Therefore, these proteins are introduced briefly in the following. AP endonucleases which excise AP sites from DNA are divided into two main classes (Levin & Demple, 1990). Class I enzymes cleave on the 3`-side of an AP site using a β-elimination mechanism, they are also referred to as AP lyases. In contrast, class II AP endonucleases hydrolytically cleave the phosphodiester bond on the 5`-side of an abasic site, thus initiating the BER pathway (Fig. II-4). Based on sequence similarity class II enzymes have been further sub-classified into two families represented by exonuclease III (ExoIII) and endonuclease IV (EndoIV) of *Escherichia coli* and thus termed accordingly. In *E. coli*, the main AP endonuclease is ExoIII exhibiting 90% of the total cellular AP endonuclease activity, whereas Endo IV accounts for the remaining 10% (Seeberg et al., 1995). The structure of endonuclease IV from *E. coli* was solved in complex with DNA and helped to understand the acquired mechanism which involves double-nucleotide flipping at the abasic site and three-metal-ion catalysis (Hosfield et al., 1999). Regarding the structurally completely unrelated ExoIII family, several homologues have been analyzed by means of X-ray crystallography - human Ape1 and Af_Exo from *Archaeoglobus fulgidus* also in complex with DNA - as described in detail in chapter 10.

The members of the ExoIII family of nucleases incise the phosphodiester backbone via an acid–base $SN_2(P)$ catalytic mechanism which is facilitated by divalent metal ions (Gerlt, 1993). In their active site, a catalytic triade is formed by a glutamate, an aspartate and a histidine residue (e.g. Glu96, Asp210 and His309 of Ape1). In particular for Ape1, several different catalytic mechanisms have been proposed, which are based on either a single or two metal ions(s) (reviewed in Wilson et al., 2001) (Fig. II-4).

The first proposed hypothesis (scheme 1 of Fig. II-4) states that His309 abstracts a proton from a water molecule to generate a hydroxyl anion serving as the nucleophile. The Mg^{2+} has an electron withdrawing effect and likely helps to orient the target phosphate (Barzilay et al., 1995), while the leaving group is stabilized by the protonated Asp210 (Erzberger & Wilson, 1999). Based on Ape1 crystal structures in complex with DNA, the mechanism was modified resulting in scheme 2 shown in Fig. II-4. Instead of His309, Asp210 abstracts a proton from a water molecule, whereas the protonated histidine orients the phosphate group of the DNA substituting for Mg^{2+} which stabilizes the leaving group in place of Asp210 (Mol et al., 2000). According to a more recently proposed hypothesis (scheme 3 of Fig. II-4) two magnesium ions are involved in the mechanism. One cation coordinates the generated nucleophilic hydroxyl ion (Beernink et al., 2001), whereas the second metal ion neutralizes the charge of the penta-covalent intermediate and / or stabilizes the 3` leaving group.

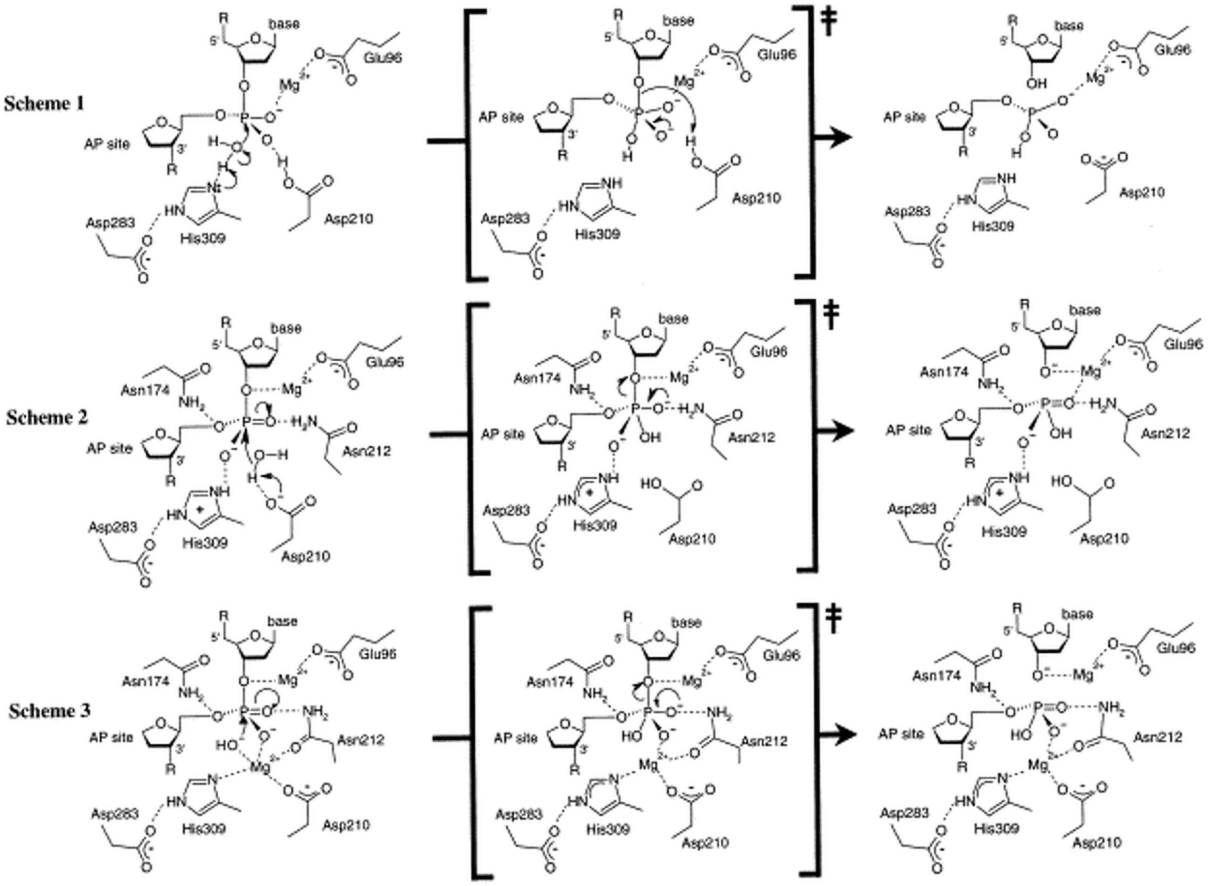


Fig. II-4. Proposed catalytic mechanisms of the removal of AP sites by the class II AP endonuclease Ape1. The bracketed central arrangement of each scheme represents the respective transition state / intermediate.

8.2.5. The exonuclease III homologue Mth0212 from the hyperthermophilic archaeon *Methanothermobacter thermoautotrophicus*

In the thermophilic archaeon *M. thermoautotrophicus*, which is devoid of a UDG, the function to initiate the repair of uracil bases in DNA is taken over by the Exo III homologue Mth0212 as demonstrated by immunodepletion studies (Fig. II-5) (Schomacher et al., 2009). However, the mechanisms of the common BER pathway and dU repair initiation by Mth0212 differ significantly. While the dU is ubiquitously removed in a two step procedure catalyzed by UDG and an AP endonuclease, Mth0212 removes the dU residue in a single step by a direct incision of the phosphodiester bond on the 5`-side of the dU. The advantage of the latter mechanism evidently is to avoid the occurrence of a toxic and pre-mutagenic AP site and not to require the transfer of the even more critical product that contains the AP site from one enzyme of the DNA repair machinery to the next. The ability of Mth012 to remove dU in a single step was evidenced with the help of pseudouridine (ψ) containing DNA substrates (Fig. II-6) (Georg et al., 2006). Mth0212 is able to excise ψ residues arguing against a cleavage of the glycosidic bond and thus excluding the possibility, that Mth0212 is a bi-functional DNA glycosylase.

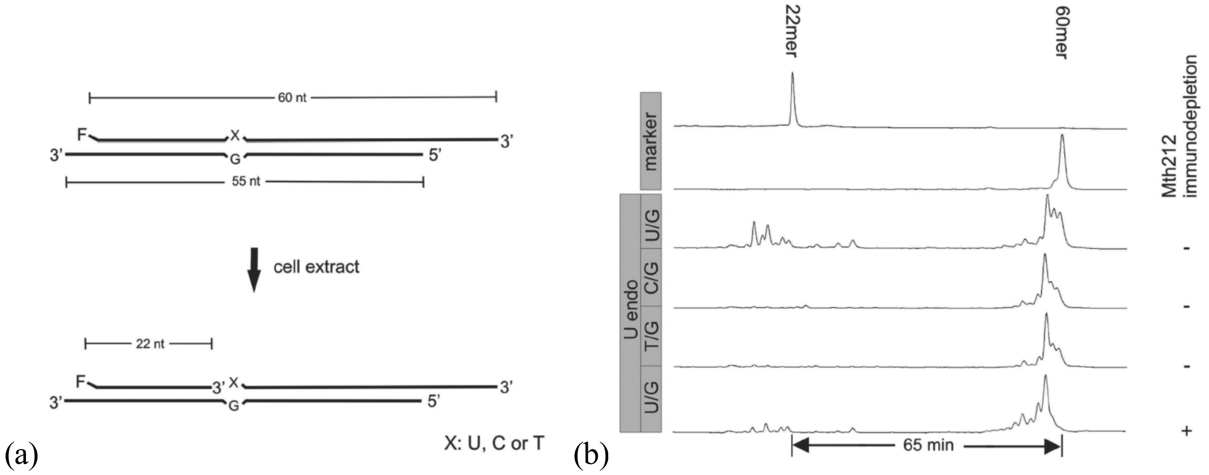


Fig. II-5. Demonstration of the dU endonuclease activity in cell extracts from *M. thermoautotrophicus* (Schomacher et al., 2009). F: fluorescein moiety, nt: nucleotides. Three substrates differing only in the identity of residue 'X' (U, C or T) were tested (3`-overhang of 10 and 5 nt in the labeled and unlabeled strand, respectively). (a) Schematic drawing of the substrate and expected reaction products. (b) Trackings of fluorescence readouts recorded by an ALF DNA sequencer with the run time difference between the marker oligonucleotides indicated below. Mth0212 immunodepletion: "+" and "-" denote pretreatment or no treatment of cell extract with α-Mth212 antibodies prior to the reaction.

In contrast to UDGs, Mth0212 can remove dU only from dsDNA (Fig. II-6). Since most of the cytosine deaminations occur in ssDNA (Lindahl, 1993; Krokan et al., 1997; Krokan et al., 2002; Barnes & Lindahl, 2004; Friedberg et al., 2006; reviewed in Yonekura et al., 2009) and thus the majority of dU residues is located therein, an enzyme responsible for the excision of dU from ssDNA might exist in the UDG lacking archaeon *M. thermoautotrophicus* as well in order to enable the cell to get rid of uracil in DNA and thus to ensure the maintenance of the correct genomic information. However, this catalytic function could not have been assigned to any enzyme so far making the role of Mth0212 in uracil DNA damage repair unique.

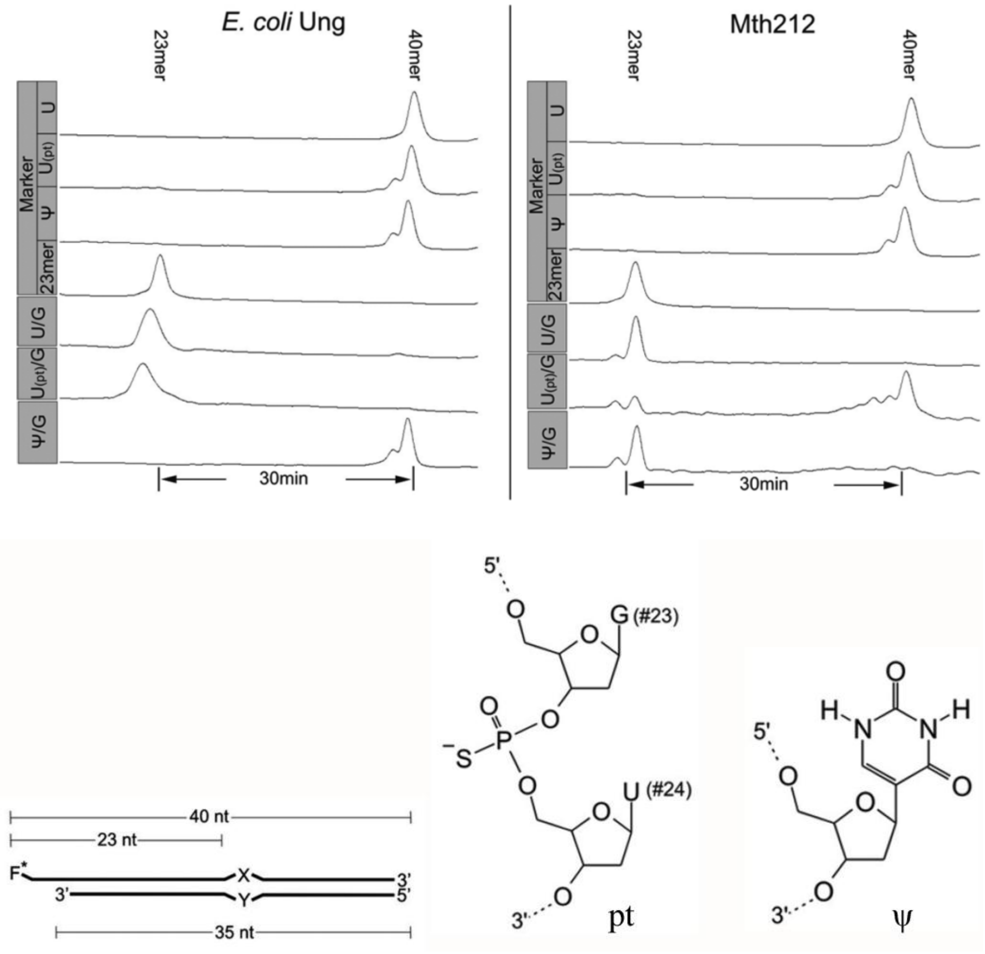


Fig. II-6. Discrimination of dU endonuclease activity of Mth0212 from possible glycosylase activity (Georg et al., 2006). At the bottom, a schematic representation of the substrates and the modifications in the sequence are shown (pt: phosphorothioate linkage between the nucleotides 23 and 24, ψ: 2`-deoxypseudouridine). The X/Y base pair was varied as denoted on the left side of each experiment. Assays using *E. coli* Ung and Mth212 are shown in the left and right panel, respectively.

A common feature between the structurally unrelated UDGs and Mth0212 is the specific recognition of dU, whereas cytidine and thymidine are not successfully recognized or at least not processed (Fig. II-5). Thus, the interactions of the enzymes with the damaged bases are very specific, although independent of any sequence context (Werner et al., 2000). This feature obviously contradicts the removal of any nucleotide type from the 3`-end by Mth0212, and such a combination might be only possible since DNA repair enzymes have evolved under the highest evolutionary pressure. With this knowledge in mind, the low *in vitro* affinity of Mth0212 for uridine in DNA is even more surprising and supports the hypothesis of an additional element enhancing the recognition efficiency *in vivo*.

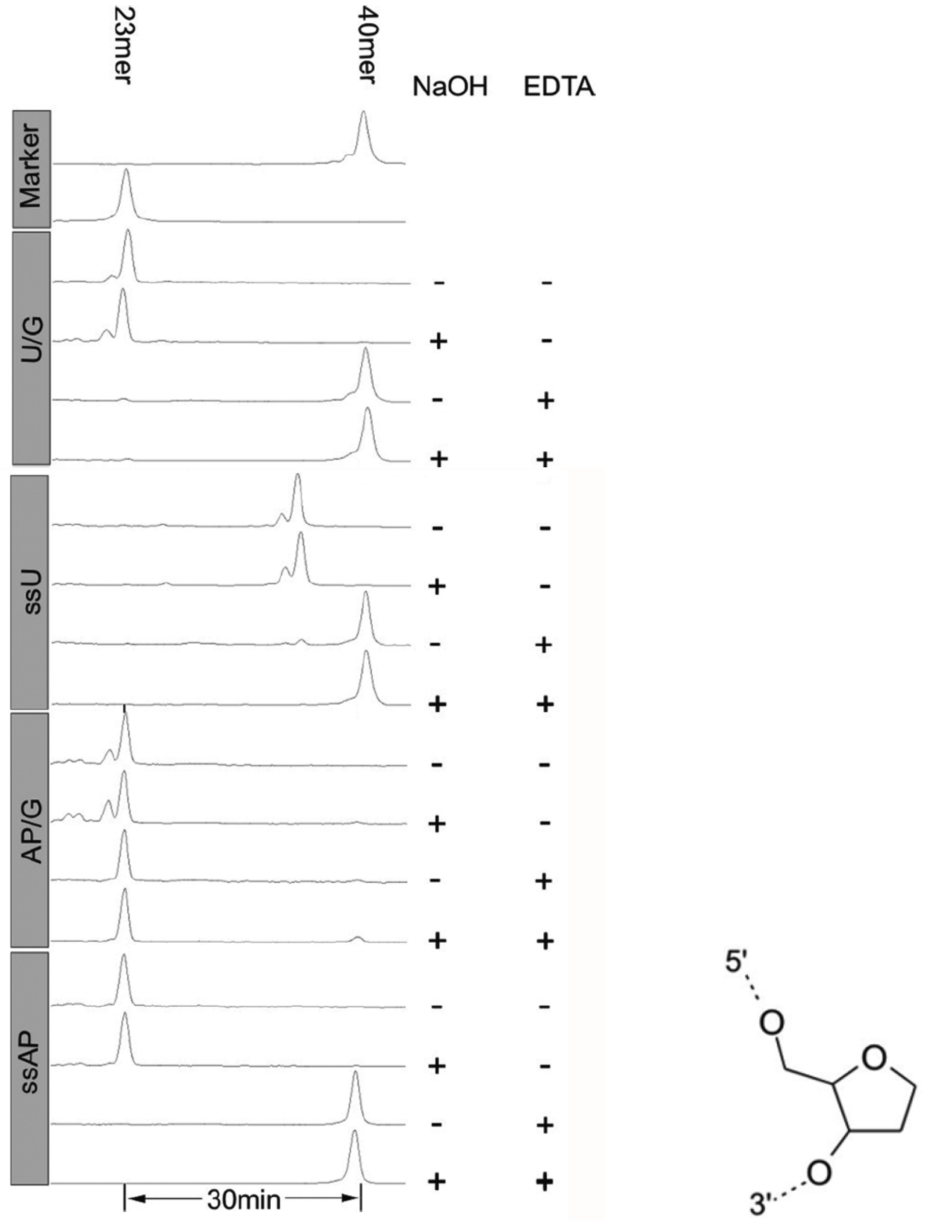


Fig. II-7. AP and dU endonuclease activity assay of Mth212 (Georg et al., 2006). Substrate as in Fig. II-6, ssDNA consisted singularly of the 40-mer. AP = chemically stable analogue of a base-free site in the DNA. The assay was performed as described previously (Georg et al., 2006). NaOH: post-reaction treatment with NaOH, EDTA: presence of 10 mM EDTA in the assay buffer to test the dependence on divalent metal ions.

Like all other members of the ExoIII family (reviewed in Demple & Harrison, 1994), Mth0212 acts as an AP endonuclease hydrolysing the phosphodiester backbone immediately 5` to AP sites in DNA (Fig. II-7) thus initiating the damage-general steps of the BER pathway. In addition to its function as an abasic endonuclease, Mth0212 serves as a 3`-5` exonuclease which becomes obvious in partial exonucleolytic degradation of dsDNA in activity assays as shown in Figs. II-5 – II-8.

Concerning the eponymous hallmark of the ExoII family, ExoIII from *E. coli* is the best studied enzyme, while the second conserved function, the AP endonuclease activity, has been extensively studied in human Ape1. Thus, most aspects of the Mth0212 structures are discussed in comparison with these two enzymes.

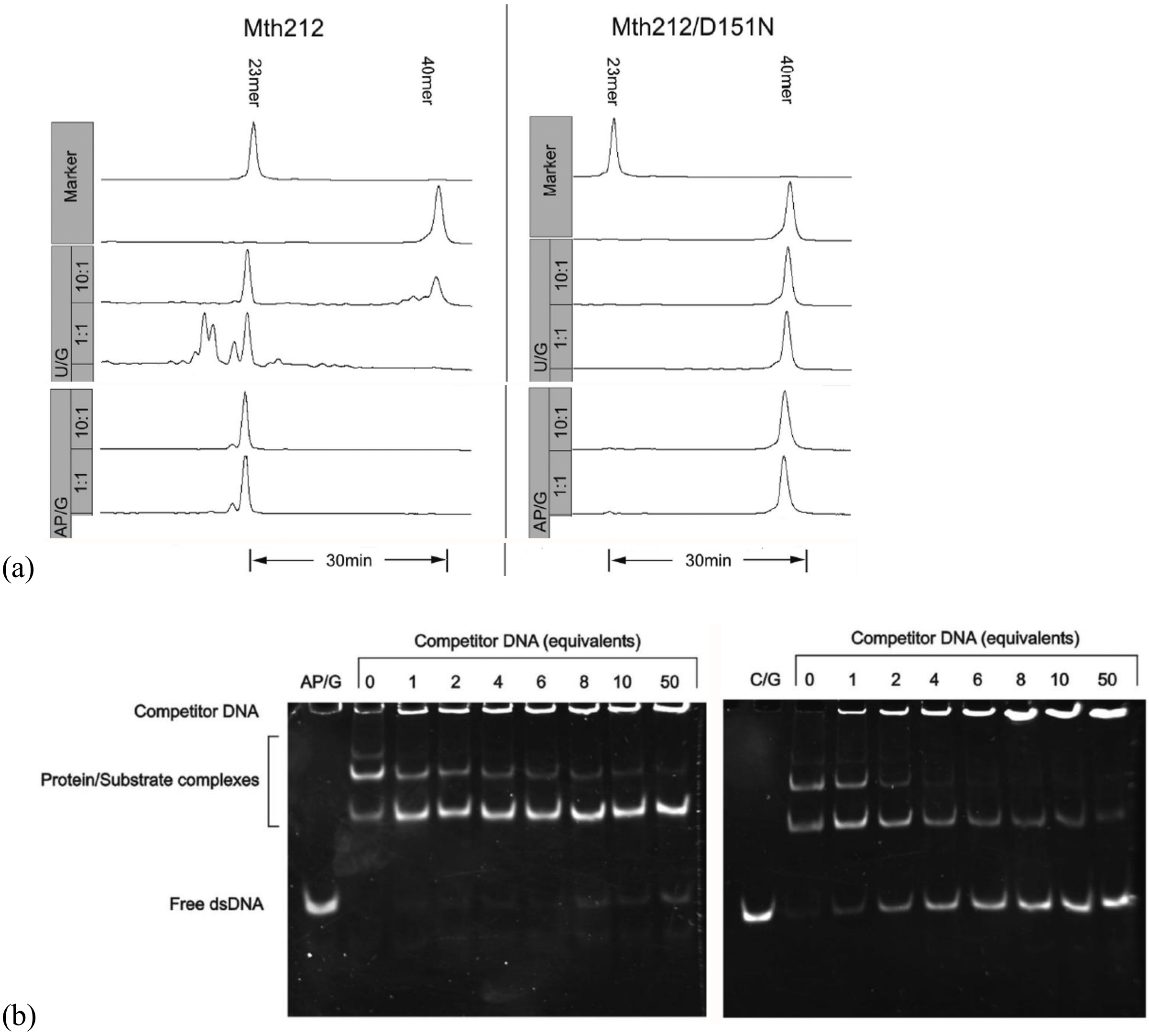


Fig. II-8. Characterization of the rationally designed mutant Mth0212(D151N) in comparison with Mth212(WT) (Georg et al., 2006). Substrates as in Fig. II-6. (a) Activity assays. Molar ratios of substrate:enzyme are denoted in the grey bars. (b) Electrophoretic mobility shift assays (EMSAs) using double-stranded oligonucleotides. As reference, the free substrates were loaded in the leftmost lanes. Amounts of competitor DNA (plasmid DNA) are indicated as nucleotide equivalents relative to substrate oligonucleotides.

It has not been examined so far whether all the other activities of a classical ExoIII homologue such as the 3`-phosphodiesterase and a 3`-phosphatase activity significantly present in the homologous LAMP from *Leishmania major* (Vidal et al., 2007) or an RNaseH activity evidenced in Exo III (reviewed in Demple & Harrison, 1994) and Ape1 (Barzilay et al., 1995) are carried out by Mth0212 as well. However, Mth0212 combines the enzymological hallmarks of the ExoIII family with the uridine endonuclease activity. The efficiency of the removal of a dU residue seems to depend on the nature of the opposing nucleotide (laboratory of Prof. H.-J. Fritz, Göttingen, unpublished data, personal communication). Asp151 of Mth0212 has been shown to be required for all nucleolytic activities (Fig. II-8).

Theoretical projections of the Mth0212 amino acid sequence onto 3D structures of the homologous repair enzyme ExoIII lacking the uridine endonuclease activity could not shed light on the structural basis for the expanded substrate specificity (Georg et al., 2006). The recognition of 2`-deoxyuridine (dU) as the specific substrate of the dU endonuclease activity of Mth0212 seems to stand in contrast to its broad spectrum binding capability regarding the classical DNA repair activities shared with the other members of the ExoIII family. In order to obtain a deeper insight into the molecular basis of the dU specificity, knowledge about structural details of Mth0212 is crucial.

Within the scope of my PhD thesis the crystal structures of Mth0212 alone as well as in complex with different substrate DNAs have been determined. The rationally designed mutant Mth0212(D151N) used for some co-crystallization trials with DNA oligonucleotides has been shown to be severely deficient in DNA uridine endonuclease, AP endonuclease and 3′→5` exonuclease activities (Fig. II-8a), but properly folded, since it binds specifically to AP/G substrates (Fig. II-8b). The results of the activity assays with this mutant suggest that all activities are carried out by a single active site (Georg et al., 2006).

9. OBJECTIVES

Theoretical projections of the Mth0212 amino acid sequence onto 3D structures of the homologous repair enzyme ExoIII lacking the dU endonuclease activity could not shed light on the structural basis for the expanded substrate specificity (Georg, 2006 #61). Thus, to gain insight into the molecular basis for the additional dU recognition, we set out to determine the crystal structures of Mth0212 alone as well as in complex with different substrate DNAs. Altough the main focus was on the dU specificity, contributions to the still on-going discussion regarding the number of bound divalent metal ions required by members of the ExoIII family were an issue as well. Additionally, the crystallographic approach was aimed to help in understanding how the 3`-5` exocnuclease, AP and dU endonuclease activities can be carried out at the same active site.

10. 3`-5` EXO COMPETES WITH 2`-DEOXYURIDINE ENDONUCLEASE ACTIVITY: STRUCTURES OF THE ARCHAEAL EXOIII HOMOLOGUE MTH0212 WITHOUT AND IN COMPLEX WITH DIFFERENT SUBSTRATE DNAS SHED LIGHT ON THE UNIQUE COMBINATION OF DNA REPAIR ACTIVITIES

10.1. OBJECTIVES AND AUTHORS` CONTRIBUTIONS

This PhD thesis presents five structures of the DNA repair enzyme in its apo form and nine structures of the protein in complex with DNA. The manuscript intended for publication of the major results is currently prepared and attached in the following in its actual state. The apo structures belong to different crystallographic space groups and contain the wild-type protein with bound magnesium or manganese ion(s) and the rationally designed mutants K116A and D151N which are catalytically active and inactive, respectively. The Mth0212-DNA complex structures comprise dsDNA of varying lengths and sequence exhibiting blunt as well as sticky ends or 4 nt long ssDNA. All complex structures reflect the eponymous exonucleolytic activity of Mth0212.

Detailed analysis of the diverse complexes in concert with comparison amongst each other and with structures of ExoIII homologues, in particular with human Ape1-DNA complexes, provided insight into the mode of divalent metal ion and DNA binding at the active site and yielded a hypothesis about the presumable 2`-deoxyuridine recognition.

For an overview, the PDB-IDs and titles of each structure are attached as an additional page in the back of this thesis.

The protein used for determination of the structures 3FZI, 3G0A and 3G91 was prepared in the laboratory of Prof. Dr. Hans-Joachim Fritz by Lars Schomacher and Elena Ciirdaeva, respectively. For the first two structures, crystallization was performed by Dr. Achim Dickmanns and Annette Berndt. Data collection and processing as well as initial Molecular Replacement and refinement procedures were contributed by Dr. Achim Dickmanns. My contribution with Prof. Dr. Ralf Ficner as my supervisor concerns the purification and crystallization of two additional Mth0212 variants alone and of all nine Mth0212-DNA complexes for the remaining eleven structures. Moreover, I have performed subsequent data collection, refinement and PDB depositions as well as structural analysis and comparison of all 14 structures. In the course of the studies, regular discussions with all contributors as well as with Swetlana Ber from the laboratory of Prof. Fritz is acknowledged. They guided the selection of co-crystallized DNA sequences. Based on the results of the structural analysis performed in this PhD work currently mutational studies are performed in the lab of Prof. Fritz. Regarding the structure 3GA6, discussions with Dr. Regine-Herbst Irmer helped to solve the crystallographic problem of twinning.

10.2. MANUSCRIPT IN PREPARATION: "3`-5` EXO COMPETES WITH 2`-DEOXYURIDINE ENDONUCLEASE FUNCTION IN MTH0212-DNA COMPLEX STRUCTURES"

For clarity reasons, the figure captions are not included in the text of the manuscript, but presented in the subsequent figure section below each respective figure.

3`-5` EXO COMPETES WITH 2`-DEOXYURIDINE ENDONUCLEASE FUNCTION IN MTH0212-DNA COMPLEX STRUCTURES

Kristina Lakomek[1], Achim Dickmanns[1], Elena Ciirdaeva[2], Lars Schomacher[2], Hans-Joachim Fritz[2], Ralf Ficner[1]*

[1] Department of Molecular Structural Biology, Institute of Microbiology and Genetics, Georg-August University Goettingen, Justus-von-Liebig-Weg 11, D-37077 Goettingen, Germany

[2] Department of Molecular Genetics and Preparative Molecular Biology, Institute of Microbiology and Genetics, Georg-August University Goettingen, Grisebachstr. 8, D-37077 Goettingen, Germany

* corresponding author
fax: +49 551 3914082
tel: +49 551 3914071
E-mail-adress: rficner@gwdg.de

Synopsis

Crystal structures of the ExoIII homolog Mth0212 from the thermophilic archaeon *Methanothermobacter thermoautotrophicus* in its apo form and in complex with different DNA substrates have identified key residues for the classical ExoIII-like activities as well as the additional 2`-deoxyuridine endonucleolytic function.

Abstract

The exonuclease III homolog Mth0212 of the thermophilic archaeon *Methanothermobacter thermoautotrophicus* displays a unique combination of DNA repair activities with an expanded substrate specificity. In addition to the biochemical hallmarks of the ExoIII family such as a 3`-5`exonuclease and an AP endonuclease function, it exhibits a 2`-deoxyuridine (dU) endonuclease activity initiating the removal of dU residues from double-stranded DNA. The reliable repair of this common and pre-mutagenic base damage is crucial for the correct maintenance of the genomic information. Mth0212 compensates for the lack of a completely unrelated, otherwise almost ubiquitous uracil DNA glycosylase (UDG) in *M. thermoautotrophicus* and substitutes a catalytic reaction catalyzed by two enzymes by a single step mechanism.

Thus, the general statement that any uracil base in DNA is removed by UDG has to be partly corrected. At least in *M. thermoautotrophicus* an alternative strategy has evolved.

In order to understand how the different nucleolytic activities of Mth0212 can be accomplished in a single active site and to get a deeper insight into the recognition of dU, we characterized the enzyme by means of X-ray crystallography. Five crystal structures of Mth0212 alone as well as nine structures in complex with different DNA substrates and products were determined to a resolution between 1.2 and 3.1 Å using wild-type or mutant proteins (K116A and D151N). Mth0212-DNA contacts across the whole interaction surface of the enzyme were revealed.

The complex structures show Mth0212 in its eponymous exonucleolytic function and give no direct insight into the mechanism of the additional dU endonuclease activity. Detailed comparison of the protein-DNA interactions and the binding of additional ligands such as phosphate and Mg^{2+} ions with each other and with homologous structures provided putative explanations for the unique combination of DNA repair activities.

Most likely tiny structural differences result in the expanded substrate spectrum of Mth0212 compared with ExoIII homologs only exhibiting the biochemical hallmarks of this nuclease family. The insertion of the side chain of Arg209 into the DNA helical base stack which resembles interactions observed in human UDG in concert with Ser171, Asn114, Asn153 and in particular Lys125 in the substrate binding pocket are supposed to play key roles in dU recognition.

Introduction

In each cell, thousands of bases of the genomic DNA are damaged per day, thereof several hundred spontaneous hydrolytic deaminations of cytosines resulting in 2`-deoxyuridine residues (dU) [1; 2; 3; 4]. Besides, dU within DNA can arise from its wrong incorporation during the replication process due to a low level of the dTTP pool in the cell [5; 6]. The presence of dU within DNA is pre-mutagenic, and substitution by the correct cytosine or thymidine is critical for the maintenance of the genomic information. Therefore enzymes acting in the dU repair play an important role in all organisms.

Commonly, dU is excised from DNA in two steps catalyzed by the successive action of a uracil DNA glycosylase (UDG) cleaving the N-glycosidic bond [7] and an AP endonuclease acting on the phosphodiester bond on the 5`-side of the generated apyrimidinic / apurinic site (AP site). Both enzymes are part of the base excision repair (BER) pathway which represents the primary defense against all major forms of DNA base damage.

UDGs are conserved from bacteria to human [8]. So far, only very few organisms are known to lack a homolog of the UDG superfamily - among them *M. thermoautotrophicus* [9; 10]. In *M. thermoautotrophicus*, the function to initiate the repair of uracil bases in DNA is taken over by the Exo III homolog Mth0212 [9]. However, the mechanisms of the two repair pathways differ significantly. The ubiquitous two step procedure catalyzed by UDG and an AP endonuclease is replaced by the direct incision of the phosphodiester bond on the 5`-side of the dU by Mth0212 in a single step [9]. The advantage of the latter mechanism evidently is to avoid the occurrence of a product containing the even more toxic and pre-mutagenic AP site and its transfer to the next enzyme of the BER pathway. The ability of Mth012 to remove dU in a single step was demonstrated with the help of pseudouridine (ψ) containing DNA substrates [9]. Mth0212 can remove dU only from dsDNA with a dependence of the efficiency on the nature of the opposing nucleotide [9]. However, since an equivalent catalytic function could not have been assigned to any enzyme so far, the role of Mth0212 in uracil DNA damage repair is unique.

Like all other members of the ExoIII family [11; 12; 13; 14; 15; 16] (reviewed in [17]), Mth0212 acts as an AP endonuclease hydrolysing the DNA phosphodiester backbone immediately 5` to AP sites [18; 19; 20]. Thus it initiates the general steps of the BER pathway which are the same for damaged bases of all four common nucleotides as reviewed in [17 20]. In addition to its function

as an abasic endonuclease, Mth0212 serves as a 3`-5` exonuclease [9; 18; 20]. Whether all the other activities of a classical ExoIII homolog such as the 3`-phosphodiesterase, 3`-phosphatase [21] or an RNaseH activity (reviewed in [17; 22]) are carried out by Mth0212 as well, has not been examined so far.
In particular the homologs Ape1 and ExoIII have been studied in detail [22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33] (reviewed in [17; 34]). These and several further ExoIII homologs have been structurally analyzed by means of X-ray crystallography as summarized in Table 1. The three-dimensional overall structure of Mth0212 additionally resembles that of bovine DNaseI and the human targeting LINE-1 retrotransposon endonuclease (Tab. 1). On the protein sequence level, Mth0212 shows the highest similarity to NAPE1 with 41 % identity of 253 aligned amino acid residues. Ape1 and Af_Exo which exhibit 40 % and 30 % sequence identity are the only homologs that have been studied in complex with DNA [35; 36]. While all Ape1-DNA complexes represent the AP endonuclease function of the enzyme, Af_ExoIII is bound to the 3`-end of dsDNA, but lacks both a phosphate group and a catalytically essential magnesium ion in the active site.

However, Mth0212 exhibits a unique combination of DNA repair activities comprising the 3`-5` exonucleolytic, the AP and additionally the dU endonucleolytic functions. They are carried out in a single active site [9] as shown by severe defects concerning all activities in the rationally designed mutant Mth0212 (D151N) [9]. The subtle discrimination between the different pyrimidine bases in order to prevent the removal of cytidine and thymidine residues by Mth0212 is shared with UDG. The structures of both enzymes are completely unrelated. Thus, most likely the dU recognition has evolved independently from each other raising the interest how the very specific interactions are achieved by Mth0212 and to compare the two recognition mechanisms. This in particular applies since the ExoIII homolog has to take an additional challenge, namely the combination with its concurrent ability to remove any nucleotide from the 3`-end of dsDNA excluding a ubiquitous exclusion of the four common nucleotides.

Theoretical projections of the Mth0212 amino acid sequence onto 3D structures of homologous ExoIII enzymes could not shed light on the structural basis for the expanded substrate specificity [9]. Thus, to get a deeper insight into the molecular basis for the additional dU recognition, we determined the crystal structures of Mth0212 alone as well as in complex with different substrate DNAs. The complex structures help to understand how the substrates of the 3`-5` exocnucleolytic as well as the AP and dU endonucleolytic incision are recognized by the same set of amino acid residues involving a significant distortion of the bound DNA duplices. Additionally, this work contributes to the still on-going discussion regarding the number of bound divalent metal ions which are catalytically essential as well as their coordination sites and thus indicates a catalytic mechanism of Mth0212 more resembling that initially suggested for Ape1 rather than a cleavage involving two metal ions.

Results and discussion

Structure determination

In order to understand, how the different DNA repair activities of Mth0212 can be accomplished in a single active site, the nuclease was crystallized alone as well as in complex with distinct substrate DNAs (Fig. 1). To prevent the degradation of the DNA during crystallization, the rationally designed catalytically inactive mutant Mth0212(D151N) was incubated with common DNA. Alternatively, non-cleavable DNA oligonucleotides were used for complex formation with the wild-type protein or the rationally designed active mutant K116A lacking one flexible side chain, respectively. The type of protein and the DNA sequences used for co-crystallization are listed separately for each determined structure in tables 2 and 3. In general, the same expression and purification procedures were used for WT and mutant enzymes. They are based on the protocol reported previously [9] and were optimized during the course of the various crystallization experiments and modified as far as required for each individual sub-project. Details are described in the materials and methods section.

The crystallization conditions for both Mth0212 alone and for the distinct Mth0212-DNA complexes are summarized in Table 3, while Table 4 lists the respective space group and unit cell parameters. The structures were solved by means of Molecular Replacement (MR) as described in Materials and Methods.

Four Mth0212-DNA complexes comprise the wild-type protein (3G2C, 3G3Y, 3G3C, 3G4T), whereas the remaining five complex structures contain the rationally designed catalytically inactive mutant Mth0212(D151N) (3G2D, 3G38, 3G0R, 3G00, 3GA6). Seven structures show Mth0212 in interaction with double-stranded DNA, only two structures (3G2C, 3G3Y) represent contacts with single-stranded DNA, which both are four nucleotides in length (Fig. 1). For the structures 3G2D, 3G38, 3G0R and 3GA6, GC-rich DNA sequences of 8-12 bp were used. In contrast, the 9 bp sequences yielding 3G2C, 3G3Y, 3G3C, 3G4T, 3G00 and 3G1K commonly were palindromic with the same amount of G-C and A-T pairs as follows. The more flexible A-T base pairs flanked the dU residue in the central region in order to facilitate the putative flipping of dU, whereas two more rigid G-C base pairs were located at both ends to stabilize the double helix for improved crystal contact formation. The base opposite of dU was varied between G, C and T. As indicated in Tab. 2, a bulky fluorescent dye was attached to both 3`-ends of some dsDNAs. It was intended to prevent these ends from binding to Mth0212 and exonucleolytic degradation. Also after transfer into cryo protecting solution, crystals displayed the pink colour characteristic of the dye interpreted as an unambiguous sign of successful complex crystallization including the label. Surprisingly, some nucleotides were not defined in the electron density and due to DNA binding mode and position seemed to have been degraded.

The structure 3GA6 was refined as a twinned crystal in the space group $P2_1$ with PHENIX [37] using the twin law “l,-k,h”, and analysis as well as refinement procedures are described in supplementary material and methods.

Structures of Mth0212 wild-type and mutant proteins alone (3FZI, 3G1K, 3G8V, 3G91, 2G0A)

Five structures of Mth0212 alone were obtained. Three apo structures represent the wild-type protein crystallized in the hexagonal space group $P6_5$ (3FZI, 3G0A) or in the monoclinic space group $P2_1$ (3G1K) and were refined to a resolution of 1.9, 2.6 and 3.1 Å, respectively. While Mth0212 binds a single naturally occurring magnesium ion in 3FZI, two anomalous scattering manganese ions are included in the structure 3G0A. The metal ion coordination is discussed separately further below. As expected due to observations in homologous proteins, the D151N mutation seems to result in the loss of metal ion binding ability of Mth0212. Accordingly, no metal ion could be placed in the structure 3G8V. In contrast, the respective mutation did not cause any significant changes in the active site arrangement or in the conformation of the DNA binding residues when compared with the wild-type apo structures 3FZI and 3G1K. The same applies to the K116A mutation, which is represented in the 1.2 Å resolution structure 3G91.

Overall structure of Mth0212

Besides the recently published structure of Af_ExoIII, the Mth0212 crystal structure represents the only structure of an ExoIII homolog of the archaeal domain. Its overall structure resembles that of other ExoIII homologs as discussed in detail further below. Thus, a detailed description of the secondary structure elements is only provided as supplemental information. Briefly, the globular α/β protein consists of two domains with similar topologies each comprising a six-stranded β-sheet surrounded by α-helices. The three specific loop regions of Mth0212 involved in DNA binding, which are shared by all ExoIII homologs but lacking in the structurally related DNaseI, comprise the residues Asn114-Arg121 (loop I), Arg163-Gly172 (loop II) and Trp205-Trp219 (loop III). They are marked with arrows in Fig. S3.

In a few structures, several residues of the C-terminal affinity tag (Leu-Glu-His_6) are clearly defined in the electron density and involved in intermolecular hydrogen bonds at crystal contacts. In the high resolution apo structure 3G91, direct hydrogen bonds are formed between the side chains of Glu259 and His261 and the backbone carbonyl oxygen of Gly201 and the Nδ2 atom of Asn216 of a symmetry equivalent molecule, respectively. In contrast, none of the complex structures include more than the first two residues of the expression tag. An analysis of contacts between symmetry equivalent molecules in the crystal using PISA [38] indicates Mth0212 to exist as a monomer. These calculations are in accordance with the protein eluting from the gel filtration column at a volume which corresponds to a molecular weight of about 31 kDa.

Metal ion coordination

When Mth0212(WT) crystals were soaked with $MnCl_2$, two divalent metal cations could be identified unambiguously based on their anomalous scattering (3G0A). They are shown in Fig. 2 and subsequently referred to as metal coordination sites A and B, which are located about 5.8 Å away from each other in the active site. However, the two manganese ions included in the structure 3G0A might exhibit different occupancies. The peak in the anomalous difference map at site A disappears only at a level of 11.7 σ and thus is significantly larger than that at site B with a height of 5.1 σ. Nevertheless, also the latter is strong enough to place a manganese ion at site B. Site A is formed by the side chains of

Asn12 and Glu38, while the Mn^{2+} ion at site B is coordinated by the side chains of Asp151, Asn153 and His248. Additionally, one or two water molecules interact with the metal ions resulting in a four- and six-fold coordination at site A and B due to the bi-dentate ligand binding of Glu38 and Asp151, respectively. The metal ion – protein ligand distances amount to only 2.15-2.53 Å at site A, whereas the weaker binding at site B might be explained by the greater distances ranging between 2.33-3.55 Å.

In none of the other structures which were yielded in presence of magnesium ions two divalent metal ions are bound in the active site of Mth0212 (Fig. 1). Only a single Mg^{2+} ion is coordinated, namely by Asn12 and Glu38 at binding site A. Interestingly, it was observed only when Asp151 was present, although this residue is part of the other metal coordination site. In the structures of the mutant Mth0212(D151N) a metal ion was not identified in the active site – neither in the apo structure (3G8V) nor in complex structures with DNA (3G2D, 3G38, 3G0R, 3G00, 3GA6). The Mg^{2+} ion in structure 3FZI must have been derived already from the expression medium indicating a strong binding which kept the cation in the active site throughout the purification procedure despite of the use of buffers lacking magnesium.

Structures of Mth0212 in complex with different substrate DNAs

In addition to the five structures of Mth0212 alone described above, nine Mth0212-DNA complex structures were obtained using WT (3G2C, 3G3Y, 3G3C, 3G4T) and mutant protein (D151N: 3G2D, 3G38, 3G0R, 3G00, 3GA6), respectively. They differ in the binding of the respective DNA substrates as contain DNA molecules of various lengths exhibiting nucleotide overhangs or distorted base pairs in some structures (Fig. 1, 3) (according to analysis with AMIGOS [39] and described in material and methods). Moreover, due to the variety in the length and sequence of the DNA, they crystallized in different monoclinic, orthorhombic and hexagonal space groups with diverse unit cells dimensions (for details cf. Tab. 4).

Although Mth0212 acts as a monomer, the binding stoichiometry of protein:dsDNA is 2:1 in all structures except for 3G2D, 3G38 and 3GA6. In all complexes in which two Mth0212 molecules bind only a single DNA duplex, the binding interface is located on the same domain of the enzyme comprising Met117 and Lys165. These observations led to the conclusion, that this domain of Mth0212 has a significantly higher affinity to the DNA sugar-phosphate backbone than the second domain and consequently provides the main DNA binding interface of the protein. Surprisingly, the surface potential of this domain appears to be less positively charged than that of the domain including Lys66 (Fig. 4a).

The complex structure 3GA6 was yielded upon incubation of the catalytically inactive mutant Mth0212(D151N) with 12 bp blunt-ended dsDNA containing a 2`-deoxyuridine / guanine mismatch. The asymmetric unit comprises two protein molecules, each of them bound to two dsDNA molecules which are both shared between the two protein molecules (or their symmetry equivalents). All amino acid residues involved in contacts to the DNA substrates in any of the eight other Mth0212-DNA complex structures interact also with the DNA in the structure 3GA6. Thus, the DNA contacting residues as well as the proposed dU recognition mechanism mainly will be described for this structure.

The structure 3GA6 represents two different Mth0212(D151N)-DNA complexes with a one nucleotide long overhang at one end of the 10 bp and 11 bp dsDNA molecule, respectively. The blunt ends of both DNA helices are bound close to the active site of Mth0212 separated by the insertion of Arg209 into the DNA duplex and forming stacking interactions via its side chain guanidinium group (Fig. 4b).

Contacts to the DNA duplices I and II, respectively, are mediated by the residues Lys116, Arg163, Asn167, Arg209, Thr210, Arg215 as well as Trp219 and Gly13, Ala16, Arg19, Lys20, Gln45, Arg65, Lys66, Tyr68, Met117 as well as Tyr208 (Fig. 4c). The amino acids and nucleotides interact via hydrogen bonds except for cation-π-stacking formed between the terminal base pair and the side chain of Arg209 and pure Van-der-Waals interactions involving Gly13 and Met117. Most residues interact directly, while Thr210 forms a water-mediated hydrogen bond to a nucleotide of a terminal base pair. Arg15 and Arg215 each contact both a phosphate moiety and the O3` atom of the 5`-adjacent nucleotide.

In the active site pocket of protein molecule B, a glycerol molecule from the cryo protecting solution participates in an extensive hydrogen bonding network directly involving the side chains of Lys125, Asn153, Ser171 and the carbonyl oxygen of Asn153 as well as water-mediated contacts to both the side chain of Trp205 and the 3`-terminal phosphate moiety of one DNA strand.

Biochemical studies have shown that the ds-specific 3`-5` exonuclease activity of Mth0212 can be significantly reduced by the use of ssDNA-overhangs at the 3`-ends [20]. Although it was not clear whether the overhang just inhibits the nucleolytic incision of the DNA or already prevents the productive binding of the enzyme to the 3`-end, an 11 bp dsDNA with a 2 nt long at the 3`-end of one strand was used with intent to block the binding in an exonucleolytic manner. However, the structure 3G0R shows the exo-complex as well with two protein molecules bound to both ends of a dsDNA helix in the asymmetric unit. The DNA helix consists of eight common Watson-Crick base pairs as well as one distorted base pair and 1-2 nt long tilted ssDNA overhangs at the ends, respectively.

The penultimate residue of the 11 nt long DNA strand, cytosine 2 of chain K, is flipped out of the helix. Due to this extra-helical position, the 3`-adjacent nucleotide, cytosine K1, occupies the plane next to the final Watson-Crick base pair but takes a position between the base stacks of both strands still feasible for stacking interactions and reminding of the inserted Arg209 side chain in the structure 3GA6. The last nucleotide of the complementary strand which afore most likely had formed a base pair with the flipped-out cytosine, remains unpaired in a strongly tilted conformation stabilized by hydrogen bonds between the base and the O4` and O5` deoxyribose atoms of cytosine K1. The distorted conformation of the terminal nucleotides which do not form a Watson-Crick base pair prevent the stacking interaction with the side chain of Arg209 observed in 3GA6. The peptide bond between Arg209 and Thr210 is flipped in both molecules of the asymmetric unit.

At metal binding site A, a water molecule rather than a magnesium ion is located in the structure 3G0R bridging the side chains of Glu38 and Asp247 with the scissile backbone phosphate of the DNA (Fig. 5a).

Under the crystallization conditions of the complexes 3GA6 and 3G0R, Mth0212 did not display its dU endonuclease activity (unpublished data, Elena Ciirdaeva, laboratory of Prof. Fritz). Therefore for subsequent co-crystallization experiments (3G2D, 3G38) the salt concentration which seemed to inhibit this activity was reduced both in the complex solution and in the crystallization reservoir solutions by preparing special low salt screens for initial crystallization trials. Finally (3G2C, 3G3Y, 3G3C, 3G4T, 3G00), the complex with DNA was formed in the same low salt concentration buffer system which enabled optimal dU activity in biochemical assays (20 mM KH_2PO_4/K_2HPO_4 pH 7.0, 50 mM KCl, 1 mM $MgCl_2$) [9].

Surprisingly, in the Mth0212(D151N) –DNA complex structure 3G2D, no phosphate anion is bound in the active site. Regarding protein molecule A, the lack might be explained by the binding of a polyethylene glycol molecule at the active site which mediates contacts between Mth0212 and the DNA, but there is no space filling solvent molecule in protein molecule B.

The structure 3G38 contains one Mth0212(D151N) and one 8 bp blunt-ended dsDNA molecule. The duplex exhibits a phosphorylated 3`-end, which is bound in the active site coordinated by Asn10, Glu38, Tyr111, Asp151, Asn153 and His248, whereas the side chains of Asn114 and Tyr68 fix the adjacent phosphate moiety. Thus, the terminal phosphate group is coordinated similarly to a free PO_4^{3-} anion such as described e.g. for the structures of Mth0212 alone (3G91) and in complex with single- (3G2C) or double-stranded DNA (Fig. 5a).

In the structure 3G0R, a phosphate anion is bound in the active sites of both protein molecules in the asymmetric unit with an occupancy of 70 % (Fig. 5b). As in several other structures such as 3GA6, Arg209 stacks with the final base pair of the dsDNA. However, the conformation of Arg209 in the structures 3G00 and 3GA6 differs. In 3G00, its side chain protrudes about 2.8 Å further into the direction of the DNA helical axis and enters the helix in an angle of about 60° to the hydrogen bonds of the terminal Watson-Crick base pair which is bound by Arg96. In contrast, in 3GA6, the entering basic side chain is oriented approximately perpendicular to the hydrogen bonds of the terminal base pair of the DNA molecule bound at the other domain of Mth0212, which contacts Arg215 and Arg163 and is only visible in 3GA6.
Like in both molecules of the complex structure 3G0R, the peptide bond between Arg209 and Thr210 is flipped in one molecule (A) of the complex 3G00. The observed flipping might be explained by the extreme distortion of the respective terminal base pairs nearby.

Using the mutant protein Mth0212(D151N), in terms of the dU endonuclease activity only non-productive binding was achieved. This result might be explained with the assumption, that the magnesium ion coordinated in wild-type enzyme is not only catalytically essential, but also required for binding of dU in the active site, which anyhow seems to exhibit a quite weak affinity. In order to rule out this possibility, Mth0212(WT) was used for co-crystallization with DNA. According to a decreased nucleolytic activity on phosphorothioate linkages [9], this modification was incorporated on the 5`-side of the target dU to prevent cleavage by Mth0212. Since the crystallization experiments were performed at 20°C, a temperature far below the optimum of the hyperthermophilic enzyme, the reaction was assumed to be completely abolished and the S-oxidation will be referred to as non-hydrolysable in the following.

The structure 3G3C represents the wild-type protein in complex with a 6 bp dsDNA containing a single one nucleotide long 3`-overhang. The active site of both protein molecules in the asymmetric unit harbours the 3`-OH group of the overhanging nucleotide as well as a phosphate and a magnesium ion. Instead of a stacking interaction, Arg209 forms hydrogen bonds with this unpaired nucleotide, namely with the N1 atom of the base (adenine), and with the base and terminal phosphate group of the first nucleotide in the other DNA strand.
The structure 3G4T looks quite similar despite of the presence of seven general base pairs resulting in two blunt ends.

In contrast to the other Mth0212-DNA complex structures described so far, the structures 3G3Y and 3G2C show Mth0212 in complex with single-stranded DNA bound adjacent to the active site. Both structures were yielded with wild-type protein.
In 3G3Y, the 4 nt long strand exhibits a 5`-phosphate end and a hydroxyl group at the 3`-end. These observations lead to the assumption that the bound ssDNA is the product of exonucleolytic degradation and that the yielded 4 bp DNA does not form a stable duplex anymore resulting in the dissociation of the two strands.
In principle, the DNA conformation in 3G2C looks similar to that in 3G3Y with one central nucleotide flipped out of the base stack and most likely represents a product as well. However, the protein – DNA and crystal contact interactions differ, and in 3G2C a phosphate anion was observed in the active site.

Comparison of the structure of Mth0212 alone with the protein-DNA complex structures

The r.m.s. deviations between the Cα atoms of Mth0212 in the free and the DNA complex structures amount to only about 0.6 Å for 250 aligned Cα atoms and thus lie in the same range as the deviations of the different apo and complex structures, respectively, among each other. Thus, as expected the conformation seems to be essentially unchanged upon DNA binding. Accordingly, a superposition of the structures of free Mth0212 with the complex structures (Fig. 5c, d) shows deviations mainly in the DNA binding loops, which has been similarly observed for human Ape1 [35; 40; 41]. This also applies to the superposition of the different Mth0212-DNA complexes which fit very well except for the loop Ser207-Arg211. When involved completely, the DNA binding interface covers an area of approximately 1300 $Å^2$ (3GA6).

Structural comparison of Mth0212 and its homologs

Based on sequence analysis and biochemical studies, Mth0212 was assigned to the ExoIII family of nucleases [9; 18; 20]. Several members of this family have been studied by means of X-ray crystallography (Tab. 1). The three-dimensional Mth0212 structure exhibits a strong overall similarity to all of these homologs as shown by the superposition in Fig. 6. These observations are supplemented by results from pairwise structural alignment of Mth0212 and its homologs with DALILite [42] (Tab. S1). Up to 251 common C_α atoms were aligned with r.m.s. deviations of about 1.3 Å, and the Z-score amounted to a maximum of 39.4 (threshold = 2.0) for Ape1. The similarity includes the five highly conserved sequence motifs of the AP endonuclease family [31] as well as additional active site residues.

The arrangement of the residues implemented in the catalytic mechanism, but not involved in DNA binding in Ape1 (Glu96, Asp283, His309) [40] and ExoIII (Glu34, Asp229, His259) [32] which both lack a uridine endonuclease activity fits well with that of the equivalent residues Glu38, Asp151, Asp222 and His248 of Mth0212. The two acidic residues and the histidine form a classical catalytic triade located at the bottom of the substrate binding pocket. Asp222 is buried deep in the pocket, at the height of the intercalating Arg209. Like the corresponding residues Glu96 of Ape1 and Glu34 of Exo III, Glu38 of Mth0212 participates in an extended hydrogen bonding network at the active site and is involved in the coordination of one Mg^{2+} ion. The similarity extends to residues involved in hydrogen bonding networks with these residues such as the conservation of a hydroxyl group containing side chain corresponding to Thr204 of Mth0212 [31] which stabilizes the Lewis acid Asp222 for abstracting a proton from His248 during the catalytic mechanism. And the same applies to a further catalytically essential amino acid, Asp151 of Mth0212.

As outlined above, the structures described here represent the nuclease in its eponymous exonucleolytic function. However, in the structure of the rationally designed, catalytically inactive mutant protein Mth0212(D151N) in complex with two DNA duplices (3GA6) all potential DNA contacting residues are involved in such contacts. Thus, most often this structure will be used for comparison with homologs in the following.

While the overall structure resembles that of its homologs, there are significant differences in the DNA interaction surface, in particular in the three DNA binding specificity loops

(Fig. 7a). Most suitable for the respective comparison are of course the structures of homologs in complex with DNA.

The structurally similar DNaseI has been co-crystallized with DNA, but lack the three specific DNA binding loops which are shared by all ExoIII homologs known so far [40; 43; 44].

Recently, the structure of the homolog Af_ExoIII from the archaeon *Archaeoglobus fulgidus* [18] has been determined in complex with dsDNA showing for the first time an ExoIII homolog bound to the end(s) of a DNA duplex [36]. Most of the amino acids important for substrate binding or catalysis are structurally conserved between Mth0212 and ExoAf as well. However, the Af_ExoIII-DNA complex structure only describes about half of the interaction surface between the enzyme and substrate DNA. Furthermore, it lacks a tightly bound target nucleotide, a phosphate ion or 5`-phosphorylated DNA as products as well as catalytically essential $Mg^{2+}$ ions at the active site. In contrast, the various Mth0212-DNA complex structures described here provide detailed insight into the eponymous exonucleolytic activity based on the characteristic binding mode of the target 3`-OH end of dsDNA and Mth0212-product complexes. Additionally, the exact localization of the metal cation(s) coordinated at the active site and required for all catalytic activities are comprised in several Mth0212 structures. The exonucleolytic activity is very important to guarantee that the DNA repair can be completed by a DNA polymerase upon removal of the damaged base (AP site / dU).

However, due to the larger number of structural and biochemical studies on Ape1, mainly this human homolog will be used for comparison with Mth0212. Several structures of Ape1 in complex with AP site containing DNA have been solved all showing the enzyme bound to the central DNA region [35; 41] (and 2ISI: Agarwal et al., to be published, deposited 2006).

The secondary structure elements are well conserved between Ape1 and Mth0212. One exception concerns loop II which is completely α-helical in Ape1, whereas it lacks almost any secondary structure in Mth0212. As mentioned above, the major differences concern the three DNA binding loops. These loops comprise the residues Asn114-Arg121 (loop I), Arg163-Gly172 (loop II) and Trp205-Trp219 (loop III) of Mth0212 and Gly176-Arg181, Asn222-Lys227 and Trp267-Lys277 of Ape1, respectively. The deviations in the protein-DNA binding interactions most likely are the predominant reason for distinct substrate specificities.

The loops vary significantly in their charge as already obvious on the amino acid sequence level [18]. In relation to the DNA helical axis, loop I and II are located on the same level as the active site. They exhibit completely distinct conformations in both nucleases. Loop I of Ape1 takes a course into the 3`-direction of the AP site, in Mth0212 the turn is oriented into the opposite direction resulting in a difference between the backbone atoms of up to 5 Å. By the distinct course of loop I and thus additional contacts to the DNA backbone in Mth0212 compared to Ape1, the binding to the DNA major groove is enhanced.

Remarkably significant differences also concern the two DNA contacting residues which have been shown to insert from the minor and major groove, respectively, Met270 and Arg177 of Ape1 [35] capping the active site when DNA is bound. They stabilize the extra-helical conformation of the abasic site and lock the enzyme onto the target DNA. However, none of them is essential for AP site flipping and activity. Thus, neither Met270 nor Arg177 represent a key element for recognition, although both directly contact the phosphate group that is located immediately 3` of the AP site.

Based on three-dimensional comparison, in Mth0212, Met270 and Arg177 correspond to Arg209 and Met117 (Fig. 8). In the complex 3GA6, the side chain of Met117 interacts with the phosphate group adjacent to the scissile target phosphodiester bond and thus might be positioned suitably to recognize conformational changes of the DNA backbone which commonly occur in AP site and dU containing nucleic acids. On the other hand, the side chains of Met117 and Arg177 point into opposite directions which causes a distance of 12 Å between the two long flexible side chains and interactions with phosphate groups of the DNA backbone separated by two nucleotides (the AP lesion and an undamaged nucleotide). In contrast, the side chain of Lys116 is positioned more similarly to that of Arg177.

Most likely, Met117 and Arg209 can take over the function of the equivalent residues of Ape1, since the appropriate prerequisite for Ape1, an extreme kink of the bound DNA which braces the backbone for double-loop insertion occurs in the Mth0212-DNA complexes as well. In Ape1, the kink is achieved predominantly by contacts of the three loops to the DNA backbone phosphate groups 3` to the AP site involving Asn222, Asn226 and Trp280 at the second adjacent 3`-phosphate and Lys276 at the next. While Asn226 and Trp280 are conserved in Mth0212 as Asn167 and Trp219, the further interactions differ since Asn222 and Lys276 are replaced by Arg163 and Arg215, respectively. The interaction with Arg215 corresponds to that of the equivalent lysine, but in contrast to Asn222, Arg163 together with Arg215 forms contacts to the third phosphate group 3` of the abasic site resulting in a stronger overall and more equally distributed binding. Consequently, the kink of the DNA might be enhanced allowing a dU to flip out as well.

The minor groove is spanned and widened about 2 Å by Gly127 and Tyr128 of Ape1 which are both structurally conserved in Mth0212. The same applies to Arg15, Ala16 and Lys20 of Mth0212 corresponding to Arg73, Ala74 and Lys78 of Ape1 which make interactions with three consecutive DNA phosphate moieties of the undamaged DNA strand on the 5`-side of the AP site.

In Ape1-DNA complex structures, the DNA exhibits an extremely distorted B-form conformation [35]. Several unusual and partially variable stacking interactions were observed. In the two Ape1-DNA complex structures which represent the status before cleavage, classical base stacking interactions are still present on both sides of the AP site (1DEW) or at least on its 5`-side (1DE8), and the base opposing the AP site stacks with the base on its 5`-side. In contrast, in the structure of Ape1 in complex with cleaved AP-DNA (1DE9), two nucleotides of the complementary strand occupy the space opposite of the damage site neither of them involved in conservative base stacking interactions. The interactions between the nucleotides include base stacking between bases arising from complementary strands of the DNA as well as hydrogen bonds between nucleotides located on different levels of the kinked helix. Rarely any Watson-Crick base pair form the classical hydrogen bonding pattern, and even purine-purine and pyrimidine-pyrimidine mismatches are observed. Within this arrangement of tilted base pairs in the Ape1-DNA product complex, the side chain of Met270 packs against the base opposite of the AP site to substitute for the lacking base.

In molecule A of the Mth0212-DNA complex 3GA6, the equivalent Arg209 has a similar role. It packs against the base of the 3`-terminal nucleotide that is bound in the active site with its phosphate moiety serving as the target of exonucleolytic cleavage which is represented in this complex instead of the incision at an AP site in the Ape1-DNA complex structures.

Additionally, Arg209 stacks with bases located on both, the 3`- and the 5`-side of the target nucleotide but once with the nucleotide immediately adjacent to the excised nucleotide and in the other case belonging to the complementary DNA strand of the AP site containing oligonucleotide. In most other Mth0212-DNA complexes, Arg209 is singularly involved in stacking interactions with the terminal base pair and positioned on the helical axis between the pairing bases. Although the arginine loop insertion first was supposed to be a unique feature of Ape1 [35], the observations of this remarkable feature in the Mth0212-DNA complex structures in total suggest that Arg209 of Mth0212 takes over a similar function. The same might apply to Arg204 of the recently structurally characterized Af_ExoIII, which forms stacking interactions with the terminal base pair of DNA duplex as well [36] but has not been analyzed in detail so far. The arginine is conserved in prokaryotes and plants, while a methionine is located at the equivalent position in almost all eukaryotes. The adjacent threonine 210 is strictly conserved in all three domains of life.

The hydrophobic pocket close to the catalytic residues is completely conserved between Ape1 and Mth0212 except for the substitution of Ape1 Phe266 by Trp205. Since the tight packing of the hydrophobic Ape1 pocket was already predicted to prevent the binding of undamaged nucleotides, the extended dU specificity of Mth0212 despite of the larger tryptophane side chain and thus an even tighter hydrophobic pocket is all the more fascinating. Furthermore, an explanation that the expanded substrate spectrum is solely based on the single varied amino acid is in conflict with the conservation of Trp205 in ExoIII from *E. coli.* And mutational studies with Ape1 rule out the possibility that substrate specificity is gained by the insertion of a hydrophobic side chain into the abasic hole since mutation of Phe266 resulted only in a six-fold reduced binding affinity but removal of AP sites with almost the same efficiency as wild-type protein [45]. However, generalizations concerning the assignment of functional roles to residues in the hydrophobic have to be done with care, because controversial results were obtained for ExoIII using the mutant Trp212Ser in which both binding to AP sites and activity are abolished completely [33] most likely due to the destruction of the structural integrity of the adjacent DNA recognition loop or the narrow hydrophobic binding pocket.

The narrow pocket of Ape1 was shown to allow only binding of AP sites in α-configuration which are specifically produced by DNA glycosylases. Whether Mth0212 exhibits the same configuration specificity or recognizes the β-anomer which is formed by racemisation in solution as well has not been investigated so far. However, the pocket bordering residues Phe266, Trp280 and Leu282 which pack against the hydrophobic side of the abasic deoxyribose moiety in Ape1-DNA complexes are conserved in Mth0212. Therefore, most likely the equivalent residues of Mth0212, namely Trp205, Trp219 and Leu221, are involved in similar interactions helping in the recognition of AP sites.

Tyr171 of Ape1 has been identified as a key element for the efficient discrimination between undamaged and AP-site containing DNA [46]. The side chain of the corresponding amino acid of Mth0212, Tyr111, directly contacts the scissile phosphate moiety of the DNA backbone (molecule A in 3GA6) and therefore most likely has a similar function. Additionally, it forms a direct hydrogen bond to the carboxyl group of Glu38 such as e.g. in the apo structure 3G91 or a water-mediated hydrogen bond to the DNA backbone phosphate of the nucleotide adjacent on the 5` to the excised residue (molecule A in 3GA6). The mediating water molecule is also bound by the backbone carbonyl of Asn114 and the side chain of Arg96. This asparagine in turn forms contacts to the backbone amine of Lys116 as well as with the free 3`- hydroxyl group of the removed nucleotide.

In contrast to mutation of Tyr111, substitution of Asp210 with alanine, asparagine or glutamate does not influence the recognition of AP sites [47] but stabilizes substrate binding. The structurally equivalent amino acid Asp151 forms a hydrogen bond to the phosphate group coordinated in the active site as well (confer 3G3C and 3G4T with bound phosphate anions) (Fig. 7) and in concert with its strict conservation in all ExoIII homologs can be assumed to have similar effects. Together with Asn10, Glu38, Tyr111, Asn114, Asn153, His248 and several water molecules, in the apo structure 3G91, Asn151 is coordinated by a phosphate anion which is bound in the active site. Most likely because of these interactions with a phosphate group (natively and in some Mth0212-DNA complexes belonging to the DNA backbone) even the conservative mutation of Asp151 to asparagine results in further reduction of the anyhow low *in vitro* activity of Mth0212 on dU residues [9].

Tyr128 of Ape1 has been shown to play a critical role in binding of the nucleotide on the 5`-side of an abasic lesion [48]. The equivalent residue Tyr68 of Mth0212 forms a hydrogen bond to the phosphate group adjacent to the scissile phosphodiester bond in the 5`-direction as well. The similarity extends to the positions of Arg156 and a bound water molecule participating in the same hydrogen bonding network. Therefore, Tyr68 of Mth0212 might fulfil the same function as Tyr128 of Ape1. Additionally, it was suggested to be a key element of the strong product binding required for processive scanning of DNA for abasic sites by Mth0212 since in the homolog Af_Exo which acts distributively the aromatic residue is replaced by Arg63 that is not involved in DNA binding at all. The suggested strong product binding by Mth0212 would be in good agreement with the observation that most of the Mth0212-DNA complex structures described here indeed contain products tightly coordinated in extended hydrogen bonding networks. The critical role and strong DNA binding of Tyr68 might cause stress in the protein backbone and thus explain the observation that the neighboured residue Ser69 only rarely is located in the expected Ramachandran regions. Ser69 is conserved in Ape1 as well (Ser129).

Magnesium ions have been shown to be catalytically important for all ExoIII homologs characterized so far including Mth0212 [9]. In order to verify the divalent metal ion coordination sites, Mth0212 was crystallized in presence of Mg^{2+} ions and crystals were soaked with solutions containing Mn^{2+}, respectively. Commonly, a single magnesium ion is bound in each active site. The presence of a second divalent metal ion at the active site became only visible in the structure 3G0A yielded from WT crystal soaked with manganese chloride (Fig. 2).

The difference between manganese and magnesium concerning the number of metal ions bound in the active site is in agreement with results from isothermal titration calorimetry (ITC) experiments, which show that Exo III from *E. coli* binds two Mn^{2+} ions but only one Mg^{2+} [49]. In analogy to ExoIII from *E. coli* [32], the magnesium ion most likely assists to orient the phosphate group in the active site and to polarize the scissile P-O3` bond as well as to stabilize the transition state.

The Mn^{2+} ions are coordinated by residues equivalent to the metal cation binding sites of human Ape1. However, based on a significantly weaker anomalous signal of the manganese ion at this second site than at the commonly used metal binding site A, it seems to be by far occupied to a smaller extent probably reflecting a lower binding affinity of site B. An alternative explanation for the different occupancies of the two metal ion binding sites would be that a magnesium ion deriving from expression is bound at site B so tightly that it could not be replaced completely by a manganese ion during the soaking procedure. But the observation of at maximum one magnesium ion in all other structures, namely singularly at site A, argues against this hypothesis.

Commonly, the magnesium ion is bound by the side chains of Glu38 and Asn12. The remaining coordination sites are occupied either by water molecules or by atoms of bound DNA. Since a divalent metal ion was not bound at site B in any of the other structures described here - neither in the absence of DNA or a phosphate anion in the active site nor in particular together with them - the partially occupied second metal ion at binding site B in 3G0A which involves Asp151 most likely does not have any physiological impact for metal ion coordination.

The binding of a metal ion at site A seems to be independent of the presence of a phosphate group or DNA strand since it was also identified in the structures 3FZI, 3G1K, (3G0A) and 3G3Y, which do not harbour any other ligand in the active site.

Vice versa, a phosphate group can also bind in the active site in the absence of a divalent metal ion such as the DNA backbone phosphate moieties in the complex structures 3G38, 3G0R and 3GA6 or the phosphate anion in 3G00. However, concerning 3G38, it is a 3`-terminal phosphate moiety of the DNA and with regard to 3G00, the phosphate anion is located at a position differing from the coordination site in all other structures by about 3.3 Å. It is bound in a bi-dentate manner by the side chains of Glu38 and Asp247 as well as the Nε atom of His248. In addition to two tightly bound water molecules it binds the terminal 3`- hydroxyl group of one DNA strand (Fig. 5a).

In all other Mth0212 apo or –DNA complex structures harbouring a phosphate group in the active site - either an anion (apo: 3G91, complex: 3G2C, 3G3C, 3C4T) or a phosphate moiety of the DNA backbone (3G38, 3G0R, 3GA6) – it is located at the coordination site B (or in a sphere of about 2 Å around it) (Fig. 5a).

Since 3G38 and molecule B of 3GA6 each contain a 3`-terminal phosphate group at site A the last but one phosphate moiety of the DNA backbone which actually represents the scissile bond of the substrate concerning the exonucleolytic activity is bound only in the active site of molecule A of the complex structure 3GA6 and in both molecules of 3G0R. These complexes therefore seem to unambiguously represent Mth0212-substrate complexes. Additionally, the structure 3G00 seems to represent an Mth0212-substrate complex since the co-crystallized DNA still includes all nucleotides. However, the DNA appears not to be bound properly for cleavage due to the coordination of a phosphate anion from the buffer solution by the catalytically essential residue Glu38.

All structures of Mth0212(D151N) in complex with a substrate DNA lack a magnesium ion in the active site. This observation and the catalytic need of magnesium ions for ExoIII homologs in concert with studies showing that the D151N mutation still enables binding to DNA, but causes a loss of all nucleolytic activities, led to the assumption that the presence of both a magnesium ion at coordination site A and Asp151 at site B are prerequisites for the productive binding of the phosphate moiety of the scissile bond in the DNA backbone at site B. Most likely they act in a cooperative manner to orient the scissile bond properly for cleavage.

The metal coordination at the active site of Mth0212 also resembles that in the N-terminal fragment of T4 DNA polymerase, which has been co-crystallized with a three nucleotides long ssDNA [50]. In both enzymes one ligand for the divalent metal ion at coordination site A is provided by a non-bridging oxygen of the scissile phosphate of the DNA backbone, while the other ligands derive from protein carboxylate groups. However, the number of the involved side chains differs. In Mth0212, only two amino acids provide hydrogen bond donors, namely Glu38 and Asp247 with one and two carboxyl groups, respectively, whereas T4 DNA polymerase binds the metal ion via three acidic residues (Asp112, Glu114, Asp324).

The complex 3G38 and the protein molecule B of the structure 3GA6 cannot be unambiguously classified as product or substrate complexes as follows. A phosphorylated 3`-terminus of DNA does not serve as a substrate of the 3`-5` exonuclease activity of Mth0212 and is neither the result of the exonucleolytic function nor of the AP and dU endonucleolytic activities. They might point out to a further activity of Mth0212, namely the removal of phosphate moieties at 3`-ends during DNA repair pathways, which has not been identified so far. This hypothesis would be also in accordance with the observation of a phosphate anion at the active site of several Mth0212 apo (3G91) and Mth0212-DNA complex structures (3G2C, 3G3C, 3G4T, 3G00). The first exonuclease III homolog discovered to accomplish such a 3`-phosphatase activity is ExoIII from *E. coli* [51].

Comparison of the Mth0212-DNA complexes with other exonucleases in complex with DNA

Structural differences between certain types of exonucleases display their variance in substrate specificity, processivity and the reaction mechanism. Since the DNA is bound by Mth0212 in an exonucleolytic manner in all complex structures the DNA binding mode was also compared with that of other 3`-5` exonucleases. The molecular mechanism of the removal of nucleoside monophosphates (dNMPs) from the DNA-3`-end was predominantly revealed by structural and mutagenic studies on the exonuclease domains of the Klenow fragment of DNA polymerase I, bacteriophage T4 DNA polymerse [52] as well as on φ29 polymerase [53] (reviewed in [54]). The Klenow fragment of DNA polymerase I from *E. coli* binds duplex DNA in a groove adjacent to the 3` to 5` exonuclease domain but in the 3` to 5` exonuclease active site interacts with single-stranded DNA, namely a 3`-extension of three nucleotides [55].

Comparison of the active site environment of Mth0212 with the amino acid arrangement in uracil DNA glycosylases

The general statement that any uracil base in DNA is removed by UDG [3] has to be partly corrected. At least in *M. thermoautotrophicus* [9; 20] and some larval stages of *Drosophila melanogaster* [10], alternative strategies have evolved.

In the terminal RNA uridylyltransferase (TUTase) TbTUT4 from *Trypanosoma brucei*, a key residue for the specific recognition of a (terminal) uridylyl residue is Arg121 which due to its hydrogen bonding interaction with the O4 atom of the pyrimidine ring can discriminate against cytosine [56]. In Mth0212, a similar role could be taken over by the side chain amino group of Lys125 since it forms a hydrogen bond with a glycerol molecule in the active site of molecule B of the complex structure 3GA6, which is bound at a similar position in the hydrophobic pocket near the active site as the abasic site in Ape-DNA complex structures [35] (Fig. 9). The glycerol molecule is located a little bit deeper in the AP site binding pocket than the deoxyribose moiety and thus could superpose very well with a uracil base linked to the sugar. Indeed, in a superposition of the respective active site residues Asp247/His248 of Mth0212 and Asp308/His309 of Ape1, the distance between the hydroxyl group hydrogen-bonded to Lys125 and the C1` atom of the deoxyribose ring amounts to 4.4 Å which is only slightly below the distance of 5.5 Å between the anomeric carbon and the carbonyl group at position 4 of the pyrimidine ring which is characteristic of uracil. Therefore, it can be assumed that a similar hydrogen bond network might also serve for the specific recognition of a dU residue when bound instead of the AP site. The involvement of hydrogen bonds in the specific recognition mechanism is further supported by the fact that also pseudouridine containing DNA is incised, most likely since this unusual nucleotide exhibits the same pattern of hydrogen bond donors and acceptors at the Watson-Crick base pairing edge [9]. However, in addition to the interaction with Lys125, the hydroxyl group of the glycerol molecule in 3GA6 also contacts the O atom of Ser171 and the carbonyl group of Asn153. While the first can interact with both an amino and a carbonyl group at cytidine and uracil bases, respectively, the latter cannot interact with uracil, but with cytidine and thus would compensate for the hydrogen bond lacking between Lys125 and cytidine.

Asn153 is conserved and fits very well in the superposition of ExoIII homologs. It might play a role in the correct positioning of the scissile bond with the help of the coordinated metal ion (Fig. 9) as well as in the discrimination of dU against cytidine and thymine bases as discussed above with regard to the glycerol molecule bound in the active site of the complex 3GA6 (molecule B). By means of the extensive hydrogen bonding network, Asn153 together with Asn10 additionally might help in adjusting the pK_a of Asp151 which most likely serves as the putative Lewis acid protonating the O3` leaving group as has been suggested for ExoIII from *E. coli* [57].

A position of the uracil base resembling the glycerol conformation could enable stacking interactions with either Trp205 or Trp219, which are arranged perpendicular to each other. The equivalent residues of Ape1, Phe266 and Trp280 have been analysed in detail concerning their stacking properties with abasic sites. Due to an orientation perpendicular to the side chain of Lys125, Trp219 seems to be more feasible for dU stacking, whereas the larger side chain of Trp205 compared with Phe266 of Ape1 might help to better shield the hydrophobic pocket from solvent as observed for a water molecule which might enter after dissociating from the DNA backbone phosphate 3` of the target nucleotide (3GA6).

On the sequence and structural level, Mth0212 is completely unrelated to both structurally defined superfamilies of UDGs known so far. While in human UDG large structural rearrangements occur upon binding to uracil in DNA [58], the structure of Mth0212 remains mainly unchanged upon DNA binding. This contrast leads to the assumption, that some smaller conformational changes might also occur in Mth0212 when actually uridine is bound instead of a 3`-OH or 5`-phosphate end of the DNA.

Both enzymes might use similar mechanisms for DNA backbone distortion as follows. The specific recognition of a dU within DNA by UDGs begins with a compression of the phosphodiester backbone due to pinching interactions of three conserved serine residues which form hydrogen bonds with three adjacent phosphodiester groups of one DNA strand. The resulting stress on the torsion angles of the DNA helix subsequently is relieved by flipping out the dU from the base stack into the active center [59]. A similar mechanism of DNA backbone compression might apply to Mth0212 implying the large number of arginine residues (instead of serine residues) for concerted contacts to the phosphate moieties of the DNA backbone (Fig. 4d). A resulting deformation of the helical torsion angles is in accordance with significant deviations from ideal DNA conformation in almost all Mth0212-DNA complex structures although crystal packing effects cannot be ruled out completely.

Discussion

The presence of 2`-deoxyuridine is pre-mutagenic. In the archaeon *M. thermoautotrophicus,* its repair is initiated by the enzyme Mth0212 compensating for the absence of an otherwise almost ubiquitous uracil DNA glycosylase (UDG). Mth0212 belongs to the ExoIII family of nucleases and - acting as a 3`-5` exonuclease and an AP and dU endonuclease - exhibits a unique combination of DNA repair activities. The structure of Mth0212 has been solved both alone and in complex with different substrate DNAs at a resolution ranging from 1.2 Å to 3.1 Å.

The *in vitro* nucleolytic activity against dU is very sensitive and requires buffer optimization to become obvious (compare [18] and [9]). However, even under co-crystallization conditions optimized in terms of buffer system, pH and salt concentration it was not possible to obtain the structure of an endo-complex, in which Mth0212 binds a target dU at the active site. Thus, another approach to circumvent the gradual affinity of Mth0212 in favour of 3`-OH ends in order to catch the Mth0212-dU endo-complex was followed. Both 3`-ends of a dsDNA substrate were blocked by a fluorescent dye. Surprisingly, the yielded complex structures represent the exonucleolytic function as well. This failure can only be explained by a strong exonucleolytic affinity and activity of Mth0212, which even allows the incision of the DNA backbone adjacent to bulky groups such as the dye. These observations remind at the results of activity assays of human Ape1 on 3,N4-Benzetheno-dC [60]. Thus, expanding the analogy, Mth0212 probably also recognizes local structural changes of the DNA B-form induced by either the terminal nucleotide, an AP site or a 2`-deoxyuridine rather than the end of the DNA or the damaged base itself. Mth0212 seems to probe for the overall DNA conformation around the target nucleotide / AP site. It is supported by further studies of the human homolog which have shown that the AP endonuclease activity is enhanced by a 3`-mismatch [61].

A similar dependence of the activity of Mth0212 would be in good agreement with biochemical and structural data. The efficiency of dU removal is increased with the ratio of A-T to G-C base pairs of the sequence context probably due to the larger local flexibility of AT-rich contexts. The subsequent exonucleolytic activity on the incised DNA strand depends on the sequence as well with an increased rate for pyrimidine bases opposite of dU (unpublished data, laboratory of Prof. Fritz).

This effect might be explained in part as follows. Although DNA helices containing an AP site assume more or less canonical B-form as evidenced by 2-D NMR experiments [62; 63; 64; 65], in the same and in additional studies [66] sequence-specific conformational effects were revealed as well. While purine bases opposite an AP site still stack with the adjacent bases, pyrimidines at this position are located intra- as well as extra-helical depending among others on the sequence context. Also the increased overall flexibility [67] and deviations from ideal B-form geometry which extend up to four nucleotides away from the abasic site [68] depend on the nature of the bases around the AP site [69] (reviewed in [34]).

The striking number of highly positively charged arginine residues particularly in loop III of Mth0212 probably reflects the requirement of a strong DNA binding affinity *in vivo* due to the high intracellular salt concentrations of *M. thermoautotrophicus* such as 500 mM K^+. On the other hand, the enzyme must not bind too tightly to the DNA in order to guarantee a processive search mechanism for damaged bases. Therefore, the presence of regulatory factors *in vivo* seems reasonable, and such elements might also enhance the dU endonuclease activity which *in vitro* is already abolished at salt concentrations of 200 mM KCl. Additionally, nucleotide modifications might regularize the different nuclease activities of Mth0212 *in vivo* in analogy to an enhanced binding affinity of Ape1 in the presence of a methylated adenine two base pairs 3` of an AP site [30], which is probably due to an enhanced water-mediated hydrogen bonding to Asn229. This asparagine is replaced by a valine in Mth0212 excluding an equivalent interaction, but similar contacts might e.g. be formed between the base three nucleotides away from the abasic lesion in the 3`-direction and Glu166 or between the second nucleotide 5` of the AP site and Lys40.

The 2:1 stoichiometry between Mth0212 and DNA in all complex structures except for 3GA6 has been also observed by Kuettner and coworkers [70] and can probably be explained by an exonucleolytic activity of two protein molecules at both ends of a dsDNA helix which is stopped by steric clashes between the two bound protein molecules at a specific length of the DNA strands. Most likely, the binding of an Mth0212 molecule to substrate DNA is enhanced when another protein molecule is already bound to the same substrate molecule due to the increased distortion of the DNA backbone geometry upon binding.

The structures obtained with wild-type protein as well as the Mth0212(D151N) complex structure 3G2D most likely represent product complexes since a free 3`-OH group is bound in their active sites.

In presence of Mg^{2+}, Asp308 of Ape1 helps to fix the enzyme onto the DNA after cleavage [71]. This role most likely is taken over by the structurally equivalent residue Asp247 of Mth0212, which forms a direct and a water-mediated hydrogen bond to the side chain of the catalytically important residue Trp205 and to the terminal 3`-hydroxyl group in molecule B of the complex structure 3GA6, respectively, that therefore might represent the enzyme-DNA product complex after incision. The tight product binding corresponds to an inhibition of the enzymatic activity by its product as reported for Ape1 [71] and probably helps to streamline the DNA repair pathway by coordinating the various enzymes involved. Asp70 supports DNA complex stability and / or the conformational changes of the DNA backbone as well [29]. In Mth0212, it is replaced by Asn12 which is involved in metal ion coordination and thus indeed most likely leads to an improved DNA binding. Additionally, Asn12 is located in hydrogen bonding distance to the Nε atom Lys40 and to several water molecules which participate in an extended hydrogen bonding network including Glu38 and the phosphate group bound at the active site (e.g. in 3G91).

The efficiency of the insertion of a so-called “arginine finger” into the DNA base stack (Arg209 of Mth0212) resulting in an enhanced DNA binding affinity has additionally been reported from other DNA repair enzymes such as for the mutated Leu276Arg of human UDG [72; 73]. Here, the penetration stabilizes the DNA conformation by occupying the space of the flipped-out ψ residue after insertion of the adjacent loop into the DNA from the minor groove. Further stabilization of the flipped-out residue in its extra-helical conformation is achieved by charged interactions with the neighboured phosphates and between a carbonyl group of the protein backbone and the opposing guanine base as well as by a hydrogen bond between the Nε atom of Arg272 and a base 5` of the flipped-out dU.

In analogy, a movement of the Arg209 containing loop of Mth0212 could be specific for productive dU binding. Subsequent recognition of the uracil base by human UDG is mainly achieved by hydrogen bonds with the side chains of Asn204, His268 and backbone atoms as well as by a stacking interaction with Phe158 [72]. A critical interaction responsible for the exclusion of thymine is the packing of Tyr147 against the C5 atom of the pyrimidine ring. Nucleotide flipping is a common mechanism which is used not only by damage-specific DNA repair enzymes but also by sequence-specific DNA methyltransferases.

According to the structure of human Ape1 in complex with DNA, AP sites are recognized due to the extra-helical conformation of their deoxyribose moiety [35; 40] involving a contact between Arg177 and the phosphate on the 3`-side of the AP site. Met117 of Mth0212 is located in a similar position and its side chain might be suitable for an insertion into the DNA base stack substituting the function of Arg177 to help in the recognition of an AP site or additionally the damaged base 2`-deoxyuridine. The whole DNA binding loop I takes significantly distinct courses probably explaining the differences in the biochemical properties. While it goes into the direction of the 3`-end of the AP containing DNA strand in Ape1, the corresponding loop of Mth0212 rather remains on the same level in relation to the DNA helical turns and is positioned further away from the DNA. The conformation of the Met117 side chain of Mth0212 is very flexible reflected in B factors above average and alternative conformations in some structures. In the Mth0212-DNA complex structure 3G00 obtained at low salt concentration, Met117 is located in close proximity to the 3`-terminal nucleotide. Most likely the recognition mechanisms of the 3`-end and a target dU are mediated by the same residues both involving specific interactions with characteristic features of the DNA backbone which significantly deviates from standard conformation. A crucial recognition element for both exo- and endonucleolytic incision probably is a destabilized region which has been shown to exhibit significant similarity at an AP site and at the terminal base pair in dsDNA [34]. Most likely, the same applies to mismatches containing 2`-deoxyuridine and all scissile target bonds are recognized by akin mechanisms. In particular the side chain of Met117 seems to play a key role in the dU recognition. A conservative mutational substitution by an arginine does not significantly affect either the dU endonuclease or another nucleolytic activity. In contrast, the absence of a space-filling side chain at the equivalent position in Ape1 (Gly178) resulting in a distance of about 6 Å more between Mth0212 and bound DNA probably prevents the removal of dU by this homolog. The side chain of Met117 lies in a straight line with Trp205 which is located directly adjacent to the third DNA binding loop and suitable for stacking interactions. The larger side chain compared to Phe266 of Ape1 can form stronger interactions with putative substrates and thus might additionally support the ability of Mth0212 to excise dU. However, this tryptophane residue is conserved in other homologs, e.g. from archaea and *E. coli*, contradicting a dU recognition mechanism solely based on Trp205.

Mutational studies concerning the residues Trp205 and Arg209 as well as some neighboured amino acids are currently carried out.

For Ape1, Gorman and co-workers suggested that the flipped-out base opposite the AP site is not required for recognition of the abasic site [35; 40]. Similarly, the activity of ExoIII from

E. coli on AP sites is scarcely influenced by the opposite base [74]. Although the dU endonuclease activity of Mth0212 has been proven for DNA containing each of the four standard DNA nucleotides opposite of dU, the efficiency of dU removal seemed to be influenced by the nature of the pairing base (unpublished results, laboratory of Prof. Fritz). When the crystallization experiments were performed, it had not been examined whether the dependence of the dU endonuclease activity of Mth0212 on the type of the base opposite of dU is caused by distinct binding affinities. Therefore, several different bases opposite of dU were tested after starting with a guanine which represents the *in vivo* situation after cytosine deamination, the most common cause of dU emergence. However, also these variations did not allow catching the dU endo-complex. Obviously, the exo-complex was favoured due to improved crystal contact formation or a greater thermodynamic stability. The latter appears to be the main reason, since the results of electrophoretic mobility shift assays (EMSAs) with Mth0212 and DNA suggest a graduate binding affinity of metal containing Mth0212 increasing from U/G pairs beyond AP sites to DNA ends. At the first glance, these observations seem to be in conflict with an efficient dU repair, but *in vivo* the number of DNA ends is so small that the strong binding affinity of Mth0212 to them does not matter at all.

The phosphate groups of all Mth0212 apo and –DNA complex structures, except for 3G00, and the divalent metal ion at the commonly used coordination site A have positions equivalent to the scissile phosphodiester bond and the manganese ion bound by Ape1 in the DNA complex structure 1DE9 [35]. Therefore, we suggest Mth0212 to act via a reaction mechanism proposed for Ape1 by Mol and co-workers involving only a single metal ion rather than to implement a two-metal-ion mechanism as discussed for Ape1 based on mutational studies equivalent to D151N (D210N) [47] involving kinetics experiments and the use of anomalous scatterers [41]. The comparison suggests that the scissile bond is bound similarly for the different nucleolytic activites, namely the AP endonuclease function represented in the Ape1-DNA complex structures and the 3`-5` exonuclease activity visualized in the Mth0212-DNA complex structures. Most likely, this resemblance applies to the complex required for an endonucleolytic cleavage at dU residues as well.

Concerning the removal of an abasic lesion by the ExoIII family of AP endonucleases, diverse hypotheses about the mechanism of the first step during recognition of the AP site have been advanced [32; 40; 51; 74]; [35; 36; 60]. Also the structure of a member of the second conserved 5` AP endonuclease family which is structurally unrelated to ExoIII and acts via a distinct chemical mechanism, namely that of endonuclease IV from *E. coli* was solved in complex with AP site containing DNA [75]. At least for the representatives Ape1 and EndoIV, respectively, it was shown that enzymes of both families orient the AP-DNA via positively charged complementary surfaces and insert loops into the helical base stack. Although the DNA conformation in each protein-DNA complex differs significantly, in both cases the resulting DNA bent and kink causes the AP site to flip out into an extra-helical conformation within the active site pocket. So far, this pocket was thought to exclude nucleotides. While this hypothesis has remained for the four common bases, it has recently been corrected for the damaged nucleotide 2`-deoxyuridine [9; 20].

Proposed mechanism of 2`-deoxyuridine recognition

A plausible model for the recognition of dU residues by Mth012 is described in the following paragraph. Double-stranded DNA is scanned for an AP site and dU. Regular Watson-Crick base pairs are stable enough to stay intact during the slight nicking of the DNA helix by Mth0212 which can be supposed to resemble the pseudo-continuous DNA helices in the complex structure 3GA6. In contrast, base pairs composed of a native nucleotide and an AP site or dU are less stable, so that the side chain of Arg209 can easily insert into the DNA helical base stack as observed between the two DNA duplices bound in an exonucleolytic manner which come into close contact on the level of Arg209. Thereby, the arginine side chain might either push away both bases or just the target nucleotide, whereas the opposite base remains in the base stack such as at molecule A of 3GA6 with the guanidinium group packed against it to compensate for the lacking base. Like an AP site, the dU flips into the hydrophobic binding pocket. In the extra-helical conformation, it is tightly bound by hydrogen bonds to Lys125, Asn153 and Ser171. The equivalent residue of Lys125 in Ape1, Arg185, could fulfil the same function and Asn153 is even conserved. In contrast, the side chain of Ala230 corresponding to Ser171 cannot participate in hydrogen bonds and thus prevents the productive binding of a (pyrimidine) nucleotide in the hydrophobic pocket. Discrimination of dU against cytidine is performed by Lys125, whereas thymidine is excluded due to steric clashes with the side chains of Asn153 or Asn114.

Materials and Methods

Protein production

In general, the same purification protocol was used for Mth0212(WT), Mth0212(D151N) and Mth0212(K116A). Cloning and protein production of Mth0212 as a full-length, C-terminal LE-His$_6$ tagged variant (pET-21d derivative) with the designed mutation T2A in *Escherichia coli* BL21_UX cells lacking the *ung* gene have been described in [9]. For production of protein used for the determination of the structures 3FZI, 3G0A and 3G91 the same protocol was applied. A final gel filtration step was added to this protocol in order to reduce the salt concentration (180 mM KCl, 10 mM KH_2PO_4/K_2HPO_4 pH 7.0) for the structure of the catalytically active mutant Mth0212(K116A) (PDB-ID 3G91). The expression and purification procedures for the protein batches used for the determination of the other structures described here were modified in order to increase the protein yield and to vary the salt concentration in the protein solutions used for co-crystallization with DNA, respectively. The protein used for determination of the structures 3G1K, 3G8V, 3G2C, 3G3Y, 3G2D, 3G3C, 3G4T, 3G38, 3G0R, 3G00 and 3GA6, respectively, was overexpressed in *E. coli* BL21-CodonPlus(DE3)-RIL cells. Regarding 3G8V and 3GA6, the same medium as described in [9] was used (2YT), while for the other structures the protein was produced using an auto-induction protocol [76]. When the auto-induction medium was used, the protein was expressed at 18°C. Subsequently, the protein was purified as described in [9] with only slight modifications. A heat denaturation step (10 min. 65°C) was followed by a Ni^+-NTA and a Heparin affinity chromatography (GE Healthcare, Germany). Except for the protein batches used for the determination of the structures 3FZI and 3G0A, gel filtration (Superdex 75 26/60, GE Healthcare, Germany) with different buffer systems was added as a final purification step. For clarity, the buffer, in which the protein was dissolved after the last purification step, is listed in table 3. After each purification step, the fractions containing Mth0212 were pooled and concentrated using Vivaspin centrifugal concentrators (Sartorius Stedim Biotech GmbH, Goettingen, Germany). All purification steps were performed at room temperature.

Complex formation

For preparation of double-stranded substrate DNAs, oligonucleotides with the sequences given in table 2 were purchased from Purimex (Grebenstein, Germany) as 2xHPLC purified and lyophilised aliquots synthesized in 1000 nmol scale. The DNA was dissolved in 5 mM

Tris-HCl pH 7.5 in a concentration of 10 mM as determined by extinction measurements at a wavelength of 260 nm. The purity of the DNA was checked by additional extinction measurements at 280 nm and calculation of the OD_{260} / OD_{280} ratio. Equal volumes of the solutions of both DNA oligonucleotides for formation of the respective DNA double-strand were mixed and annealed by application of a slow cooling protocol, in which a heat denaturation step of 10 minutes incubation at 95°C was followed by slow cooling to room temperature yielding a 5 mM dsDNA stock solution. DNA and purified Mth0212 were mixed in a molar ratio varying between 1.1:1.0 and 1.4:1.0 and the complexes were formed during incubation at 20°C for 15 minutes in a solution with varying composition (for details confer Tab. 3). In contrast to the gel filtration buffer used in the last step of protein purification, for most structures the buffer for complex formation additionally contained 2 mM $MgCl_2$.
Subsequently, the solution was centrifuged at 20 000 g, 4°C for 10 minutes to remove possibly precipitated macromolecules.

Crystallization

All crystals were obtained at 293 K using the sitting drop vapour diffusion method. In general, droplets were prepared by mixing 0.7-1.0 µl concentrated protein or complex solution with the same volume of reservoir solution and equilibrated against 300-500 µl of reservoir solution which commonly contained polyethylene glycol or MPD as precipitating agents. Most crystals grew within several days. Single crystals with a size of e.g. about 50 x 50 x 200 μm^3 for the hexagonal crystal of Mth0212(WT) were used for data collection. For the structure 3G0A, Mth0212 (WT) crystals were soaked with a $Mn2^+$ ions containing solution for five minutes.

Data collection

Data were collected at the beamlines X13 of EMBL at the “Deutsche Elektronensynchroton” (DESY), Hamburg, Germany as well as at the beamlines BL14.1 and BL14.2 of the “Berliner Elektronen Synchroton” (BESSY), Berlin, Germany, and on a rotating anode (3G0A), which were equipped with different CCD detector systems and a MAR35 image plate, respectively, as specified in each PDB entry. Commonly images were taken in rotation steps of 0.5-1° at

100 K. Thus, prior to data collection - if required - each single crystal was transferred into a cryo cooling solution, which most often consisted of a 2:1 - 3:1 (v/v) mixture of the crystallization reservoir with glycerol. The composition of the cryo protecting solution used for each crystal yielding a structure is listed in table 3. Most crystals were flash-cooled in liquid nitrogen before mounting them into the cryo stream, while some crystals were mounted directly from the cryo protecting solution. For the structure 3G3C, the crystal was annealed in cryo protecting solution for 20 s before mounting it back into the cryo stream.

Data processing

Data integration and scaling were performed with either the HKL2000 package (HKL Research, Inc., Charlottesville, Virginia, USA) or a combination of XDS [77] and XSCALE [77], XDS and SCALA [78] or IMOSFLM [79] and SCALA depending on the detector used (e.g. XDS for a MAR MOSAIC 225 mm detector) and on crystal parameters. The programs used for each structure are given in the header of the respective PDB entry. If required, data reduction was performed using TRUNCATE [78; 80]. In general, 5 % of the reflections were omitted from the refinement and used for calculation of a free R factor. Space group and cell contents of each structure are summarized in table 4. Table 5 lists the data processing and refinement statistics.

Structure determination

The structure 3 FZI was solved by means of Molecular Replacement (MR) with Molrep [78; 81] using the protein coordinates of the human Ape1-DNA complex structure deposited with the PDB code 1DE9 [35] with all non-conserved residues mutated to alanine as a search model. The signal to noise ratio of the rotation function was $I/\sigma = 9.6$ with the next false peak exhibiting $I/\sigma(I)$ of 4.2. The translation function exhibited a correlation coefficient of Scor = 0.45 and an R factor of 55.3 %. Subsequent rigid body refinement resulted in a CC of 0.61. All other structures described here were solved by MR using MOLREP and the template structure 3FZI.

Refinement

Refinement was performed using Refmac5 [82], PHENIX [37] and SHELXL [83], respectively. For use of the latter, the coordinate file was prepared with SHELXPRO, whereas the reflection file format was converted with XDSCONV before the assignment of free R flags with XPREP (Bruker AXS Inc., Madison, Wisconsin, USA). Except for the structure 3G91 which was anisotropically refined with SHELXL all structures were refined isotropically. A random set of 5 % of reflections was excluded from refinement to monitor R_{free} for all structures except for 3GA6, for which the free R flags were assigned in thin shells using DATAMAN v. 040701/6.3.5 [84]. Cycles of refinement were alternated with manual building in COOT [85]. In general, water molecules were assigned for peaks above 3.0 σ in F_oF_c difference maps and retained when represented in the $2F_oF_c$ map at a level of 1.0 σ as well as located not further than 3.5 Å away from a putative hydrogen bonding partner. The quality of the model was analyzed using PROCHECK [86]. The residues Ser69, Asp103 and Phe173 were found to be located in the generally allowed or disallowed Ramachandran regions in most structures. SIOCS (version 2007/07 alpha_test 0.1; Heisen & Sheldrick, in preparation) was used for prediction of the amide / imidazole orientations of asparagine, glutamine and histidine side chains. The refinement statistics are summarized in table 5 and listed in detail in each PDB entry. The number of molecules in the asymmetric unit as well as the number of amino acids, DNA residues, solvent and water molecules included in each final structure are given in table 6.

The structure 3G0R does not include any bromine atom at the residue 2 of chain G, although the oligonucleotide used for co-crystallization contained the nucleotide 5-Bromo-2`dC, since there was no hint for such a heavy atom at the appropriate position in the electron density map. Reasons for the absence might either be an inefficient incorporation of the modified nucleotide during DNA synthesis, a break of the bond between the appropriate C5 of residue 2 in chain G and the bromine atom caused by heating or light during the annealing procedure or crystallization. Another explanation for the absence of the bromine label might be radiation damage during data collection.

The refinement of the perfectly twinned Mth0212-DNA complex structure 3GA6 is described in more detail in the results section. The arrangement of the DNA helices cannot be correctly described in the apparent space group $C222_1$ due to the fact that one of the 2-fold axes would divide the pseudo-continuous helices in the middle of its length. Since the oligonucleotides used for crystallization did not represent palindromic sequences they cannot be depicted by half of the sequence in one asymmetric unit and its symmetry equivalents in the adjacent asymmetric unit (separated by the 2-fold axis). Though disorder of the DNA molecules within the crystal cannot be excluded but is even assumed, there was no possibility to determine the occupancies of the DNA oligonucleotides bound in opposite directions undoubtedly (with their 5`- and 3`- end next to the intercalating residue R209 of the two protein molecules in the asymmetric unit). That`s why the final model was refined in $P2_1$ with PHENIX including the twin law "l,-k,h". The choice of the monoclinic space group was supported further by the final value 0.497 of the refined twin fraction, although the difference to the description of a perfect twin (0.5) is only small. However, the interpretation of the model is not influenced by these crystallographic methodical considerations.

Accession numbers

The coordinates and structure factors were deposited in the Protein Data Bank with the accession codes listed in table 6.

Structure analysis

The multiple sequence alignment was performed with T-COFFEE. For superposition, the chains B, X and Y of the structure 1DEW, chain B of 1VYB and chain A of 2J63, 2O3C and 2VOA were used. Superpositions were performed with SSM superpose [78; 87; 88], DALI or DALILite v.3 [42]. Alternatively, structures were aligned directly in PyMOL [94]. For the superposition of the Mth0212-DNA complex structure 3GA6 with human Ape1 in complex with DNA the structural alignment matrix from DALILite v.3 [42] using the residues 4-255 of chain A of 3GA6 and 62-316 of chain A of the PDB entry 1DEW was applied in order to exclude flexible N- and C-terminal residues to get the best fit. Calculations of the electrostatic surface potential were performed with DELPHI 4.1 (Rocchia et al., 2002). The solvent accessible surface buried upon DNA binding was calculated with AREAIMOL [87; 89] from the CCP4 suite [78] using a 1.4 Å probe radius [90]. A modified version of the program AMIGOS [39] kindly provided by Tim Grüne (Department of Structural Chemistry, Institute of Inorganic Chemistry, Georg-August-University of Goettingen) was used for analysis of the DNA backbone torsions and for comparison with common standard values in analogy to regions in the Ramachandran plot which are used for validation of protein geometry. The program NUCPLOT [91; 92] was used for schematic representations of protein-DNA interactions (Fig. 4c). Anomalous difference maps were calculated with FFT [93] using a 5.0 Å resolution cut off. In addition to the sites A and B representing manganese ions which are discussed in the results section, only two further but less significant peaks were detected caused by Cys35/Cys149 (4.3 σ) and Cys249 (4.1 σ).

Figure preparation

Figures 1-9, S1c, S1d, S2, S4 and S5 were created with PyMOL [94], whereas CCP4 MOLECULAR GRAPHICS (CCP4MG) was used for Figure S3. COOT was used to generate Fig. S1b.

Acknowledgements

We thank the synchrotron staff for excellent support during data collection: Uwe Müller, Jörg Schulz and Georg Zocher from BESSY, Berlin, Germany as well as Paul Tucker, Manfred Weiss from EMBL at DESY, Hamburg, Germany. We are grateful to Dr. Regine Herbst-Irmer and Dr. Piotr Neumann from the University of Goettingen, Germany, for help with the refinement of the twinned structure 3GA6. We thank PD Dr. Winfried Kramer and Swetlana Ber from the University of Goettingen, Germany, for fruitful discussions.

Table 1. Members of the ExoIII family of nucleases which have been studied by means of X-ray crystallography (*) as well as additional structures exhibiting a significant similarity on the level of the three-dimensional overall structure ($^{\#}$). ExoIII = exonuclease III from *Escherichia coli* [32], Ape1 = human Ape1 (formerly HAP1, Ref-1) [35; 40; 41; 95] (cf. also 2ISI: Agarwal & Naidu, to be published, deposited 2006), NEXO = 3`-5` exonuclease from *Neisseria meningitides* (2JC4, Carpenter et al., to be published, deposited 2006), NAPE1 = AP endonuclease from *Neisseria meningitides* (2JC5, Carpenter et al., to be published, deposited 2006), Dr_Ape1 = Ape from *Danio reo* (zebrafish) (2O3C, Georgiadis et al., to be published, deposited 2006), LMAP = LMAP from *Leishmania major* [21], Af_Exo = exonuclease III from *Archaeoglobus fulgidus* [36], DNaseI = bovine DNaseI [43; 44; 96; 97; 98; 99], LINE-1 = hairpin exchange variant of the human targeting LINE-1 retrotransposon endonuclease [100; 101] (see also Tab. S1).

Abbreviation of the enzyme name	PDB entries
ExoIII	1AKO
Ape1	1BIX, 1DE9, 1DEW, 1E9N, 1HD7, 2ISI, 2O3H
NEXO	2JC4
NAPE1	2JC5
Dr_Ape1	2O3C
LMAP	2J63
Af_Exo	2VOA
DNaseI	1DNK, 2DNJ, 3DNI
LINE-1	2V0R, 1VYB, 2V0S

Table 2. Sequences of the DNA oligonucleotides which were used for complex formation with Mth0212 and subsequent co-crystallization yielding a structure deposited in the PDB. The nucleotides visible in the respective complex structure are marked in grey (,when only the phosphate moiety is included in the final structure, the respective nucleotide is not highlighted). 2`-deoxyuridine (U) and 2`-deoxy-5`-O-thiophosphonouridine (sU) residues are highlighted in orange. $C^{Br}$ = 5-Bromo-2`-deoxycytidine, # = ROX NHS 5/6-mixture connected to the nucleotide at the 3`-end of the DNA strand via the amino linker C6 CPG. The second column specifies whether wild-type (WT) or the mutant protein Mth0212(D151N) was crystallized. # The mutation T2A which was introduced due to the cloning strategy occurs in all protein batches and thus is not listed separately. The ratio between GC and AT base pairs is given in column 3 together with the base opposite of the 2`-deoxyuridine residue.

PDB-ID	Mth0212: WT / D151N #	base pair ratio GC:AT; opposite U	DNA oligonucleotides used for co-crystallization
3G2C	WT	4:4 G	5` - CGTA sU TACG – 3` 3` - GCAT GATGC – 5`
3G3Y	WT	4:4 C	5` - CGTA sU TAC G# – 3` 3` - #GCAT C ATGC – 5`
3G2D	D151N	6:1 G	5`- CCTG U GCGAT – 3` 3`- CGGACGCGC – 5  5`- CCTG U GCGAT – 3` 3`- CGGACGCGC – 5
3G3C	WT	4:4 C	5` - CGTA sU TAC G# – 3` 3` - #GCAT CATGC – 5`
3G4T	WT	4:4 C	5` - CGTA sU TAC G# – 3` 3` - #GCAT C ATGC – 5`
3G38	D151N	6:1 G	5`- CCTG U GCGAT – 3` 3`- CGGACGCGC – 5
3G0R	D151N	7:2 G	5` - CCCTG U GCAG C – 3` 3` - GCGGGACGCGTC$^{Br}$G - 5`
3G00	D151N	4:4 T	5` - CGTA U TACG – 3` 3` - GCATTATGC – 5`
3GA6	D151N	9:2 G	5` - GCCCTG U GCAGC – 3` 3` - CGGGACGCGTCG - 5` 5` - GCCCTG U GCAGC – 3` 3` - CGGGACGCGTCG - 5`
apo structure:			
3G1K		4:4 G	5` - CGTA sU TACG – 3` 3` - GCAT GATGC – 5`

Table 3. Conditions of protein solution, complex formation and crystallization experiment used for structure determination. “DNA:protein” indicates the molar ratio between protein and dsDNA used for co-crystallization. * “v/v”: vol. (probe) + vol. (res.) [µl].

PDB-ID	protein / complex solution (probe)	reservoir solution (res.)	v/v*	cryo protecting solution
3FZI	600 mM NaCl 20 mM HEPES-KOH pH 7.6 2 mM DTT 12.6 mg/ml Mth0212(WT)	20 % (w/v) PEG 1500 100 mM HEPES pH 7.5		20 % (w/v) PEG 1500 100 mM HEPES-KOH pH 7.5 2 mM DTT
3G1K	1 mM MgCl$_2$ 50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 12.7 mg/ml Mth0212(WT) DNA : protein: = 1.1:1.0	10 % (w/v) PEG 400 100 mM sodium cacodylate pH 6.5	0.7 + 0.7	7 % (w/v) PEG 400 75 mM sodium cacodylate pH 6.5 25 % (v/v) glycerol
3G8V	120 mM NaCl 2 mM DTT 8 mM HEPES-KOH pH 7.6 8.7 mg/ml Mth0212(D151N) dUMP:Mth0212 = 83:1	15 % (w/v) PEG 3350 33 mM MnCl$_2$; in the drop: 2x concentrated and buffered reservoir: 30 % PEG3350, 100 mM MnCl$_2$, 200 mM sodium cacodylate pH 6.5	3.34+ 0.66	11 % (w/v) PEG 3350 25 mM MnCl$_2$ 25 % (v/v) glycerol
3G91	180 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 12.0 mg/ml Mth0212 (K116A) DNA : protein: = 1.3:1.0	20 % (v/v) MPD 40 mM MgAc 50 mM sodium cacodylate pH 6.0	0.7 + 0.7	35 mM MgAc 45 mm sodium cacodylate pH 6.0 27 % (v/v) MPD 3 % (v/v) glycerol
3G0A	600 mM NaCl 20 mM HEPES-KOH pH 7.6 2 mM DTT 12.6 mg/ml Mth0212(WT)	10 % (w/v) PEG 20000 100 MES pH 6.5	1.0 + 1.0	6 % (w/v) PEG 20000 23 % (v/v) glycerol 60 mM MES-NaOH pH 6.5 200 mM MnCl$_2$
3G2C	50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 1 mM MgCl$_2$ 12.7 mg/ml Mth0212(WT) DNA : protein = 1.1:1.0	25 % (v/v) MPD 100 mM HEPES-KOH pH 7.0	0.7 + 0.7	18 % (v/v) MPD 75 mM HEPES-KOH pH 7.0 25 % (v/v) glycerol
3G3Y	50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 1 mM MgCl$_2$ 12.0 mg/ml Mth0212(WT) DNA : protein = 1.2:1.0	40 mM MgAc 50 mM sodium cacodylate pH 6.0 20 % (v/v) MPD	0.7 + 0.7	35 mM MgAc 45 mm sodium cacodylate pH 6.0 27 % (v/v) MPD 3 % (v/v) glycerol

(Table 3 continued:)

PDB-ID	protein / complex solution (probe)	reservoir solution (res.)	v/v*	cryo protecting solution
3G2D	240 mM NaCl 8 mM HEPES-KOH pH 7.6 4 mM DTT 2 mM MgCl2 1 mM KH_2PO_4/K_2HPO_4 pH 7.0 19.0 mg/ml Mth0212 (D151N) DNA : protein = 1.3:1.0	5 % (w/v) PEG 4000 50 mM KCl 50 mM MES-NaOH pH 5.8 10 mM $MgCl_2$	0.7 + 0.7	3 % (w/v) PEG 4000 30 mM KCl 30 mM MES-NaOH pH 5.8 7 mM $MgCl_2$ 33 % (v/v) glycerol
3G3C	50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 1 mM $MgCl_2$ 12.0 mg/ml Mth0212(WT) DNA : protein = 1.2:1.0	40 mM MgAc 50 mM sodium cacodylate pH 6.0 20 % (v/v) MPD	0.7 + 1.0	35 mM MgAc 45 mm sodium cacodylate pH 6.0 27 % (v/v) MPD 3 % (v/v) glycerol
3G4T	50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 1 mM $MgCl_2$ 8.0 mg/ml Mth0212(WT) DNA : protein = 1.2:1.0	40 mM MgAc 50 mm sodium cacodylate pH 6.0 30 % (v/v) MPD	0.7 + 0.7	40 mM MgAc 50 mm sodium cacodylate pH 6.0 30 % (v/v) MPD
3G38	180 mM NaCl 6 mM HEPES-KOH pH 7.6 2 mM MgCl2 2 mM DTT 10 mM KCl 2 mM KH_2PO_4/K_2HPO_4 pH 7.0 14.9 mg/ml Mth0212 (D151N) DNA : protein = 1.4:1.0	5 % (w/v) PEG4000 50 mM KCl 100 mM MES pH 5.6 10 mM $MgCl_2$	0.9 + 0.9	5 % (w/v) PEG4000 45 mM KCl 90 mM MES pH 5.6 10 mM $MgCl_2$ 10 % (v/v) glycerol
3G0R	234 mM NaCl 7 mM HEPES-KOH pH 7.6 2 mM $MgCl_2$ 3 mM DTT 13.4 mg/ml Mth0212 (D151N) DNA : protein = 2.1:1.0	189 mM KCl 10 mM $MgCl_2$ 4.7 % (w/v) PEG 8000 47 mM MES-NaOH pH 5.6 2.1 % (w/v) 1,4-butandiol 2 mM DTT	1.0 + 2.0	3.1 % (w/v) PEG 8000 126 mM KCl 6 mM MgCl2 32 mM MES-NaOH pH 5.6 1.4 % (w/v) 1,4-butandiol 33 % (v/v) glycerol
3G00	50 mM KCl 10 mM KH_2PO_4/K_2HPO_4 pH 7.0 1 mM $MgCl_2$ 8.8 mg/ml Mth0212 (D151N) DNA:protein = 1.2:1.0	30 % (v/v) MPD 40 mM $MgCl_2$ 50 mM KH_2PO_4/K_2HPO_4 pH 7.0	0.7 + 0.7	30 % (v/v) MPD 40 mM $MgCl_2$ 50 mM KH_2PO_4/K_2HPO_4 pH 7.0
3GA6	245 mM NaCl 8 mM HEPES-KOH pH 7.6 2 mM MgCl2 3 mM DTT 19 mg/ml Mth0212 (D151N) DNA:protein = 1.4:1.0	5 % (w/v) PEG 8000 200 mM KCl 10 mM MgCl2 50 mM MES-NaOH pH 5.6 2 mM DTT	0.6 + 0.6	4 % (w/v) PEG 8000 150 mM KCl 8 mM MgCl2 40 mM MES-NaOH pH 5.6 25 % (v/v) glycerol

Table 4. Space groups and unit cell parameters of the Mth0212 apo and –DNA complex structures.

PDB-ID		
	space group	unit cell [Å, °]
3FZI	$P6_5$	56.040 / 56.040 / 161.320 90.00 / 90.00 / 120.00
3G1K	$P2_1$	44.635 / 80.345 / 81.839 90.00 / 90.32 / 90.00
3G8V	$P6_5$	60.310 / 60.310 / 149.450 90.00 / 90.00 / 120.00
3G91	$P2_1$	44.330 / 72.110 / 46.330 90.00 / 117.96 / 90.00 °
3G0A	$P6_5$	56.381 / 56.381 / 162.932 90.00 / 90.00 / 120.00
3G2C	$P3_2$	80.325 / 80.325 / 79.611 90.0 / 90.0 / 120.0
3G3Y	$P3_2$	80.487 / 80.487 / 79.749 90.00 / 90.00 / 120.00
3G2D	$P2_1$	44.603 / 81.301 / 97.091 90.00 / 90.00 / 90.00
3G3C	$P2_{1}2_{1}2$	100.354 / 79.337 / 98.550 90.00 / 90.00 / 90.00
3G4T	$P2_{1}2_{1}2$	100.764 / 79.600 / 99.409 90.00 / 90.00 / 90.00
3G38	$P2_{1}2_{1}2$	79.960 / 107.150 / 44.270 90.0 / 90.0 / 90.0
3G0R	$P2_1$	44.751 / 80.760 / 105.211 90.00 / 94.03 / 90.00
3G00	$P2_1$	48.966 / 79.510 / 87.750 90.00 / 97.76 / 90.00
3GA6	$P2_1$	54.834 / 126.655 / 54.826 90.00 / 93.14 / 90.00

Table 5. Data collection and refinement statistics. Numbers in parentheses refer to the outer resolution shell. $R_{sym} = 100\ \Sigma h\ \Sigma i\ |\Sigma\ Ii(h) - \langle I(h)\rangle|\ /\ \Sigma h\ I(h)$, where Ii(h) is the *i*th measurement of reflection h of the average reflection intensity ‹I(h)›. * hs = RIGAKU MICROMAX-007 (home source).

PDB-ID	beamline, wavelength [Å]	resolution [Å]	R_{sym} [%]	I/ sig(I)	redundancy	completeness [%]	unique reflections	R_{work} R_{free} [%]
3FZI	EMBL/DESY X13 0.80150	50.0-1.90 (1.97-1.90)	5.8 (55.8)	36.3 (4.0)	7.5 (7.3)	99.8 (100.0)	22592	19.8 26.2
3G1K	EMBL/DESY X11 0.80148	39.28-3.10 (3.18-3.10)	14.8 (33.2)	9.8 (4.0)	4.0 (3.5)	99.0 (94.1)	10535	19.7 29.0
3G8V	EMBL/DESY 0.97900	50.0-2.40 (2.49-2.40)	9.3 (39.1)	11.7 (3.3)	3.2 (3.2)	96.7 (100.0)	73402	22.0 28.0
3G91	BESSY BL14.2 0.91841	19.40-1.23 (1.27-1.23)	2.9 (13.1)	28.1 (9.0)	4.06 (3.66)	97.9 (92.7)	297294	12.2 17.2
3G0A	hs 1.5418	50.0-2.60 (2.69-2.60)	9.5 (35.1)	17.0 (2.4)	6.5 (2.4)	96.6 (82.6)	8864	19.0 25.2
3G2C	EMBL/DESY X11 0.80150	30.0-2.30 (2.38-2.30)	8.5 (34.7)	17.1 (4.1)	5.5 (4.5)	99.8 (99.3)	25521	26.3 20.4
3G3Y	BESSY BL14.2 1.00605	50.0-2.50 (2.59-2.50)	6.4 (35.5)	60.1 (4.6)	4.9 (4.9)	100.0 (100.0)	19991	21.1 30.0
3G2D	EMBL/DESY X13 0.80150	50.00-2.30 (2.38-2.30)	5.5 (19.1)	18.0 (4.2)	2.9 (2.6)	93.8 (76.3)	29120	20.8 29.8
3G3C	BL14.2 0.91841	50.0-3.04 (3.15-3.04)	8.2 (50.9)	26.3 (4.1)	7.3 (6.8)	100.0 (100.0)	15706	22.0 30.1
3G4T	BESSY BL14.2 0.91841	50.0-2.64 (2.73-2.64)	4.7 (41.7)	36.2 (2.5)	7.6 (4.6)	97.9 (83.1)	23676	22.7 30.9
3G38	EMBL/DESY X11 0.80150	36.42-2.74 (2.89-2.74)	11.7 (47.1)	5.9 (1.6)	7.6 (7.8)	100.0 (100.0)	10553	25.6 32.8
3G0R	EMBL/DESY X12 0.97623	50.0-2.40 (2.49-2.40)	6.1 (19.1)	22.7(5.2)	4.7 (3.2)	92.8 (72.3)	27308	17.8 24.3
3G00	BESSY BL14.1 0.91838	17.0-1.74 (1.80-1.74)	3.3 (26.8)	15.7 (3.2)	2.1 (1.8)	94.9 (72.0)	65383	16.7 21.7
3GA6	BESSY BL14.2 0.91800	50.0-1.90 (1.97-1.90)	7.5 (60.7)	22. (1.8)	6.7 (4.1)	95.8 (72.2)	56272	15.5 20.3

Table 6. Structures deposited in the PDB. The kind of protein (WT/D151N/K116A) used for each PDB-ID (column 1) is listed in the second column (“Mth0212”). The mutation T2A which was introduced due to the cloning strategy occurs in all protein batches and is not listed separately. The column “protein molecules” specifies the number of protein molecules per asymmetric unit followed by the amino acid residues included in the final model in parentheses. The column “DNA strands” gives the number of DNA strands and indicates whether it is single- (ss) or double-stranded (ds) DNA as well as the number of nucleotides (nt) in each strand with the number of non-pairing bases (“nucleotide overhangs”) labelled with “o” and phosphate moieties at the 3`- or 5`-end marked by “3P” and “5P”, respectively. The number of ligand, solvent and water molecules or ions is given in the last column using the following abbreviations: “g” = glycerol, “P” = PO_4^{3-}, “pg4” = tetraethylene glycol, “w” = water.

PDB-ID	Mth 0212	title	protein molecules	DNA strands	solvent molecules
3FZI	WT	1.9 Angstrom structure of the thermophilic exonuclease III homolog Mth0212	1 (A2-E259)	-	1 Mg^{2+}, 292 w
3G1K	WT	Mth0212 (WT) crystallized in a monoclinic space group	2 (A2-L257, A2-L257)	-	2 Mg^{2+}, 38 w
3G8V	D151N	The rationally designed catalytically inactive mutant Mth0212(D151N)	1 (A2-E256)	-	3g, 1 pg4, 100 w
3G91	K116A	1.2 Angstrom structure of the exonuclease III homolog Mth0212	1 (A2-H261)	-	1 Mg^{2+}, 1 P, 1 g, 1 pg4, 1 PEG, 433 w
3G0A	WT	Mth0212 with two bound manganese ions	1 (V3-E259)	-	2 Mn^{2+}, 1 g, 1 P, 70 w
3G2C	WT	Mth0212 in complex with a short ssDNA (CGTA)	2 (A2-L258, V3-L257)	1 (ss, 4nt, 5P)	2 Mg^{2+}, 7 g, 3 P, 168 w
3G3Y	WT	Mth0212 in complex with ssDNA in space group P32	2 (A2-L258; A2-E256)		2 Mg^{2+}, 4 g, 65 w
3G2D	D151N	Complex of Mth0212 and a 4 bp dsDNA with 3`-overhang	2 (A2-L257; A2-E256)	4 (ds+ds, 8/8/9/9 nt, 4o+5o)	7 g, 4 pg4, 3 P, 329 w
3G3C	WT	Mth0212 (WT) in complex with a 6bp dsDNA containing a single one nucleotide long 3`-overhang	2 (A2-E256; A2-L258)	2 (ds, 6/7 nt, 1o)	2 Mg^{2+}, 1 MRD, 2 P, 6 w
3G4T	WT	Mth0212 (WT) in complex with a 7bp dsDNA	2 (V3-E256; M1-I255	2 (ds, 7/7 nt)	2 Mg^{2+}, 3 P, 16 w
3G38	D151N	The catalytically inactive mutant Mth0212 (D151N) in complex with an 8 bp dsDNA	1 (A2-L258)	2 (ds, 8/8 nt)	6 g, 48 w
3G0R	D151N	Complex of Mth0212 and an 8bp dsDNA with distorted ends	2 (V3-L257, V3-E256)	2 (ds, 8/8 nt)	2 Na^{+}, 12 g, 3 pg4, 305 w
3G00	D151N	Mth0212 in complex with a 9bp blunt end dsDNA at 1.7 Angstrom	2 (A2-L258, V3-L257)	2 (ds, 9/9 nt)	4 P, 3 g, 2 MPD, 659 w
3GA6	D151N	Mth0212 in complex with two DNA helices	2 (V3-E256, A2-E256)	4 (ds/ds, 12/11/12/11 nt, 1o/1o)	1Na^{+}, 2P, 19 g, 354 w

Figures

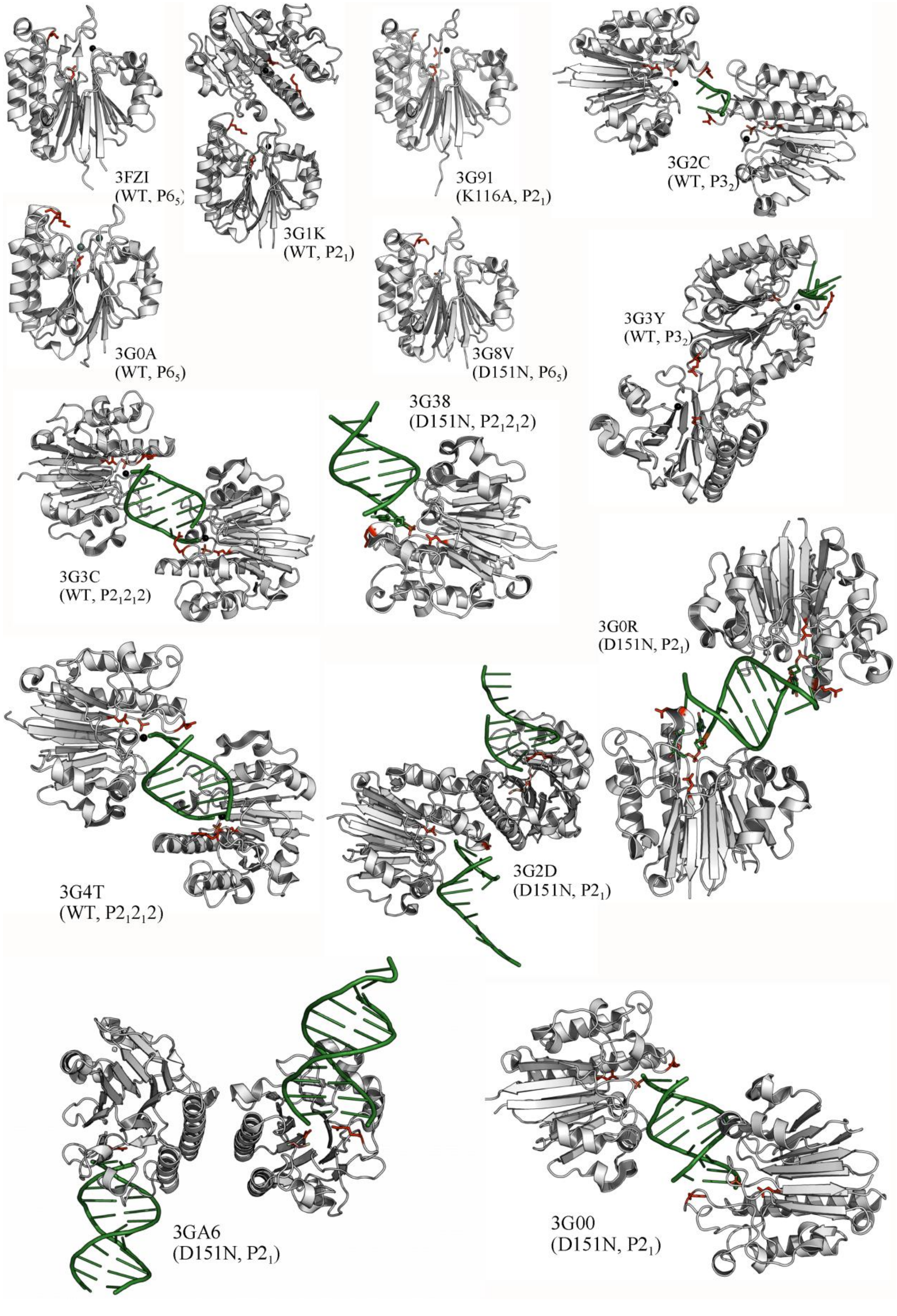

3FZI
(WT, $P6_5$)
3G1K
(WT, $P2_1$)
3G91
(K116A, $P2_1$)
3G2C
(WT, $P3_2$)
3G0A
(WT, $P6_5$)
3G8V
(D151N, $P6_5$)
3G3Y
(WT, $P3_2$)
3G38
(D151N, $P2_12_12$)
3G3C
(WT, $P2_12_12$)
3G0R
(D151N, $P2_1$)
3G4T
(WT, $P2_12_12$)
3G2D
(D151N, $P2_1$)
3GA6
(D151N, $P2_1$)
3G00
(D151N, $P2_1$)

Fig. 1. Overview of the structures of Mth0212 alone as well as in complex with distinct DNA substrates and products. Cartoon representations with protein and DNA molecules coloured in grey and dark green, respectively. Metal and phosphate ions bound in the active site are included as black spheres and orange-red sticks, respectively. The side chains of the residues 116 and 151 which have been replaced for the mutational studies are highlighted as red sticks. For structures with phosphate moieties of the DNA bound in the active site, the respective nucleotides are represented as sticks as well.

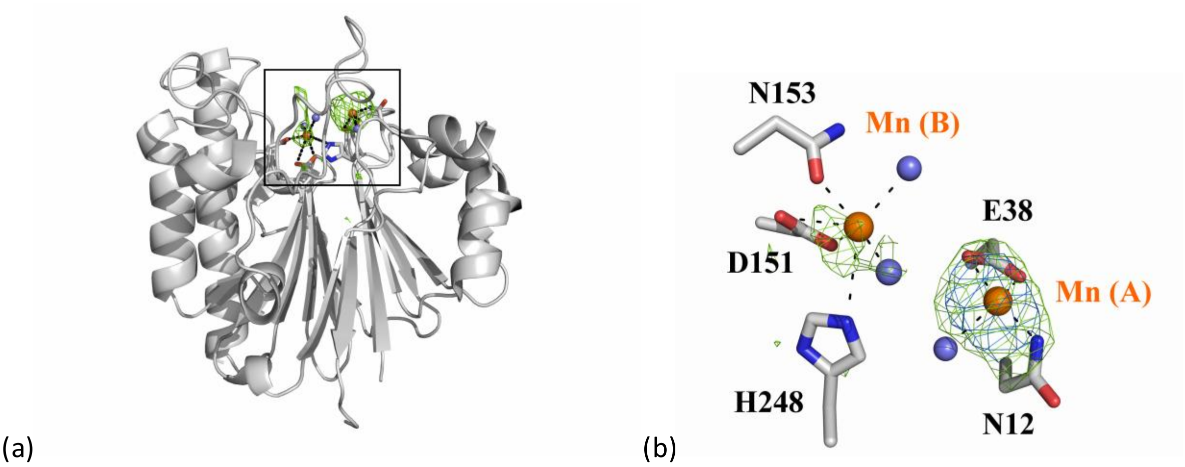
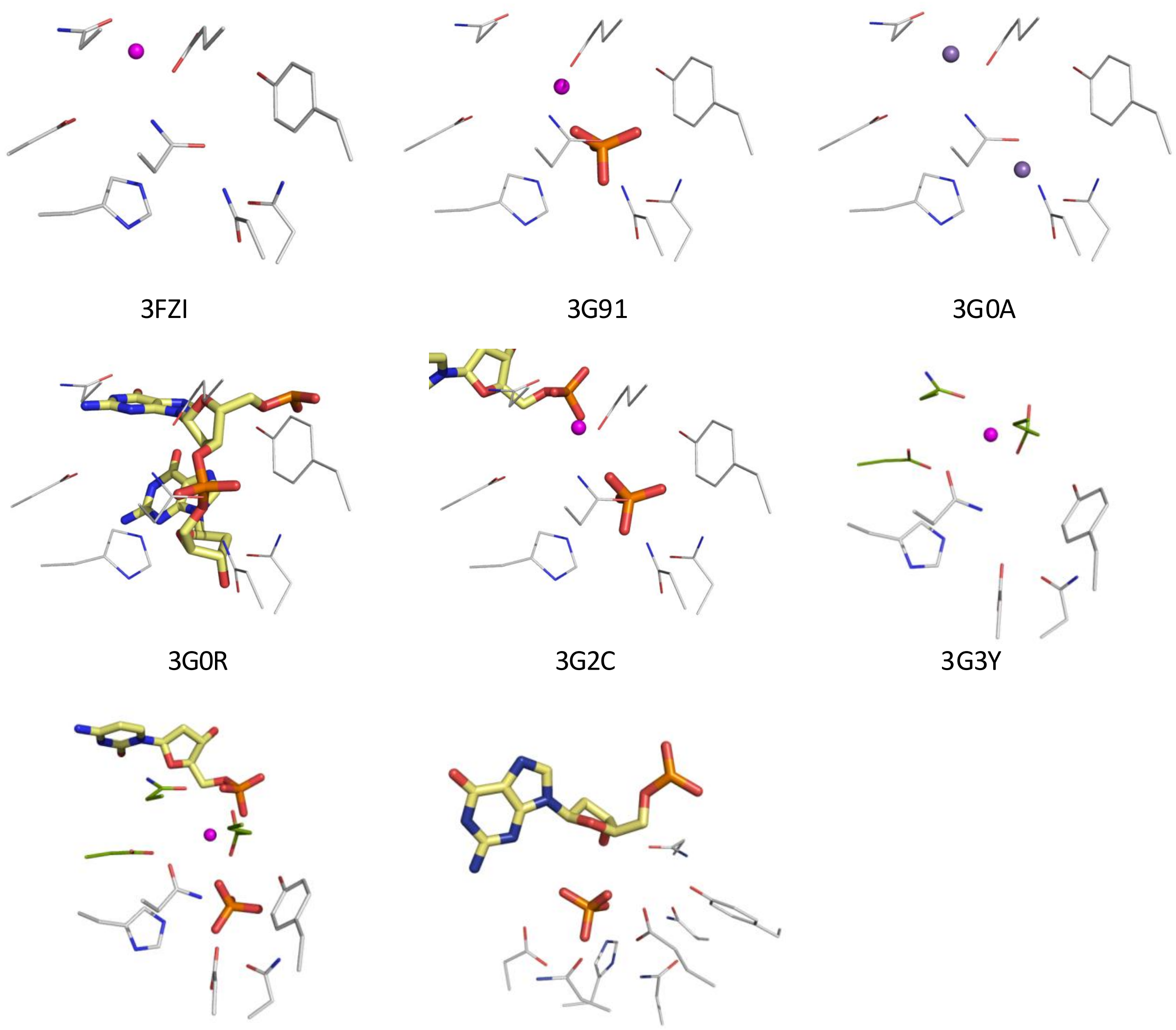


Fig. 2. Mth0212 (WT) with two bound manganese ions (3G0A). The anomalous difference map is contoured at a level of 5.0 σ and 3.0 σ in blue and green, respectively, the peaks represent the anomalously scattering metal ions bound at the coordination sites A and B. (a) Overall structure. The black rectangle specifies the region which is represented in an enlarged view in (b). (b) Close up on the metal coordination site, the location in the overall structure is indicated in (a).



3FZI 3G91 3G0A

3G0R 3G2C 3G3Y

3G2C 3G00

[Fig. 3, caption see below]

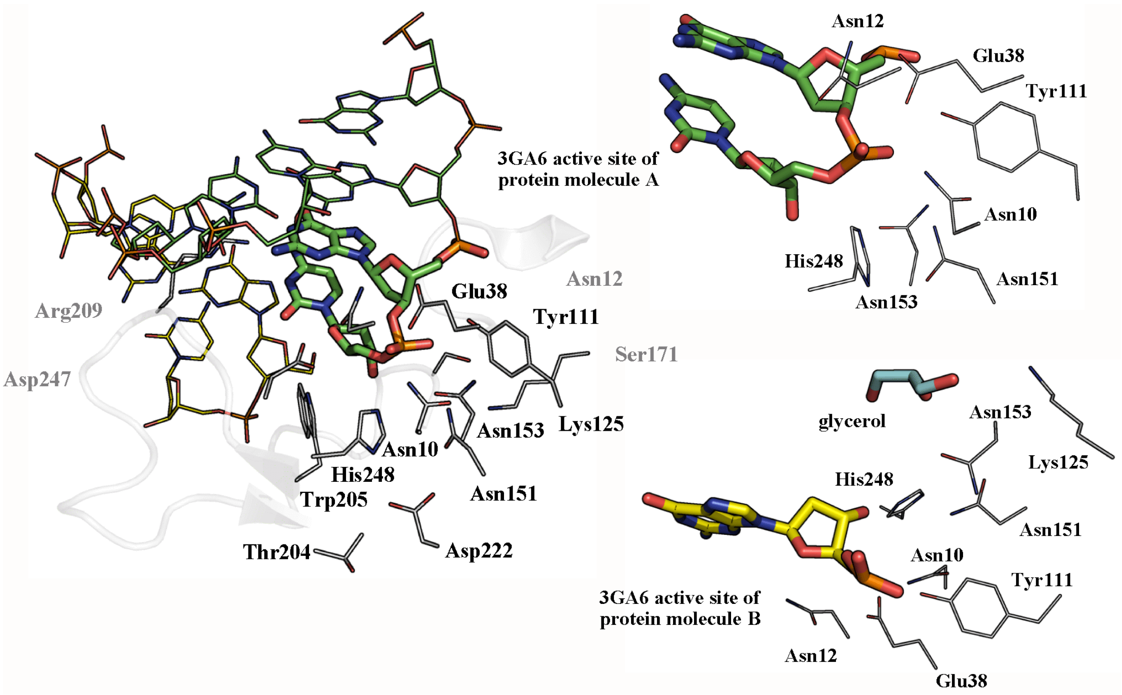
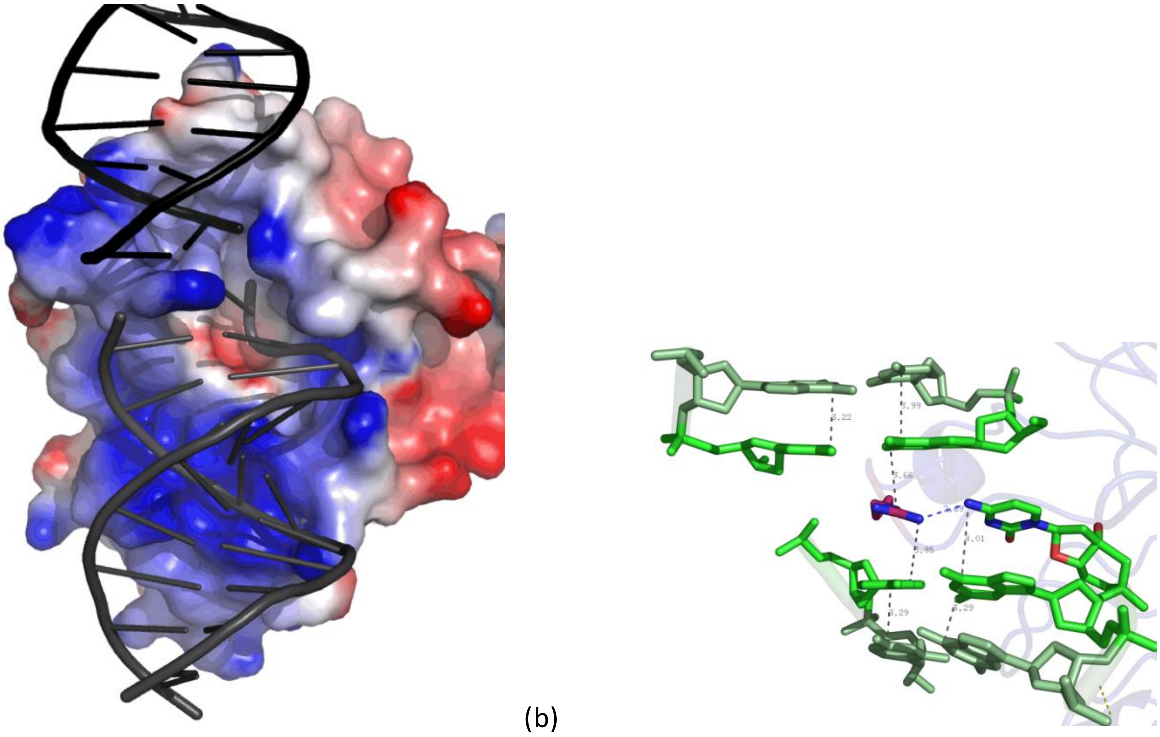
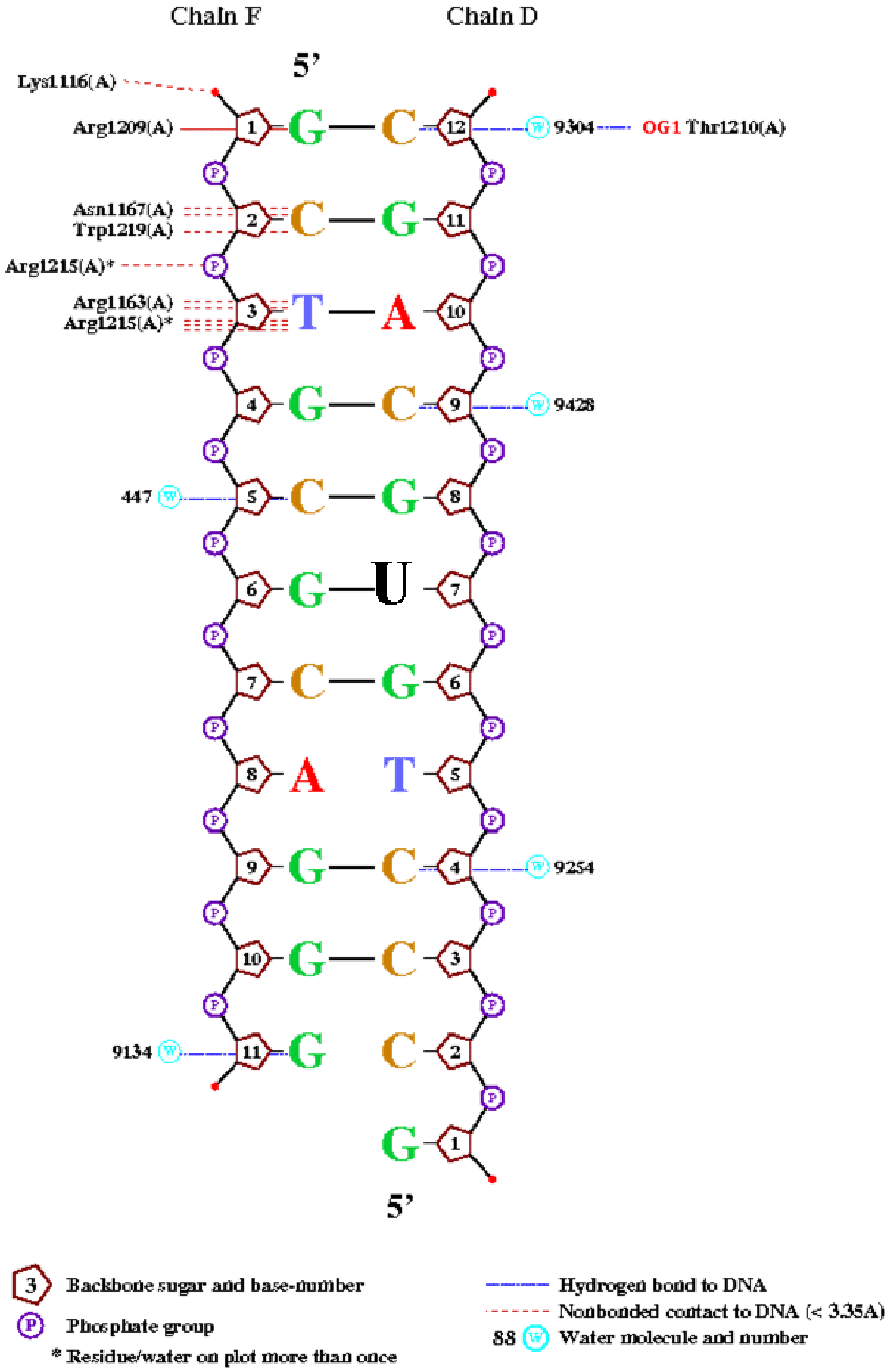


Fig. 3. Comparison of ligands bound in the active site of selected Mth0212 apo and –DNA complex structures. The respective PDB code is denoted with each representation. Except for the structure 3GA6 the amino acid residues only serve for orientation and thus are taken by the superposed structure 3G00, they are depicted in very similar orientations. In the representations of the complex structure 3GA6, different oritentations are shown, all residues derive from this PDB entry.

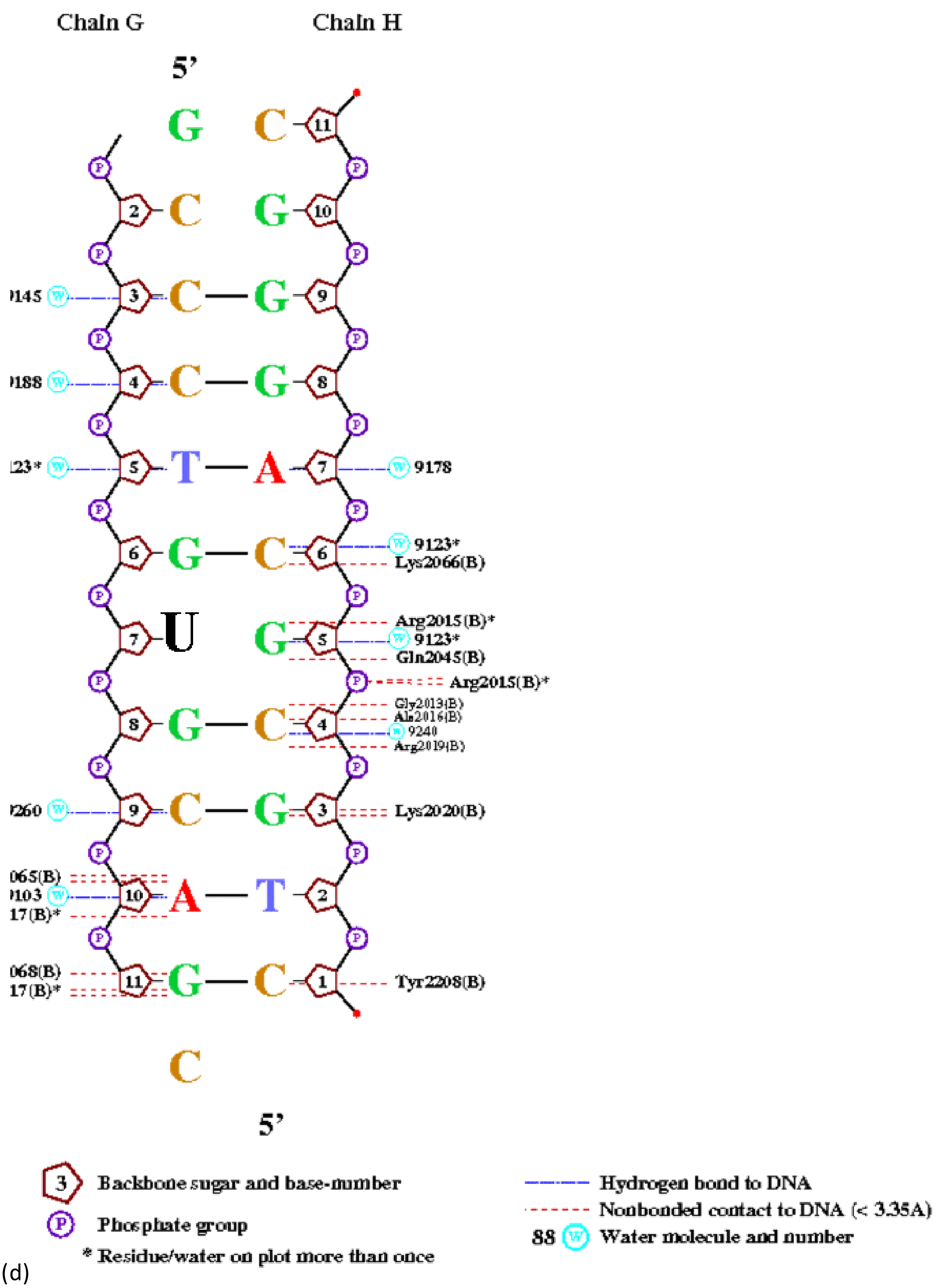


(a) (b)

[Fig. 4, caption see below]



(c) [Fig. 4, caption see below]



(d)

Fig. 4. Structure of Mth0212 in complex with two exonucleolytically bound DNA helices (3GA6). (a) Surface representation of the protein residues (molecule A) coloured according to the electrostatic surface potential. The bound DNA is represented as cartoon model and coloured in black. (b) The intercalating side chain Arg209 (coloured by atom in magenta and blue, respectively). Stacking interactions are indicated by broken lines with the distances roughly given in Å. (c, d) Schematic representation of the Mth0212-DNA interactions for the two DNA duplexes formed by (c) chains D and F and (d) chains G and H, resepctively, that are bound by the two protein molecules (A and B) of the asymmetric unit.

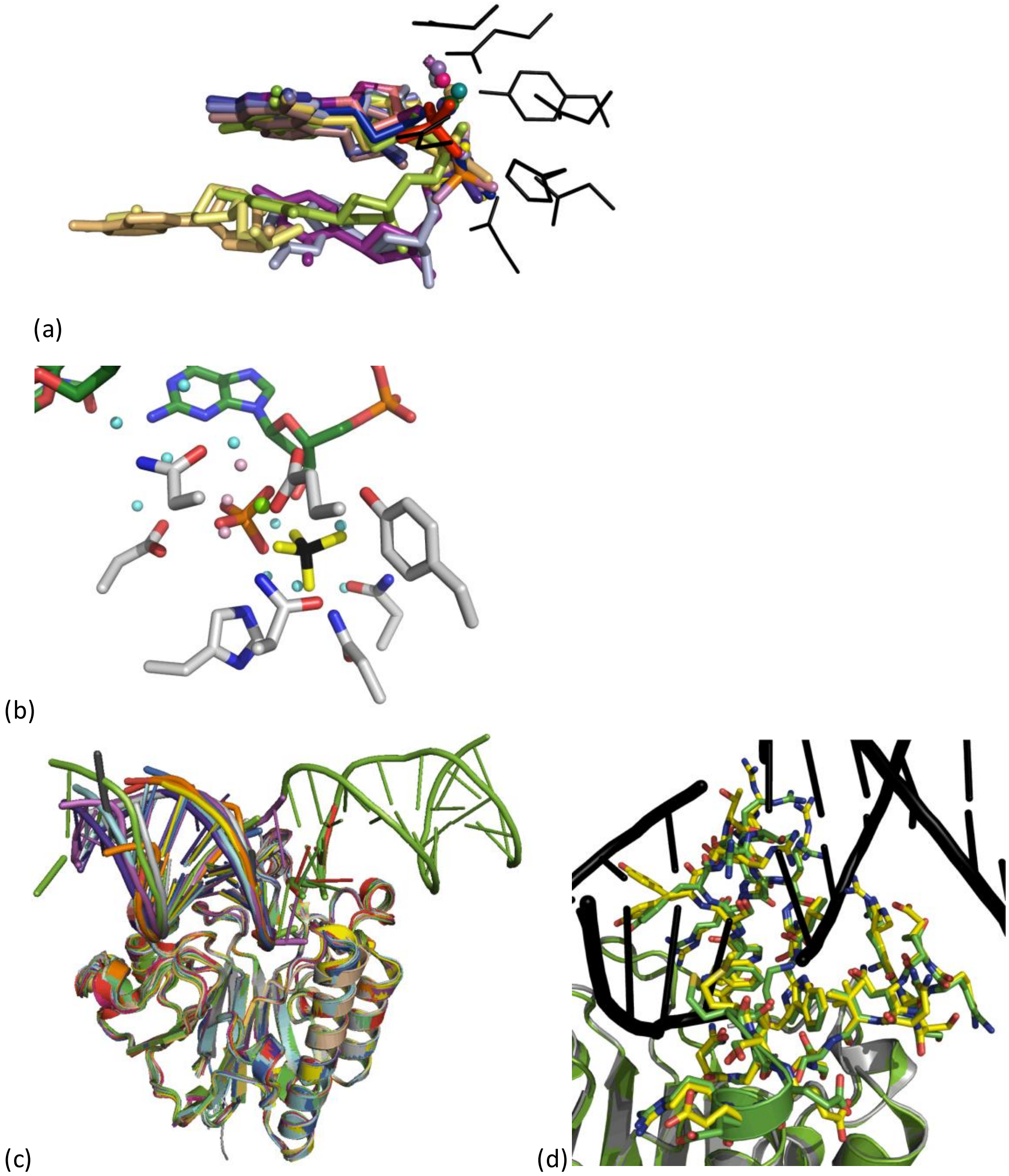


Fig. 5. (a) Superposition of all Mth0212 apo and Mth0212-DNA complex structures harbouring a phosphate group in the active site. Only this phosphate group and if present residues bound to it are shown in stick mode. Magnesium and manganese ions are represented as spheres. For orientation the side chains of the coordinating amino acids are shown as black lines. Except for the phosphate anion in 3G00 (coloured in red) all phosphate groups fit well with each other and are bound at coordination site B. (b) Superposition of the phosphate groups coordinated in the high resolution structure of Mth0212 alone (3G91) and the Mth0212-DNA complex 3G00. The sticks are coloured by atom, residues of the coordinating side chains are shown for 3G00. Ordered solvent molecules (in blue and rose for 3G00 and 3G91, respectively) and the Mg^{2+} ion coordinated in 3G91 (in green) are shown as spheres. (c) Superposition of all Mth0212 apo and –DNA complex structures in cartoon representation coloured by each PDB entry (cf. Tab. 6). (d) Close up on the three DNA binding specificity loops which are shown in stick mode. For clarity, only the high resolution apo form 3G91 and the complex 3GA6 were superposed.

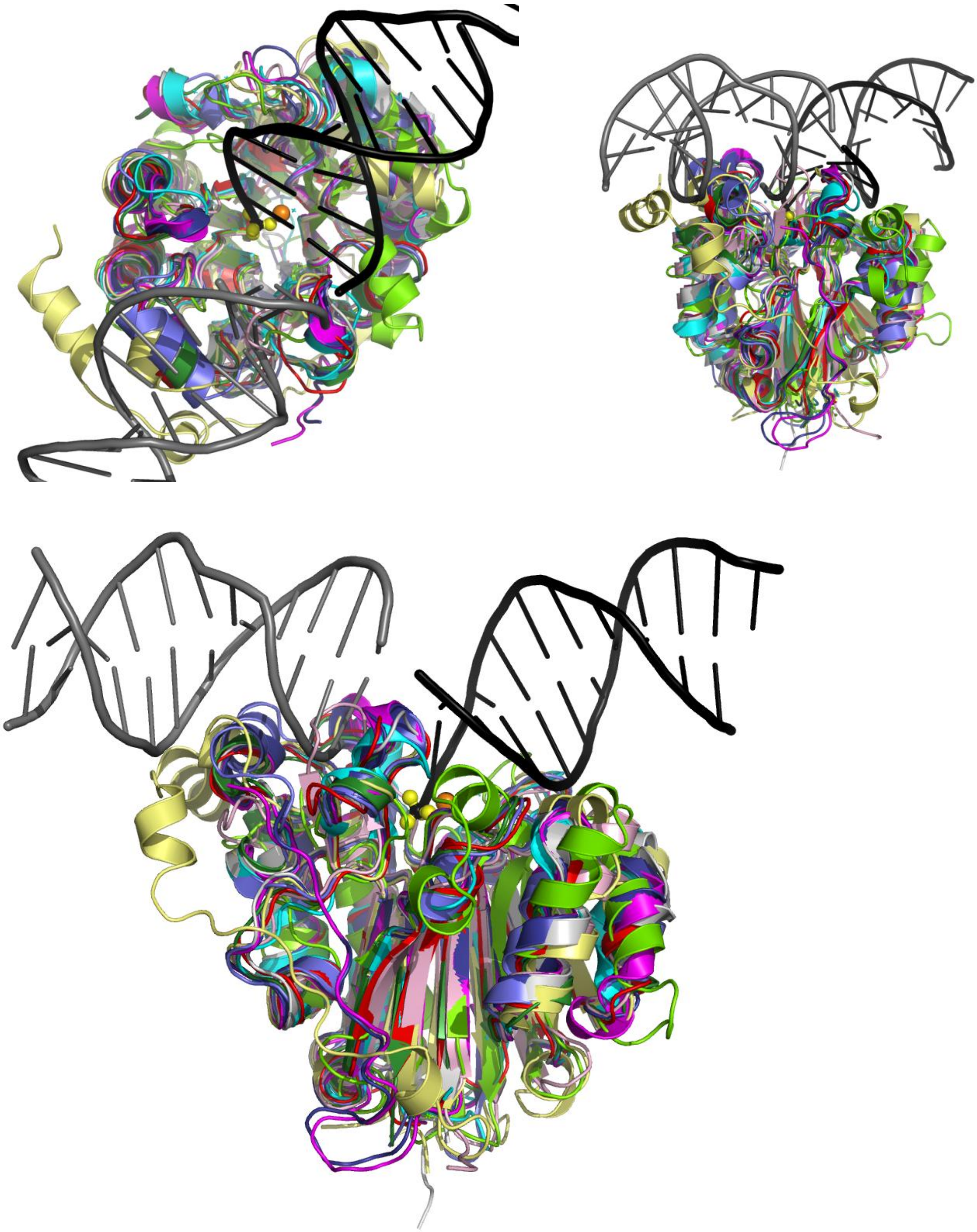


Fig. 6. Superposition of the crystal structures representing Mth0212 with homologuous structures in cartoon mode in different oritentations. The PDB codes of the used structures are given in brackets. For orientation, the phosphate anion and the magnesium ion bound in 3G91 as well as the DNA duplices bound in 3GA6 are shown as sticks, sphere and cartoon, respectively. The following colour code was applied: Mth0212 in dark grey, human Ape1 in complex with DNA (1DEW) in light green, ExoIII from *Escherichia coli* (1AKO) in dark blue, an AP endonuclease from *Archaeoglobus fulgidus* (2VOA) in light blue, LMAP from *Leishmania major* (2J63) in dark green, zebrafish Ape (2O3C) in magenta as well as the 3`-5` exonuclease NEXO from *Neisseria meningitidis* (2JC4) in red and NAPE from *N. meningitidis* (2JC5) in orange, the endonuclease domain of human LINE1 ORF2P (1VYB) in yellow and bovine DNaseI in complex with DNA (1DNK) in light grey.

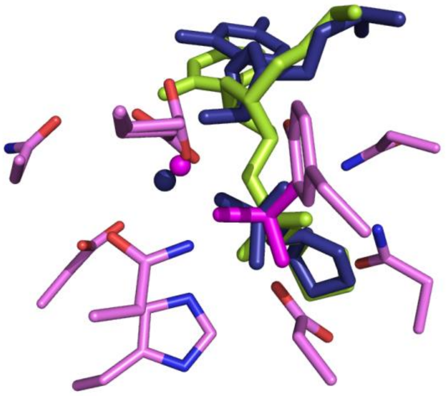


Fig. 7. Superposition of the structures of Mth0212 alone (3G91) with Ape1 in complex with uncleaved and incised AP-DNA (1DEW and 1DE9 in green and blue, respectively). The phosphate groups fit well.

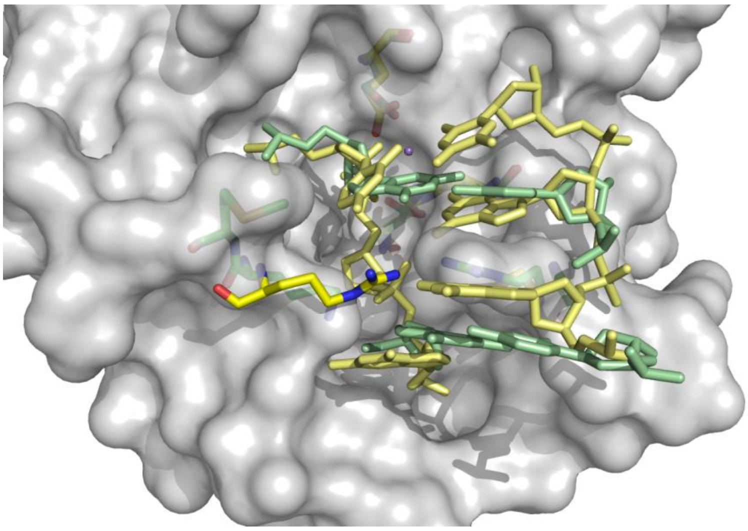


Fig. 8. Superposition of the Mth0212-DNA complex 3GA6 in partially transparent surface representation of the protein residues with the intercalating / DNA contacting residues Met117 and Arg209 as well as the active site residues and the nucleotides close to the active site in stick mode. Arg177 of an Ape1-DNA complex is shown in stick mode as well with carbon atoms coloured in bright yellow).

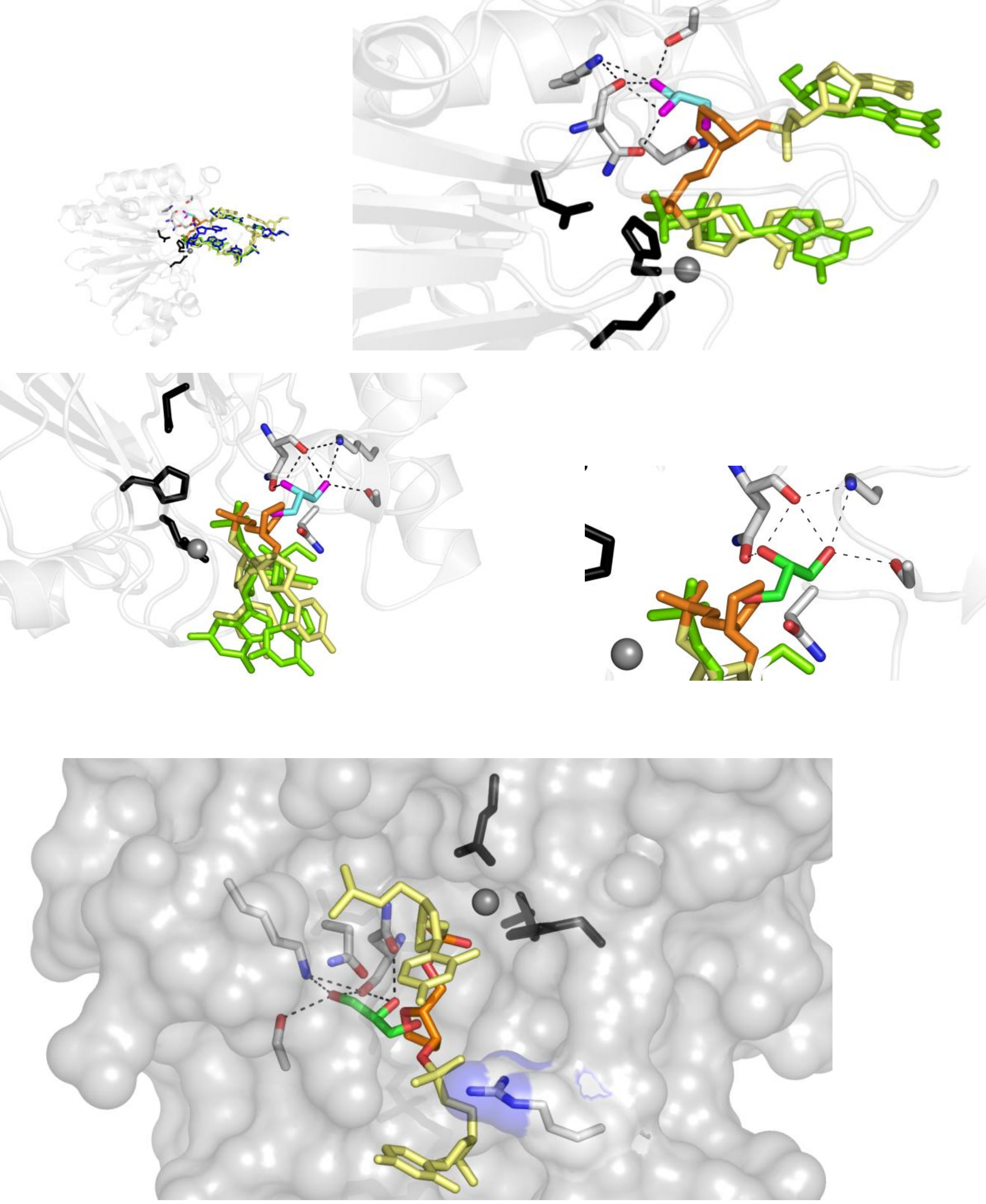


Fig. 9. Glycerol molecule in the active site pocket of molecule B of the Mth0212-DNA complex 3GA6 (carbon atoms in cyan and oxygen atoms in magenta). For comparison, Ape1 in complex with uncleaved AP-DNA (1DEW) is superposed, the DNA residues are shown in light yellow with the AP site highlighted in orange. Additionally, in the overview the DNA residues which are bound at protein molecule A in the same complex are shown (in blue). The catalytically essential residues Glu38, Asp222 and His248 are shown as black sticks. The amino acids which form hydrogen bonds with the bound glycerol molecule are shown in stick mode and coloured by atom (Lys125, Asn114, Asn153, Ser171) (for some residues only the side chains).

References

1. Bernstein, C. B., H. (1991). *Aging, Sex and DNA Repair*, Academic Press.
2. Ames, B. N., Shigenaga, M. K. & Hagen, T. M. (1993). Oxidants, antioxidants, and the degenerative diseases of aging. *Proc Natl Acad Sci U S A* **90**, 7915-22.
3. Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature* **362**, 709-15.
4. Fondufe-Mittendorf, Y. N., Harer, C., Kramer, W. & Fritz, H. J. (2002). Two amino acid replacements change the substrate preference of DNA mismatch glycosylase Mig.MthI from T/G to A/G. *Nucleic Acids Res* **30**, 614-21.
5. Tye, B. K., Nyman, P. O., Lehman, I. R., Hochhauser, S. & Weiss, B. (1977). Transient accumulation of Okazaki fragments as a result of uracil incorporation into nascent DNA. *Proc Natl Acad Sci U S A* **74**, 154-7.
6. Wist, E., Unhjem, O. & Krokan, H. (1978). Accumulation of small fragments of DNA in isolated HeLa cell nuclei due to transient incorporation of dUMP. *Biochim Biophys Acta* **520**, 253-70.
7. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. & Sperens, B. (1977). DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *J Biol Chem* **252**, 3286-94.
8. Aravind, L. & Koonin, E. V. (2000). The alpha/beta fold uracil DNA glycosylases: a common origin with diverse fates. *Genome Biol* **1**, RESEARCH0007.
9. Georg, J., Schomacher, L., Chong, J. P., Majernik, A. I., Raabe, M., Urlaub, H., Muller, S., Ciirdaeva, E., Kramer, W. & Fritz, H. J. (2006). The Methanothermobacter thermautotrophicus ExoIII homologue Mth212 is a DNA uridine endonuclease. *Nucleic Acids Res* **34**, 5325-36.
10. Bekesi, A., Pukancsik, M., Muha, V., Zagyva, I., Leveles, I., Hunyadi-Gulyas, E., Klement, E., Medzihradszky, K. F., Kele, Z., Erdei, A., Felfoldi, F., Konya, E. & Vertessy, B. G. (2007). A novel fruitfly protein under developmental control degrades uracil-DNA. *Biochem Biophys Res Commun* **355**, 643-8.
11. Robson, C. N., Milne, A. M., Pappin, D. J. & Hickson, I. D. (1991). Isolation of cDNA clones encoding an enzyme from bovine cells that repairs oxidative DNA damage in vitro: homology with bacterial repair enzymes. *Nucleic Acids Res* **19**, 1087-92.
12. Robson, C. N. & Hickson, I. D. (1991). Isolation of cDNA clones encoding a human apurinic/apyrimidinic endonuclease that corrects DNA repair and mutagenesis defects in E. coli xth (exonuclease III) mutants. *Nucleic Acids Res* **19**, 5519-23.
13. Robson, C. N., Hochhauser, D., Craig, R., Rack, K., Buckle, V. J. & Hickson, I. D. (1992). Structure of the human DNA repair gene HAP1 and its localisation to chromosome 14q 11.2-12. *Nucleic Acids Res* **20**, 4417-21.
14. Johnson, R. E., Torres-Ramos, C. A., Izumi, T., Mitra, S., Prakash, S. & Prakash, L. (1998). Identification of APN2, the Saccharomyces cerevisiae homolog of the major human AP endonuclease HAP1, and its role in the repair of abasic sites. *Genes Dev* **12**, 3137-43.
15. Shatilla, A., Ishchenko, A. A., Saparbaev, M. & Ramotar, D. (2005). Characterization of Caenorhabditis elegans exonuclease-3 and evidence that a Mg2+-dependent variant exhibits a distinct mode of action on damaged DNA. *Biochemistry* **44**, 12835-48.
16. Shatilla, A., Leduc, A., Yang, X. & Ramotar, D. (2005). Identification of two apurinic/apyrimidinic endonucleases from Caenorhabditis elegans by cross-species complementation. *DNA Repair (Amst)* **4**, 655-70.

17. Demple, B. & Harrison, L. (1994). Repair of oxidative damage to DNA: enzymology and biology. *Annu Rev Biochem* **63**, 915-48.
18. Pfeifer, S. & Greiner-Stoffele, T. (2005). A recombinant exonuclease III homologue of the thermophilic archaeon Methanothermobacter thermautotrophicus. *DNA Repair (Amst)* **4**, 433-44.
19. Deuschl, F., Kollmann, K., von Figura, K. & Lubke, T. (2006). Molecular characterization of the hypothetical 66.3-kDa protein in mouse: lysosomal targeting, glycosylation, processing and tissue distribution. *FEBS Lett* **580**, 5747-52.
20. Schomacher, L., Chong, J. P., McDermott, P., Kramer, W. & Fritz, H. J. (2009). DNA uracil repair initiated by the archaeal ExoIII homologue Mth212 via direct strand incision. *Nucleic Acids Res.*
21. Vidal, A. E., Harkiolaki, M., Gallego, C., Castillo-Acosta, V. M., Ruiz-Perez, L. M., Wilson, K. & Gonzalez-Pacanowska, D. (2007). Crystal structure and DNA repair activities of the AP endonuclease from Leishmania major. *J Mol Biol* **373**, 827-38.
22. Barzilay, G., Walker, L. J., Robson, C. N. & Hickson, I. D. (1995). Site-directed mutagenesis of the human DNA repair enzyme HAP1: identification of residues important for AP endonuclease and RNase H activity. *Nucleic Acids Res* **23**, 1544-50.
23. Demple, B., Herman, T. & Chen, D. S. (1991). Cloning and expression of APE, the cDNA encoding the major human apurinic endonuclease: definition of a family of DNA repair enzymes. *Proc Natl Acad Sci U S A* **88**, 11450-4.
24. Chen, D. S., Herman, T. & Demple, B. (1991). Two distinct human DNA diesterases that hydrolyze 3'-blocking deoxyribose fragments from oxidized DNA. *Nucleic Acids Res* **19**, 5907-14.
25. Xanthoudakis, S. & Curran, T. (1992). Identification and characterization of Ref-1, a nuclear protein that facilitates AP-1 DNA-binding activity. *Embo J* **11**, 653-65.
26. Walker, L. J., Craig, R. B., Harris, A. L. & Hickson, I. D. (1994). A role for the human DNA repair enzyme HAP1 in cellular protection against DNA damaging agents and hypoxic stress. *Nucleic Acids Res* **22**, 4884-9.
27. Barzilay, G. & Hickson, I. D. (1995). Structure and function of apurinic/apyrimidinic endonucleases. *Bioessays* **17**, 713-9.
28. Barzilay, G., Mol, C. D., Robson, C. N., Walker, L. J., Cunningham, R. P., Tainer, J. A. & Hickson, I. D. (1995). Identification of critical active-site residues in the multifunctional human DNA repair enzyme HAP1. *Nat Struct Biol* **2**, 561-8.
29. Erzberger, J. P. & Wilson, D. M., 3rd. (1999). The role of Mg2+ and specific amino acid residues in the catalytic reaction of the major human abasic endonuclease: new insights from EDTA-resistant incision of acyclic abasic site analogs and site-directed mutagenesis. *J Mol Biol* **290**, 447-57.
30. Wilson, D. M., 3rd, Takeshita, M. & Demple, B. (1997). Abasic site binding by the human apurinic endonuclease, Ape, and determination of the DNA contact sites. *Nucleic Acids Res* **25**, 933-9.
31. Kaneda, K., Sekiguchi, J. & Shida, T. (2006). Role of the tryptophan residue in the vicinity of the catalytic center of exonuclease III family AP endonucleases: AP site recognition mechanism. *Nucleic Acids Res* **34**, 1552-63.
32. Mol, C. D., Kuo, C. F., Thayer, M. M., Cunningham, R. P. & Tainer, J. A. (1995). Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* **374**, 381-6.
33. Shida, T., Kaneda, K., Ogawa, T. & Sekiguchi, J. (1999). Abasic site recognition mechanism by the Escherichia coli exonuclease III. *Nucleic Acids Symp Ser*, 195-6.

34. Wilson, D. M., 3rd & Barsky, D. (2001). The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA. *Mutat Res* **485**, 283-307.
35. Mol, C. D., Izumi, T., Mitra, S. & Tainer, J. A. (2000). DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected]. *Nature* **403**, 451-6.
36. Schmiedel, R., Kuettner, E. B., Keim, A., Strater, N. & Greiner-Stoffele, T. (2009). Structure and function of the abasic site specificity pocket of an AP endonuclease from Archaeoglobus fulgidus. *DNA Repair (Amst)* **8**, 219-31.
37. Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58**, 1948-54.
38. Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-97.
39. Duarte, C. M. & Pyle, A. M. (1998). Stepping through an RNA structure: A novel approach to conformational analysis. *J Mol Biol* **284**, 1465-78.
40. Gorman, M. A., Morera, S., Rothwell, D. G., de La Fortelle, E., Mol, C. D., Tainer, J. A., Hickson, I. D. & Freemont, P. S. (1997). The crystal structure of the human DNA repair endonuclease HAP1 suggests the recognition of extra-helical deoxyribose at DNA abasic sites. *Embo J* **16**, 6548-58.
41. Beernink, P. T., Segelke, B. W., Hadi, M. Z., Erzberger, J. P., Wilson, D. M., 3rd & Rupp, B. (2001). Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape1: implications for the catalytic mechanism. *J Mol Biol* **307**, 1023-34.
42. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780-1.
43. Weston, S. A., Lahm, A. & Suck, D. (1992). X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J Mol Biol* **226**, 1237-56.
44. Lahm, A. & Suck, D. (1991). DNase I-induced DNA conformation. 2 A structure of a DNase I-octamer complex. *J Mol Biol* **222**, 645-67.
45. Erzberger, J. P., Barsky, D., Scharer, O. D., Colvin, M. E. & Wilson, D. M., 3rd. (1998). Elements in abasic site recognition by the major human and Escherichia coli apurinic/apyrimidinic endonucleases. *Nucleic Acids Res* **26**, 2771-8.
46. Melo, L. F., Mundle, S. T., Fattal, M. H., O'Regan, N. E. & Strauss, P. R. (2007). Role of active site tyrosines in dynamic aspects of DNA binding by AP endonuclease. *DNA Repair (Amst)* **6**, 374-82.
47. Rothwell, D. G., Hang, B., Gorman, M. A., Freemont, P. S., Singer, B. & Hickson, I. D. (2000). Substitution of Asp-210 in HAP1 (APE/Ref-1) eliminates endonuclease activity but stabilises substrate binding. *Nucleic Acids Res* **28**, 2207-13.
48. Mundle, S. T., Fattal, M. H., Melo, L. F., Coriolan, J. D., O'Regan, N. E. & Strauss, P. R. (2004). Novel role of tyrosine in catalysis by human AP endonuclease 1. *DNA Repair (Amst)* **3**, 1447-55.
49. Casareno, R. L. B. C., J.A. (1996). Magnesium vs. manganese cofactors for metallonuclease enzymes. A critical evaluation of thermodynamic binding parameters and stoichiometry. *Chem. Commun.*, 1813-1814.
50. Wang, J., Yu, P., Lin, T. C., Konigsberg, W. H. & Steitz, T. A. (1996). Crystal structures of an NH2-terminal fragment of T4 DNA polymerase and its complexes with single-stranded DNA and with divalent metal ions. *Biochemistry* **35**, 8110-9.

51. Weiss, B. (1976). Endonuclease II of Escherichia coli is exonuclease III. *J Biol Chem* **251**, 1896-901.
52. Reha-Krantz, L. J. N., R.L. (1993). Genetic and biochemical studies of bacteriophage T4 DNA polymerase 3′–5′ exonuclease activity. *Genomics* **36**, 449-458.
53. Esteban, J. A., Soengas, M. S., Salas, M. & Blanco, L. (1994). 3'-->5' exonuclease active site of phi 29 DNA polymerase. Evidence favoring a metal ion-assisted reaction mechanism. *J Biol Chem* **269**, 31946-54.
54. Shevelev, I. V. & Hubscher, U. (2002). The 3' 5' exonucleases. *Nat Rev Mol Cell Biol* **3**, 364-76.
55. Beese, L. S., Derbyshire, V. & Steitz, T. A. (1993). Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science* **260**, 352-5.
56. Stagno, J., Aphasizheva, I., Aphasizhev, R. & Luecke, H. (2007). Dual role of the RNA substrate in selectivity and catalysis by terminal uridylyl transferases. *Proc Natl Acad Sci U S A* **104**, 14634-9.
57. Mol, C. D., Arvai, A. S., Slupphaug, G., Kavli, B., Alseth, I., Krokan, H. E. & Tainer, J. A. (1995). Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. *Cell* **80**, 869-78.
58. Parikh, S. S., Walcher, G., Jones, G. D., Slupphaug, G., Krokan, H. E., Blackburn, G. M. & Tainer, J. A. (2000). Uracil-DNA glycosylase-DNA substrate and product structures: conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. *Proc Natl Acad Sci U S A* **97**, 5083-8.
59. Werner, R. M., Jiang, Y. L., Gordley, R. G., Jagadeesh, G. J., Ladner, J. E., Xiao, G., Tordova, M., Gilliland, G. L. & Stivers, J. T. (2000). Stressing-out DNA? The contribution of serine-phosphodiester interactions in catalysis by uracil DNA glycosylase. *Biochemistry* **39**, 12585-94.
60. Singer, B. & Hang, B. (1997). What structural features determine repair enzyme specificity and mechanism in chemically modified DNA? *Chem Res Toxicol* **10**, 713-32.
61. Wong, D., DeMott, M. S. & Demple, B. (2003). Modulation of the 3'-->5'-exonuclease activity of human apurinic endonuclease (Ape1) by its 5'-incised Abasic DNA product. *J Biol Chem* **278**, 36242-9.
62. Cuniasse, P., Sowers, L. C., Eritja, R., Kaplan, B., Goodman, M. F., Cognet, J. A., LeBret, M., Guschlbauer, W. & Fazakerley, G. V. (1987). An abasic site in DNA. Solution conformation determined by proton NMR and molecular mechanics calculations. *Nucleic Acids Res* **15**, 8003-22.
63. Kalnik, M. W., Chang, C. N., Johnson, F., Grollman, A. P. & Patel, D. J. (1989). NMR studies of abasic sites in DNA duplexes: deoxyadenosine stacks into the helix opposite acyclic lesions. *Biochemistry* **28**, 3373-83.
64. Goljer, I., Kumar, S. & Bolton, P. H. (1995). Refined solution structure of a DNA heteroduplex containing an aldehydic abasic site. *J Biol Chem* **270**, 22980-7.
65. Coppel, Y., Berthet, N., Coulombeau, C., Coulombeau, C., Garcia, J. & Lhomme, J. (1997). Solution conformation of an abasic DNA undecamer duplex d(CGCACXCACGC) x d(GCGTGTGTGCG): the unpaired thymine stacks inside the helix. *Biochemistry* **36**, 4817-30.
66. Kalnik, M. W., Chang, C. N., Grollman, A. P. & Patel, D. J. (1988). NMR studies of abasic sites in DNA duplexes: deoxyadenosine stacks into the helix opposite the cyclic analogue of 2-deoxyribose. *Biochemistry* **27**, 924-31.
67. Stivers, J. T. (1998). 2-Aminopurine fluorescence studies of base stacking interactions at abasic sites in DNA: metal-ion and base sequence effects. *Nucleic Acids Res* **26**, 3837-44.

68. Wang, K. Y., Parker, S. A., Goljer, I. & Bolton, P. H. (1997). Solution structure of a duplex DNA with an abasic site in a dA tract. *Biochemistry* **36**, 11629-39.
69. Barsky, D., Foloppe, N., Ahmadia, S., Wilson, D. M., 3rd & MacKerell, A. D., Jr. (2000). New insights into the structure of abasic DNA from molecular dynamics simulations. *Nucleic Acids Res* **28**, 2613-26.
70. Kuettner, E. B., Pfeifer, S., Keim, A., Greiner-Stoffele, T. & Strater, N. (2006). Crystallization and preliminary X-ray characterization of two thermostable DNA nucleases. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **62**, 1290-3.
71. Masuda, Y., Bennett, R. A. & Demple, B. (1998). Rapid dissociation of human apurinic endonuclease (Ape1) from incised DNA induced by magnesium. *J Biol Chem* **273**, 30360-5.
72. Slupphaug, G., Mol, C. D., Kavli, B., Arvai, A. S., Krokan, H. E. & Tainer, J. A. (1996). A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. *Nature* **384**, 87-92.
73. Chen, C. Y., Mosbaugh, D. W. & Bennett, S. E. (2004). Mutational analysis of arginine 276 in the leucine-loop of human uracil-DNA glycosylase. *J Biol Chem* **279**, 48177-88.
74. Shida, T., Noda, M. & Sekiguchi, J. (1996). Cleavage of single- and double-stranded DNAs containing an abasic residue by Escherichia coli exonuclease III (AP endonuclease VI). *Nucleic Acids Res* **24**, 4572-6.
75. Hosfield, D. J., Guan, Y., Haas, B. J., Cunningham, R. P. & Tainer, J. A. (1999). Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double-nucleotide flipping at abasic sites and three-metal-ion catalysis. *Cell* **98**, 397-408.
76. Studier, F. W. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* **41**, 207-34.
77. Kabsch, W. (1993). Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795-800.
78. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**, 760-3.
79. Leslie, A. G. W. (1992). Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography.
80. French, G. S. a. W., K.S. . (1978). On the treatment of negative intensity observations. *Acta. Cryst. A* **34**, 517-525.
81. Vagin, A. A., Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30**, 1022-1025.
82. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**, 240-55.
83. Sheldrick, G. M. & Schneider, T. R. (1997). SHELXL: high-resolution refinement. *Methods Enzymol* **277**, 319-43.
84. Kleywegt, G. J. & Jones, T. A. (1996). xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr D Biol Crystallogr* **52**, 826-8.
85. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32.
86. Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). Main-chain bond lengths and bond angles in protein structures. *J Mol Biol* **231**, 1049-67.
87. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379-400.

88. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**, 2256-68.
89. Saff, E. B., and Kuijlaars, A.B.J. . (1997). Distributing many points on a sphere. *The Mathematical Intelligencer* **19**, 5-11.
90. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709-13.
91. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (1997). NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res* **25**, 4940-5.
92. Luscombe, N. M., Laskowski, R. A., Westhead, D. R., Milburn, D., Jones, S., Karmirantzou, M. & Thornton, J. M. (1998). New tools and resources for analysing protein structures and their interactions. *Acta Crystallogr D Biol Crystallogr* **54**, 1132-8.
93. Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr D Biol Crystallogr* **55**, 247-55.
94. DeLano, W. L. (2008). The PyMOL Molecular Graphics System.
95. Mol, C. D., Hosfield, D. J. & Tainer, J. A. (2000). Abasic site recognition by two apurinic/apyrimidinic endonuclease families in DNA base excision repair: the 3' ends justify the means. *Mutat Res* **460**, 211-29.
96. Suck, D., Oefner, C. & Kabsch, W. (1984). Three-dimensional structure of bovine pancreatic DNase I at 2.5 A resolution. *Embo J* **3**, 2423-30.
97. Oefner, C. & Suck, D. (1986). Crystallographic refinement and structure of DNase I at 2 A resolution. *J Mol Biol* **192**, 605-32.
98. Suck, D. & Oefner, C. (1986). Structure of DNase I at 2.0 A resolution suggests a mechanism for binding to and cutting DNA. *Nature* **321**, 620-5.
99. Weston, S. & Suck, D. (1993). X-ray structures of two single-residue mutants of DNase I: H134Q and Y76A. *Protein Eng* **6**, 349-57.
100. Weichenrieder, O., Repanas, K. & Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-86.
101. Repanas, K., Zingler, N., Layer, L. E., Schumann, G. G., Perrakis, A. & Weichenrieder, O. (2007). Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* **35**, 4914-26.
102. Rudolph, M. G., Wingren, C., Crowley, M. P., Chien, Y. H. & Wilson, I. A. (2004). Combined pseudo-merohedral twinning, non-crystallographic symmetry and pseudo-translation in a monoclinic crystal form of the gammadelta T-cell ligand T10. *Acta Crystallogr D Biol Crystallogr* **60**, 656-64.
103. Mueller, U., Muller, Y. A., Herbst-Irmer, R., Sprinzl, M. & Heinemann, U. (1999). Disorder and twin refinement of RNA heptamer double helices. *Acta Crystallogr D Biol Crystallogr* **55**, 1405-13.

Supplemental

Results

Overall structure: secondary structure elements and topology

Mth0212 is a globular α/β protein consisting of two domains with similar topologies each comprising a six-stranded β-sheet surrounded by α-helices. The following description of the overall structure is based on the high resolution structure 3G91 but is valid for all other structures as well with only negligible deviations.

Domain I comprises the residues 2-76, 197-217 and 234-257, while domain II consists of the amino acids 80–192 and 218-228. Both domains display a similar topology and together form a four-layered α/β –sandwich. One β-sheet is formed by the strands β1-β4, β11 and β12, while the other β–sheet is composed of β5-β10. The β-strands β1 (3-10), β2 (33-37), β3 (57-61), β5 (81-83), β7 (106-111), β8 (146-151), β9 (189-191) and β11 (234-239) are parallel to each other and anti-parallel to β4 (71-75), β6 (97-101), β10 (222-227) and β12 (251-255) with the topology β3-β4-β2-β1-β12-β11 and β5-β6-β7-β8-β10-β9, respectively. The overall dimensions amount to approximately 47 x 44 x 38 $Å^3$. The two six-stranded β-sheets of the two domains are flanked by two long α-helices on each side comprising the residues 13-30 (α1), 42-52 (α2), 118-142 (α4) and 174-187 (α7) as well as five shorter α-helices ranging from 88-93 (α3), 157-161 (α5), 164-168 (α6), 192-197 (α8) and 229-232 (α9), respectively.

Materials and Methods

The structure 3GA6 was refined as a twin in the space group $P2_1$ with PHENIX [37] using the twin law "l,-k,h". The analysis as well as refinement procedures are described in detail in the following. An indication of possible hemihedral twinning in the 3GA6 complex crystal was the observation that the data could be indexed, processed and scaled with comparable overall R_{sym} values in the space groups $P2_1$ and $C222_1$ as has been reported e.g. for the γδ T-cell ligand T10 [102]. The cell parameters amounted to 54.83 Å / 126.66 Å / 54.83 Å / 90.00° / 93.14° / 90.00° for $P2_1$ and 75.38 Å / 79.64 Å / 126.65 Å and all angles 90.0° for the apparent space group $C222_1$, respectively. The equal length of the axes a and c in the monoclinic space group is one of the common warning signs of twinned crystals. Calculation of the packing density based on the monoclinic data yielded two complexes in the asymmetric unit (asu) with a Matthews coefficient of about 2.5 $Å^3$/Da (55.2 % solvent content). The even number of complexes in the asu is reasonable due to the bisection of the volume in $C222_1$. Since the same solvent content was estimated in the $C222_1$ setting when one complex was assumed, packing density considerations did not provide indications of twinning. The same applies to an L-test performed with DATAMAN [84] (Fig. S1a) as well as the $|E^2$-$1|$ value of 0.784 calculated with XPREP (Bruker AXS Inc., Madison, Wisconsin, USA) (0.736 expected for an untwinned non-centrosymmetric structure and lower values for twinned data).

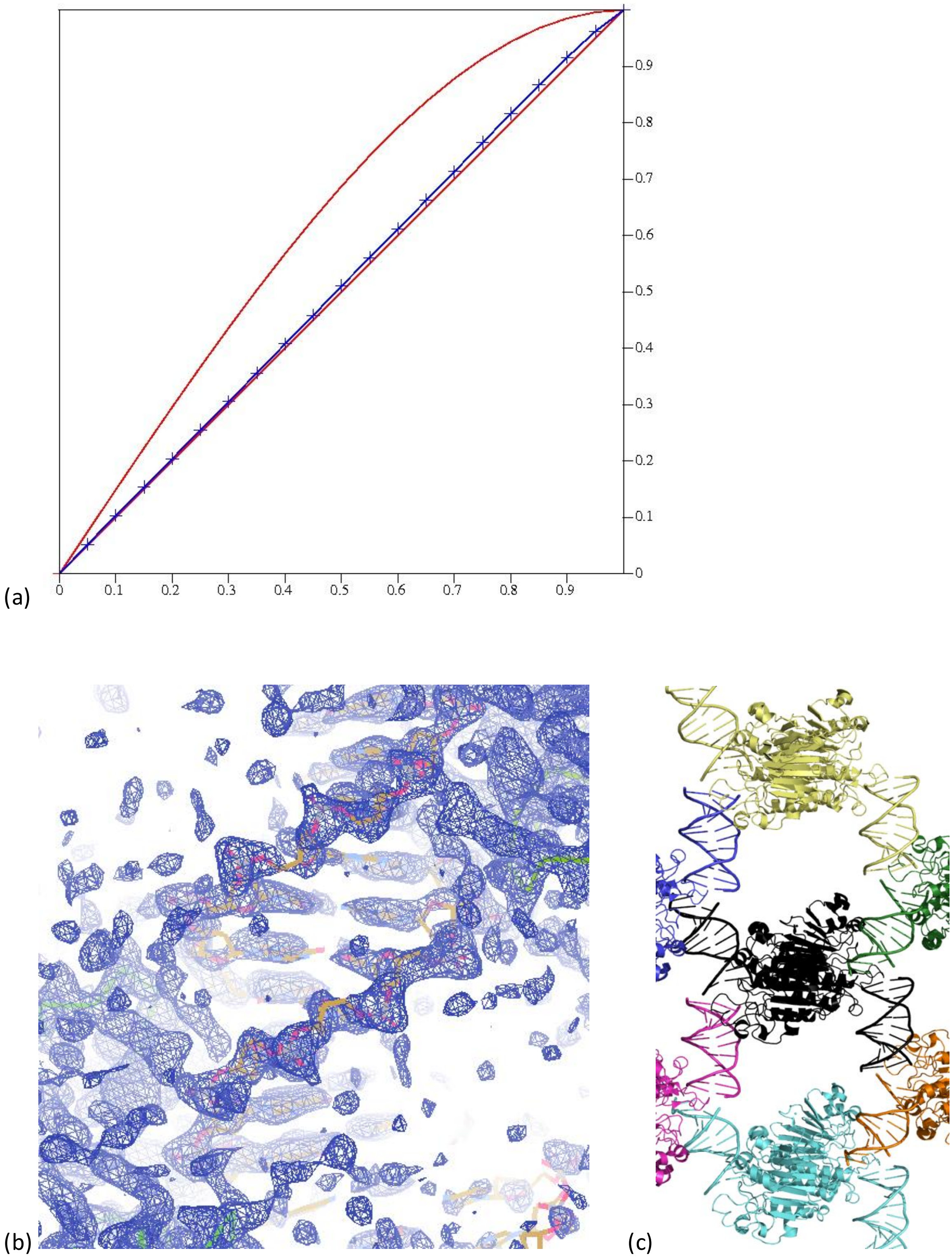
Thus, the space group ambiguity had to be resolved during structure determination. According to the common practice, first the data in the higher symmetry space group $C222_1$ were used. After MR and the first round of refinement using only the protein coordinates, clear density for the DNA duplex was observed (Fig. S1b). But obviously the density representing the dsDNA was clearly continuous over the interchange between two adjacent asymmetric units. Therefore, either the space group was incorrect or the DNA which was arranged in pseudo-continuous helices throughout the crystal was disordered. The disorder did not only concern a shift with regard to the position of the protein in the asu but additionally a rotation around the same axis since the 12bp DNA sequence was not palindromic. The absence of hints for twinning in the data statistics could be due to the low number of reflexes twinned, since the molecular differences concern only a few atoms located in the bases which differ between the two protein-DNA complexes in $P2_1$.

In analogy to similar observations for RNA heptamer double helices [103], both the refinement in $C222_1$ modelling the residues of the four DNA strands with an occupancy of 0.5 each to reflect their disorder and twin refinement in $P2_1$ applying the twin law “l,-k,h” were possible. The two strategies yielded structures of reasonable quality which did not show significant systematic differences. Nevertheless, at some nucleotide positions the electron density did not support the building of equally occupied purine and pyrimidine bases, which commonly becomes obvious in an ambiguous – since mixed – density reflecting opposing bases of the pairing DNA strands. Based on this analysis and for clarity of the resulting molecular structure which represents two different kinds of Mth0212-DNA exo-complexes, twin refinement in $P2_1$ was performed. The crystal packing in $P2_1$ is shown in Figs. S1c and S1d. Finally, the twin fraction was refined to about 0.5 indicating perfect twinning. At an intermediate stage of refinement with ShelxL [83] the structure built in $P2_1$ was analyzed for residual $C222_1$ symmetry in the model. For this purpose, structure factors (F_c) were re-calculated from the coordinates and subsequently used for space group determination with XPREP. The procedure was performed with both protein and DNA coordinates and with either exclusively the protein or the DNA coordinates, respectively. Obtained figure of merit (FOM) values were compared. As expected, the complex structures did not exhibit significant $C222_1$ symmetry anymore. In contrast, the protein molecules bore slight $C222_1$ symmetry, while the two DNA helices of the asu clearly lacked any higher orthorhombic symmetry at all.

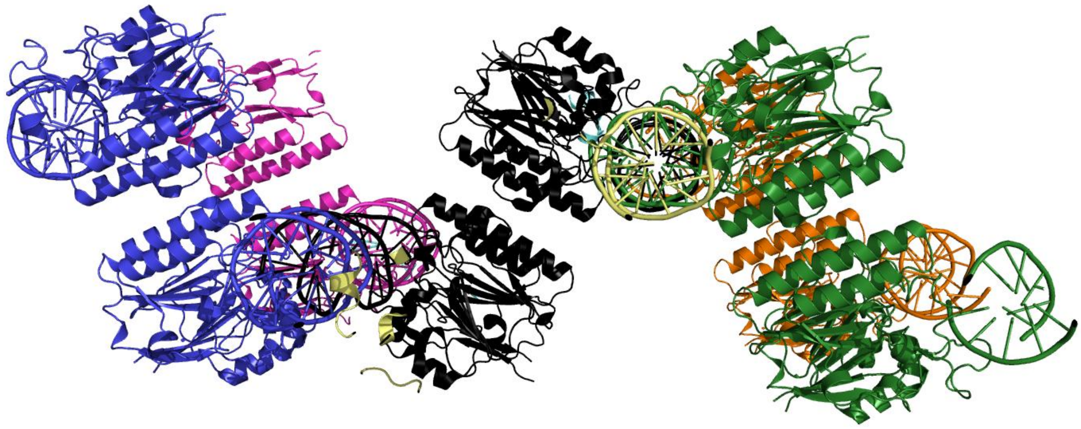
Table S1. Members of the ExoIII family of nucleases which have been studied by means of X-ray crystallography (*) as well as additional structures exhibiting a significant similarity on the level of the three-dimensional overall structure ($^{\#}$). For each homolog of which several structures are available only one representative PDB entry is listed. For abbreviations and references confer Tab. 1. [1] Results of an NCBI blast against the PDB corresponding to the aligned residues, -: no meaningful value due to the small number of aligned residues, [2] results from DALILite, ()[2] sequence identity of the residues aligned with DALILite, [2]* A Z-score ≤ 2 indicates a spurious hit.

abbreviation of the enzyme name	PDB-ID	total number of amino acids	number of aligned residues[2]	sequence identity to Mth0212 [%][1] , ()[2]	sequence similarity to Mth0212 [%][1]	Z-score[2] *	r.m.s.d. [Å][2]
ExoIII	1AKO	268	250	30 (30)	50	33.2	1.9
Ape1	1BIX	276	255	40 (41)	59	38.4	1.4
NEXO	2JC4	256	249	26 (27)	59	34.6	1.8
NAPE1	2JC5	259	256	41 (41)	45	37.6	1.5
Dr_Ape1	2O3C	282	255	41 (42)	56	39.5	1.2
LMAP	2J63	467	252	31 (35)	40	32.9	1.9
Af_Exo	2VOA	257	247	30 (30)	45	33.8	1.8
DNaseI	1DNK	260	207 (34)[1]	- (15)	-	19.5	2.7
LINE-1	2V0R	238	222	23 (23)	42	25.9	2.6

Supplementary figures



[Fig. S1, caption see below]



(d)

Fig. S1. Twin refinement for the structure 3GA6. (a) Local intensity plot (cumulative N(|L|) vs. |L| (acentrics)) as obtained by the "L-test" using DATAMAN [84]. Horizontal axis: |L|, vertical axis: N(|L|) acentrics. Theoretical values: untwinned: <|L|> = 0.500, <L^2> = 0.333 (red line); perfectly twinned: <|L|> = 0.375, <L^2> = 0.200 (red curve); values observed for the data set yielding 3GA6: <|L|> = 0.491, <L^2> = 0.322 (blue curve and points). (b) Electron density map after MR and the first round of refinement using data in $C222_1$ and only the protein coordinates. For orientation, protein and DNA residues of the final model are shown as C_α trace and as sticks, respectively. Obviously the density is continuous between the two asymmetric units indicated by two protein molecules of adjacent asymmetric units, that are represented as green C_α traces. (c, d) Crystal packing in $P2_1$ represented in cartoon mode. The two protein molecules and two DNA helices included in the PDB entry 3GA6 are coloured in black, while symmetry equivalent complexes are coloured as an entity each corresponding to the symmetry operation. The two views are rotated by about 90°.

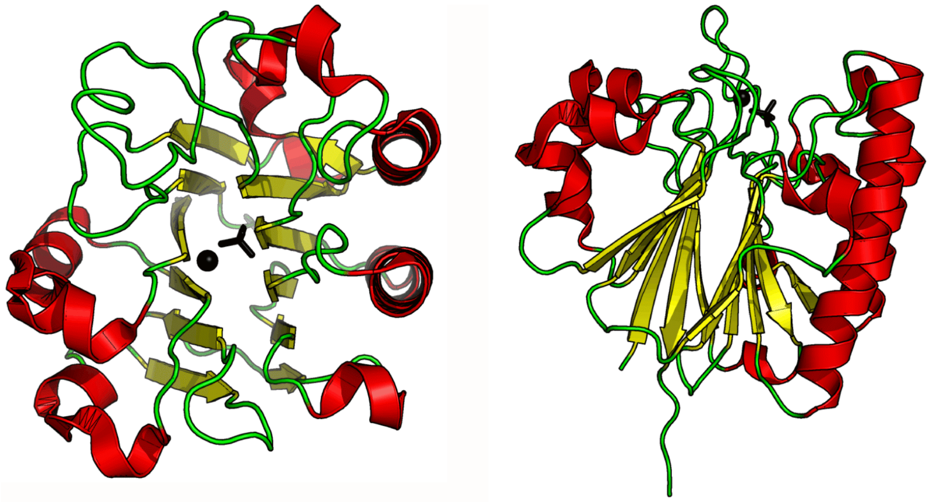


Fig. S2. Overall structure of Mth0212 (3G91). The cartoon representation is coloured according to the secondary structure: α-helices in red, β-strands in yellow, loop regions in green. The magnesium and phosphate ions are shown as black sphere and sticks, respectively. In the representations on the right and left side the view is rotated by 90°.

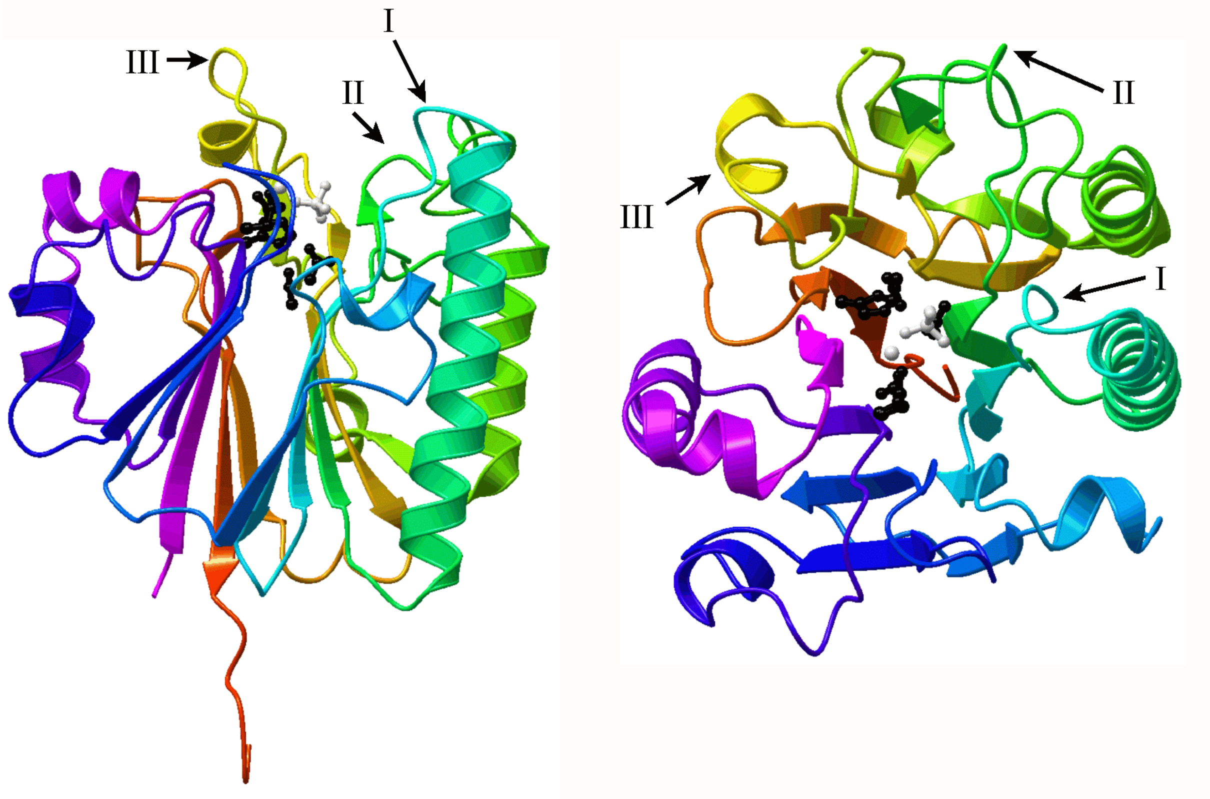


Fig. S3. Overall structure of Mth0212 (3G91). The residues are rainbow-coloured according to their position in the polypeptide chain from the N-terminus (magenta) to the C-terminus (red) and represented in cartoon mode. The side chains of the catalytically essential residues Glu38, Asp151, Asp222, His248 and the magnesium and phosphate ions bound in the active site are represented in ball and stick mode and coloured in black and light grey, respectively. The three specific DNA binding loops (I-III) are indicated by arrows.

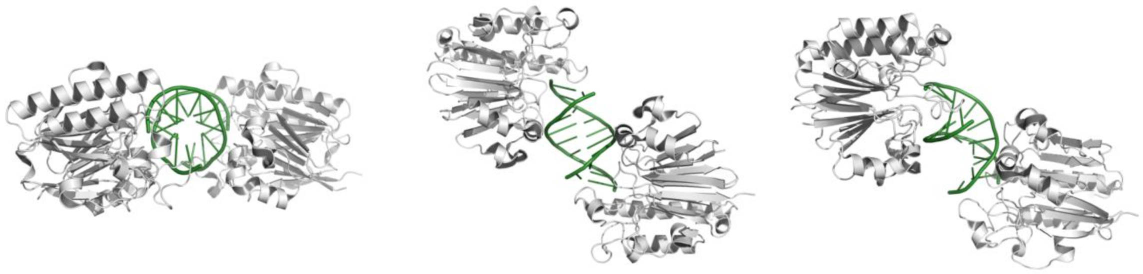

(a) (b) (c)

Fig. S4. Mth0212-DNA complex structure 3G00. The three different orientations show the distortion of the DNA. The cartoons are slightly tilted and rotated by 90° with respect to each other, respectively. In particular in (c) the significant deviation from ideal B-form DNA conformation is visible, whereas it is only slightly recognizable in (b) and not abundant in (a). The same applies to the orientation of the representation of most Mth0212-DNA complex structures. Thus, this figure additionally serves as representative for the orientations used in most figures in order to get a better impression of the protein and DNA molecules in the three-dimensional space.

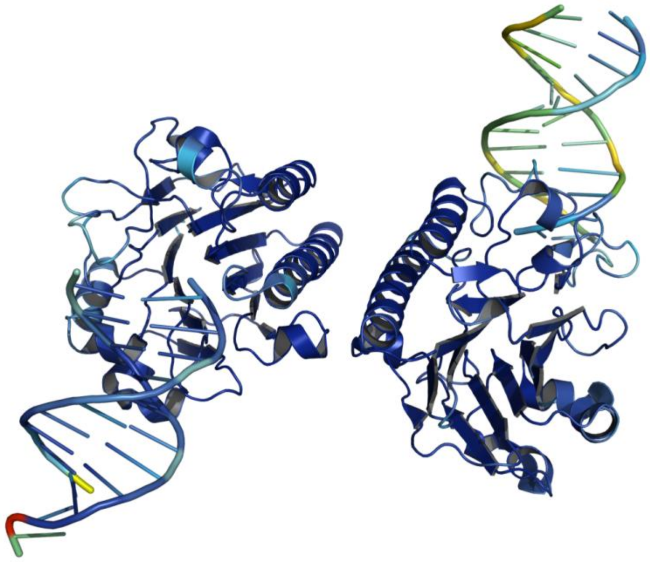


Fig. S5. Mth0212-DNA complex structure 3GA6 rainbow-coloured according to atomic B factors with an increase from blue to red. One of the two DNA duplices in the asymmetric unit exhibits a relatively high degree of flexibility compared with the total structure, while the other double helix is well defined due to a large number of crystal contacts.

11. SUPPLEMENTARY RESULTS AND DISCUSSION

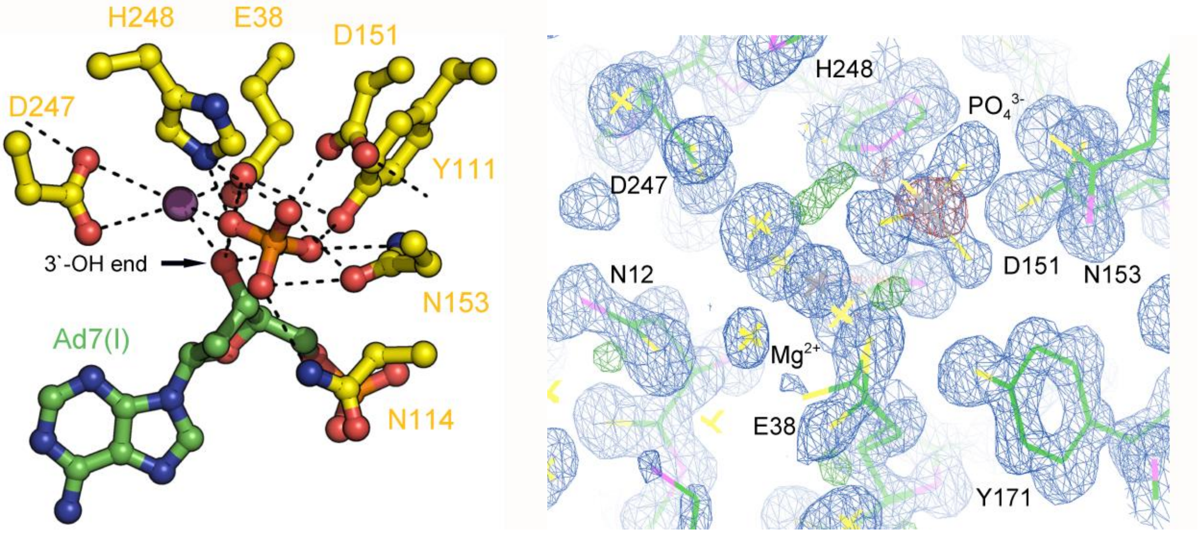
The following text and figures were not included in the submitted version of the publication due to limited space. They are not essential for the manuscript but contain complementary information which might help to visualize some of the described results in greater detail.

Right here at the very beginning of this section, the initial aim of the project should be outlined again in order to make the rationale of the various Mth0212-DNA complex structure determinations comprehensible. The most interesting feature of the DNA repair enzyme Mth0212 is its unique dU endonuclease activity. Thus, the crystallographic studies have been geared towards an Mth0212-DNA complex showing the target dU in the substrate binding pocket. However, during the course of the project, this objective turned out to be more complicated than expected. As discussed in the manuscript this was mainly due to the obvious graduation of endo- and exo-complexes with regard to their thermodynamic stability in concert with a strong binding affinity of the enzyme which also acts as a 3`-5` exonuclease to the ends of the DNA. To tackle this problem, a large number of different DNA oligonucleotides was used for co-crystallization experiments with Mth0212 as described in the following.

After obtaining the structure of Mth0212 alone the question concerning the coordination of catalytically essential divalent metal ions could not be answered unambiguously. To tackle this problem, several experiments were performed intended to make use of the anomalous scattering of manganese ions. As described in the manuscript this approach was successful using Mth0212(WT) and yielded the structure 3G0A. Additionally, experiments were performed with the mutant protein Mth0212(D151N) as follows and confirmed expectations that were based on biochemical data. By mutational studies (Georg et al., 2006) Asp151 of Mth0212 has been shown to be crucial for all nucleolytic activities. In analogy to homologous proteins, the D151N mutation has been assumed to result in the loss of metal ion binding ability. This expectation is in agreement with observations in the Mth0212(D151N) mutant structure 3G8V as outlined in the following. The crystal yielding the structure 3G8V was crystallized in the presence of $MnCl_2$ with intent to allow the binding of Mn^{2+} ions which can be localized due to their anomalous signal. As a first estimate of this feature, calculations with SHELXC (Schneider & Sheldrick, 2002) were performed. According to the program, the scaled diffraction data did not exhibit any anomalous signal. Since the absence of a significant signal could be due to the low redundancy of 3.2 an anomalous difference map was calculated and searched for a small peak caused by an only faint signal. The inspection by eye confirmed the absence of a manganese ion bound at any of the two metal coordination sites. These observations support the biochemical results mentioned above that this respective mutant (D151N of Mth0212 and equivalent mutations in other members of the ExoIII family) is not able to bind divalent metal ions which are implicated in the cleavage mechanism and therefore required for catalytic activity.

In wild-type protein, the metal ion is tightly bound. In the structure 3FZI, the Mg^{2+} ion has been derived already from the expression medium and caught throughout the purification procedure. This might apply to other structures as well. The crystals used for the respective data collections were grown either from a magnesium supplemented complex (3G1K, 3G2C) or reservoir solution (3G91) or both (3G3Y, 3G3C, 3G4T) so that the Mg^{2+} ion could also have been bound only during complex formation or crystallization. As an example, the coordination of the Mg^{2+} ion in the Mth0212-dsDNA complex structure 3G4T is shown in Fig. II-9a. Analysis of the structure reveals that the metal ions is involved in an extended hydrogen bonding network and bridging between the active side residues as well as the free 3`-hydroxyl group of the DNA backbone after incision by Mth0212. In the structure 3G4T, the terminal 3`-hydroxyl group is coordinated in the active site. This binding mode has been observed in most structures. In contrast, in the complexes 3G38, 3G0R and 3GA6 the phosphate moiety of the scissile phosphodiester bond is bound. In order to highlight these features, the respective nucleotides are represented in stick mode in Fig. II-10 regarding 3G38 and 3G0R. In 3GA6, this binding mode can be visualized only with the help of the representation of symmetry equivalent molecules, thus confer Figs. 4a, 5a and 8 of the manuscript. A phosphate ion at the active site could be unambiguously localized in the high resolution apo structure 3G91 as well (Fig. II-9b).

The structures 3G38 and 3G0R are represented in Fig. II-10 which also provides an overview of the additionally obtained Mth0212-DNA complex structures (similar to Fig. 1 of the manuscript but in an enlarged view). Fig. II-10 might benefit for the following detailed reports on several Mth0212-DNA complex structures which have been selected as supplemental to the summarized description in the manuscript.



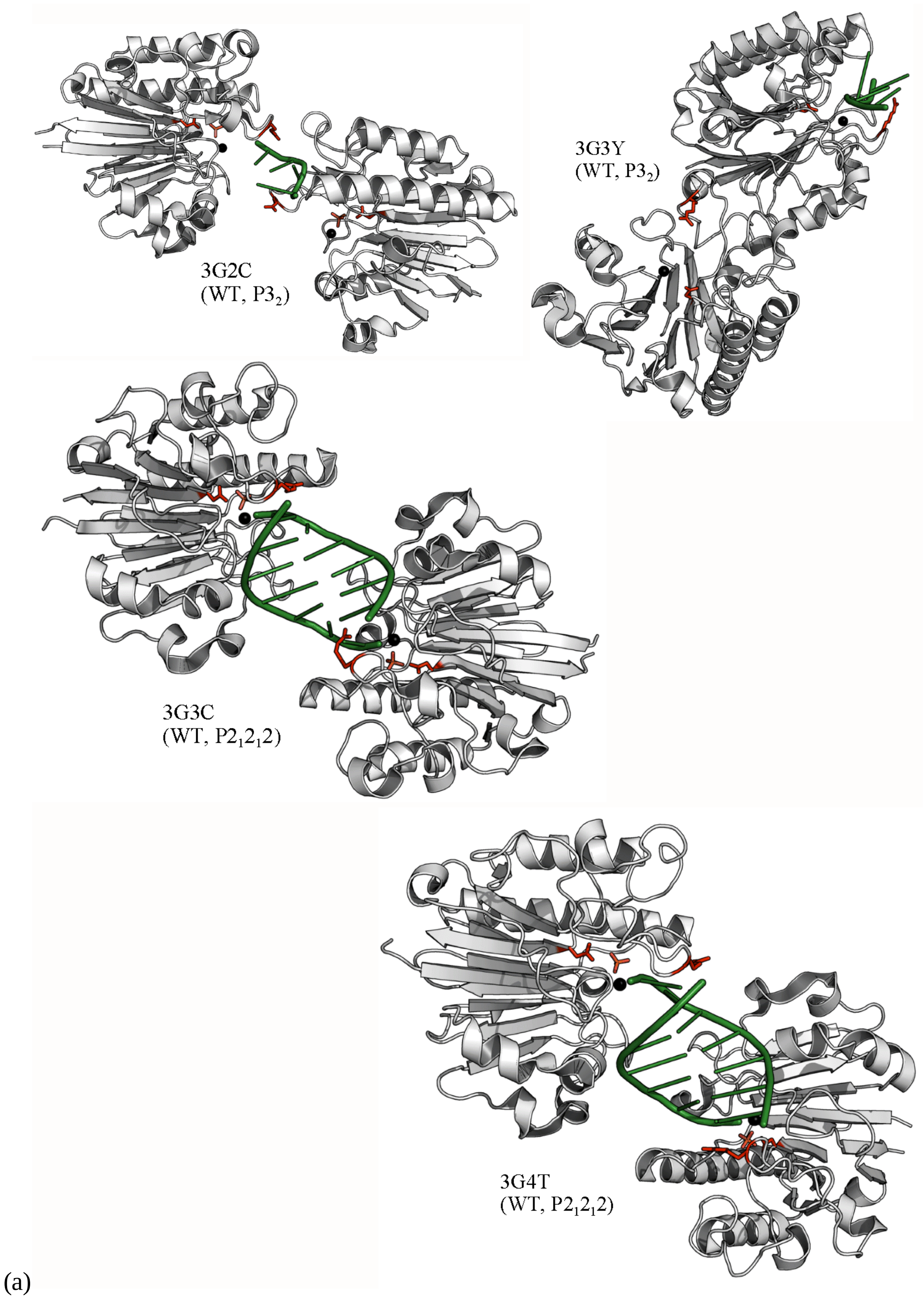
(a) (b)

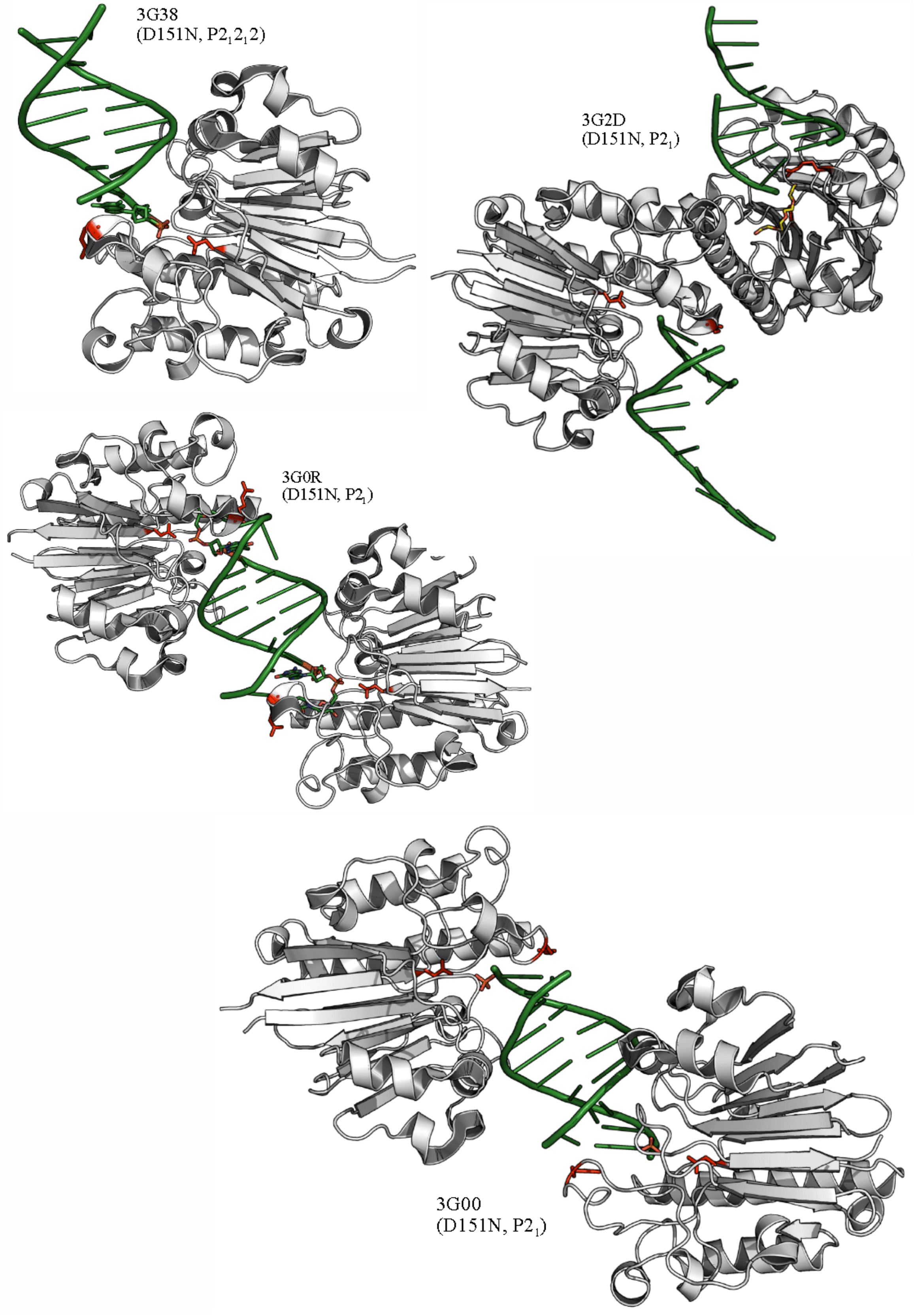
Fig. II-9. Coordination of the Mg^{2+} and phosphate ions in the active site. (a) The Mth0212-DNA complex structure 3G4T. The metal ion is represented as a green sphere, while the phosphate anion is coloured by atom and shown in stick mode. The coordinating residues (labeled in the one letter code) are represented in stick mode as well as is the 3`-terminal nucleotide (AD7(I) = adenosine 7 of chain I) which is bound to both ions via its free hydroxyl group. Putative hydrogen bonding interactions are indicated as broken lines. (b) Mth0212 apo structure 3G91. The final $2F_oF_c$ and F_oF_c electron density maps are contoured at a level of 1.5 σ and 3.5 σ and coloured in blue and red (-) / green (+), respectively. Residues are shown as thick lines and labeled according to the one letter code. The magnesium ion and the phosphate atom are coloured in grey.

3G3Y

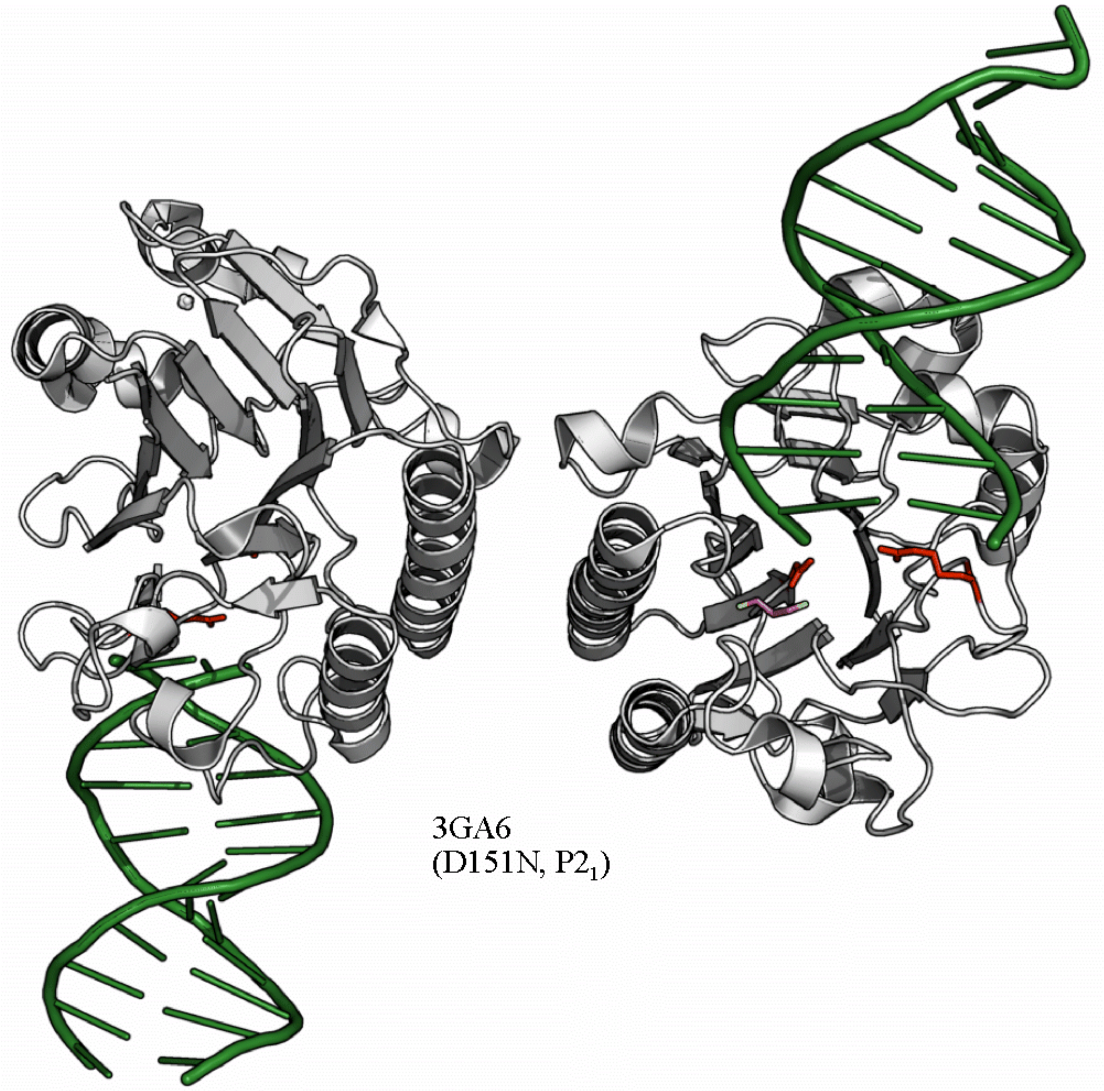
The structure 3G3Y represents Mth0212(WT) in complex with a four nucleotides long ssDNA. The DNA backbone is hold in a fixed position by hydrogen bonds with the side chains of Asn167, Ser207 and Arg215 of protein molecule B as well as with the guanidinium group of Arg209 of a symmetry equivalent molecule A`. Further hydrogen bonds are formed between the side chain of Arg209 and the deoxyribose oxygen atom O4` and the pyrimidine ring O2 atom of the 3`-terminal nucleotide, which additionally stacks with the side chain of Arg144 of a symmetry equivalent molecule A` and forms water-mediated hydrogen bonds with A` Asp101 and Asp103. An intermolecular π-stacking contact stabilizing crystal packing is also observed between the 5`-terminal nucleotide and the aromatic ring system of A` Tyr208.

Due to these interactions, only the first two nucleotides form a base stack, whereas the third nucleotide, an adenosine, is flipped out - the base in close contact (3.6 Å) to Val 170 suitable for hydrophobic interactions. The 5`-terminal nucleotide is tilted due to the interactions with two protein molecules as described above.


3G2C
(WT, $P3_2$)
3G3Y
(WT, $P3_2$)
3G3C
(WT, $P2_12_12$)
3G4T
(WT, $P2_12_12$)
(a)


3G38
(D151N, P2_12_12)
3G2D
(D151N, P2_1)
3G0R
(D151N, P2_1)
3G00
(D151N, P2_1)

(b)



(c)

Fig. II-10. Overview of the Mth0212-DNA complex structures (enlarged view representing selected structures of Fig. 1 of the manuscript). (a) Mth0212(WT) in complex with ssDNA (upper row) and dsDNA (lower row), respectively. (b) Mth0212(D151N) in complex with dsDNA. (c) The two Mth0212(D151N)-dsDNA complexes of 3GA6 differing significantly in the DNA binding mode obtained from a single crystal.

3G2C

The structure 3G2C comprises two wild-type protein molecules but only a single four nucleotides long DNA strand as 3G3Y. At the first glance, the DNA conformation in 3G2C looks similar to that in 3G3Y with one central nucleotide flipped out of the base stack. The DNA sequence could derive from both strands of the palindromic dsDNA used and thus represents the product of either the dU endonuclease activity or an exonucleolytic degradation of the dU containing or the opposite strand. The DNA strands in the structures 3G2C and

3G3Y do not only differ in their sequence (CGTA and dUTAC, respectively), but also in their interactions with Mth0212. While Asn167, Ser207 and Arg215 and the same two π-stacking contacts again fix the DNA backbone and the terminal bases, respectively, further interactions with the backbone phosphates involve the side chains of Arg163, Arg209, which contacts the sugar moiety and base in 3G3Y, Thr210 and Trp219 as well as – mediated by water - the backbone carbonyl of Val217. Interactions occur also between the free 3`-OH group and the carbonyl oxygen of a symmetry equivalent molecule A` and the side chain of A` Lys116 directly and mediated by a water molecule, respectively. In addition to the water-mediated hydrogen bonds between the side chains of A` Asp101 and Asp103 and the base of this 3`-terminal nucleotide already described for 3G3Y, the side chain of A` Ser80 and the backbone carbonyl of A` Asp103 are involved in such contacts. The phosphate ion bound in the active site is located about 2.3 Å away from the Mg^{2+} ion. It is bound by the same side chains as the phosphate anion in the structure 3G38 and additionally by Asn114 and might either derive from the DNA or from the potassium phosphate buffer, in which the Mth0212-DNA complex was formed.

3G0R

In the structure 3G0R, a flipped out cytosine close to one end of the DNA double helix results in a significant distortion of this DNA helical end. In contrast, the conformation of the last base pair at the DNA end bound to protein molecule B only slightly deviates from standard values. According to the distances, a water molecule rather than a magnesium ion is coordinated in the active site by the side chains of Glu38 and Asp247. The scissile backbone phosphate is bound in the active site of each of the two protein molecules in the asymmetric unit.

3G00

As in the structure 3G0R, the asymmetric unit of the complex structure 3G00 (Fig. II-11) contains two Mth0212(D151N) molecules but only one dsDNA molecule, which are arranged similarly to 3G0R. All nucleotides of the 9 bp palindromic blunt end dsDNA used are clearly defined in the electron density map (Fig. II-11). The conformational difference between molecules A and B of 3G00 concerning the (non-) flipped peptide bond Arg209-Thr210 might be caused by the DNA binding interactions as follows. The ends of the DNA bound by molecule A are significantly distorted, whereas the final nucleotides near molecule B only slightly deviate from ideal geometry.

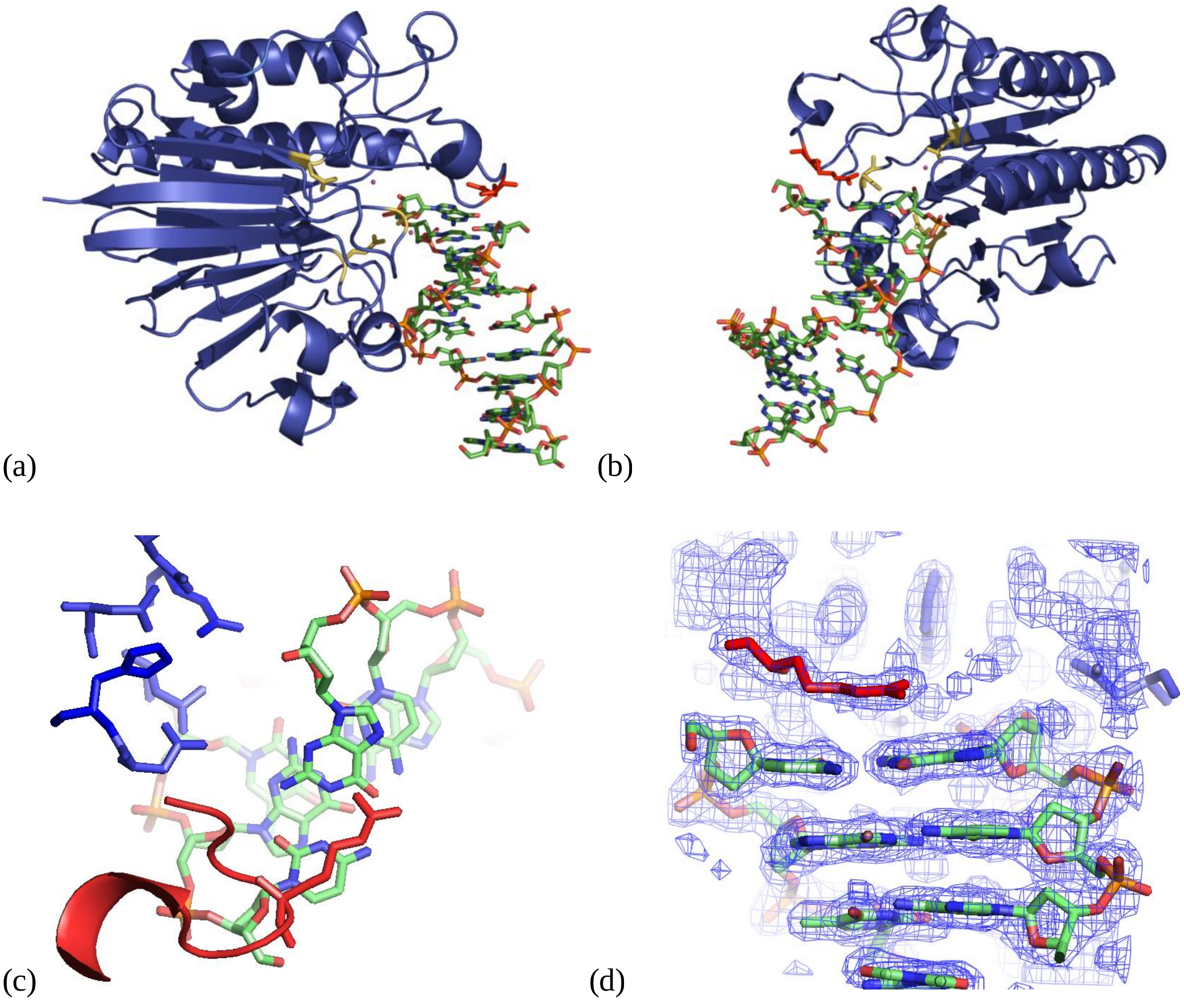


Fig. II-11. The 1.7 Å Mth0212-DNA complex structure 3G00. (a, b) Overview of the structure in two different orientations (rotated by almost 180°). For clarity, only one protein molecule of the asymmetric unit and the DNA duplex are represented in cartoon and atom-type-coloured stick mode, respectively. (c, d) Cation-π-stacking interaction between the side chain of Arg209 and the terminal base pair of the DNA duplex viewed approximately along the DNA helical axis (c) and rotated about 90° (d). The $2F_oF_c$ electron density map in (d) is contoured at a level of 1.0 σ.

3G2D

The structure 3G2D comprises two Mth0212(D151N) molecules and two DNA duplices which both exhibit one blunt and one sticky end. The double stranded region is formed between the four bases 5` of dU and the respective Watson-Crick partners of the opposite strand. The dU containing strand further proceeds into the direction of its 3`-end and includes all nine nucleotides in the DNA molecule bound at protein molecule B, while the last nucleotide could not be placed in the DNA bound to the protein molecule A. An inorganic phosphate ion is bound at the non-phosphorylated 5`-end of the dU containing DNA strand at protein molecule A. It is involved in the formation of crystal contacts due to a water-mediated hydrogen bond with the side chain of Asp92 of a symmetry equivalent protein molecule A`. In addition to the polyethylene molecules bound at the active site in molecule B, a water molecule is visible as it is in molecule A. In contrast, the magnesium lacking complex 3G2D is also devoid of a phosphate group in the active site, although phosphate anions could be placed at other positions in the solvent model and the DNA reaches the active site, but instead of a phosphate the 3`-terminal hydroxyl group is bound (in both, protein molecules A and B).

3G38

In the complex structure 3G38, Arg209 is rather oriented into the direction opposite of the bound DNA and forms a hydrogen bond with Ser207 than stacking with the terminal base pair. The other blunt end of the DNA duplex is involved in stacking interactions, namely with the DNA molecules of the adjacent asymmetric unit resulting in pseudo-continuous helices of 16 bp interrupted by breaks of about the same distance throughout the crystal. These interactions stabilize the crystal contacts significantly.

3G3C

The structure 3G3C represents the wild-type protein in complex with a 6 bp dsDNA containing a single one nucleotide long 3`-overhang. At the blunt end, the distance of 4.4 Å between the terminal base pair at one end of the DNA duplex and the guanidinium group of Arg209 which is oriented in parallel is too far for a stacking interaction.

3GA6

Almost all complex structures involve only part of the whole DNA binding area of Mth0212. On the contrary, in the structure 3GA6 additionally the second domain of Mth0212 forms contacts with DNA so that the whole protein-DNA interaction sphere could be analyzed (Fig. II-12). The structure has been obtained in almost the same crystallization condition as 3G0R except for the use of DNA of distinct length and with varied 5`- and 3`-ends. In one DNA duplex, only the terminal nucleotide at the 3`- end of one strand could not be placed in the electron density, while the other DNA double helix lacks one base pair as well as a single nucleotide at the 5`- end of one strand. In the longer DNA duplex, base pair 8 (counting started from the blunt end) as well as the last base pair before the overhang are slightly tilted and do not form classical Watson-Crick base pairs anymore. The dU endonuclease activity of Mth0212 turned out to be inhibited in the conditions of about 200-250 mM salt yielding the first Mth0212-DNA complex structures such as 3GA6 (Fig. II-13; unpublished data, Elena Ciirdaeva, laboratory of Prof. Fritz). Thus, for subsequent crystallization experiments, Mth0212 was dissolved in activity assay buffer, and specifically prepared low salt reservoir solutions were used. Nevertheless, Mth0212 was bound to the DNA ends in all yielded complex structures. Obviously, the exo-complex is thermodynamically more stable and tends more to form crystals – at least under the about 1000 tested conditions.

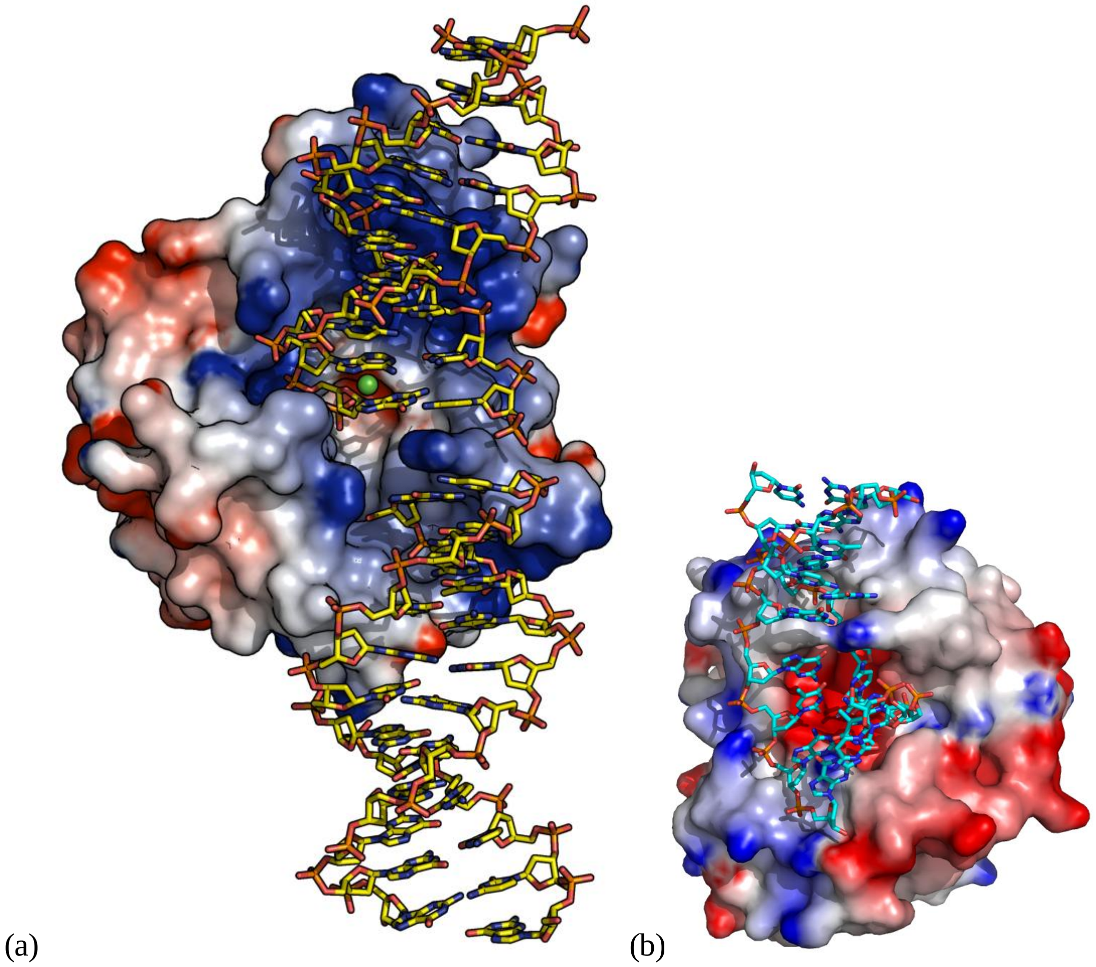


Fig. II-12. Mth0212-DNA complex 3GA6. (a) Protein molecule A coloured according to the electrostatic surface potential (calculated with DELPHI). The bound DNA helices (one duplex as included in the PDB entry and one duplex of a symmetry equivalent complex) are represented in stick mode and coloured by atom. For orientation, the magnesium ion bound in the superimposed Mth0212 apo structure 3G91 is shown as a green sphere. The side chain of Arg209 enters from the right side. (b) Ape1-DNA complex structure 1DE9 represented accordingly.

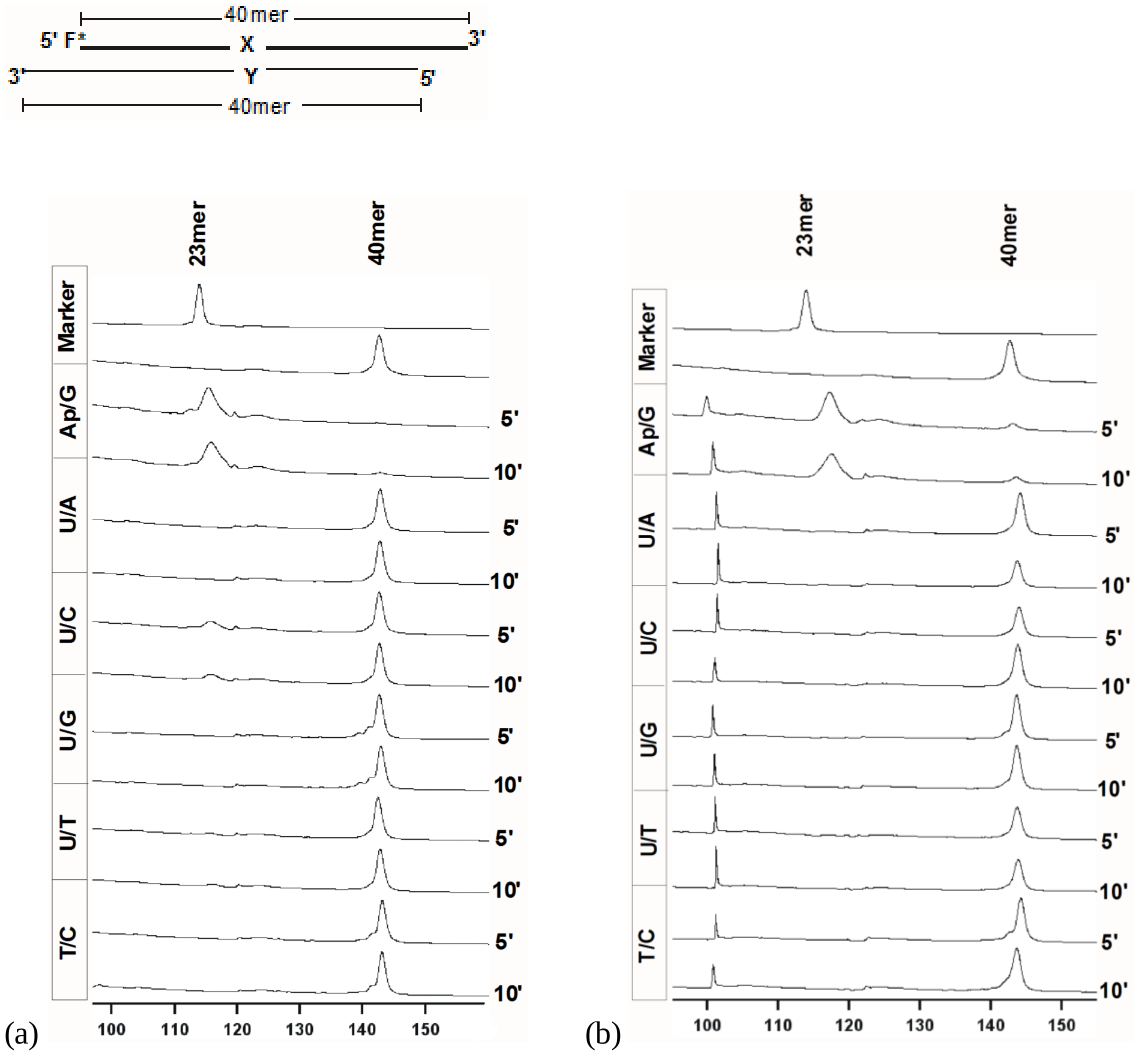
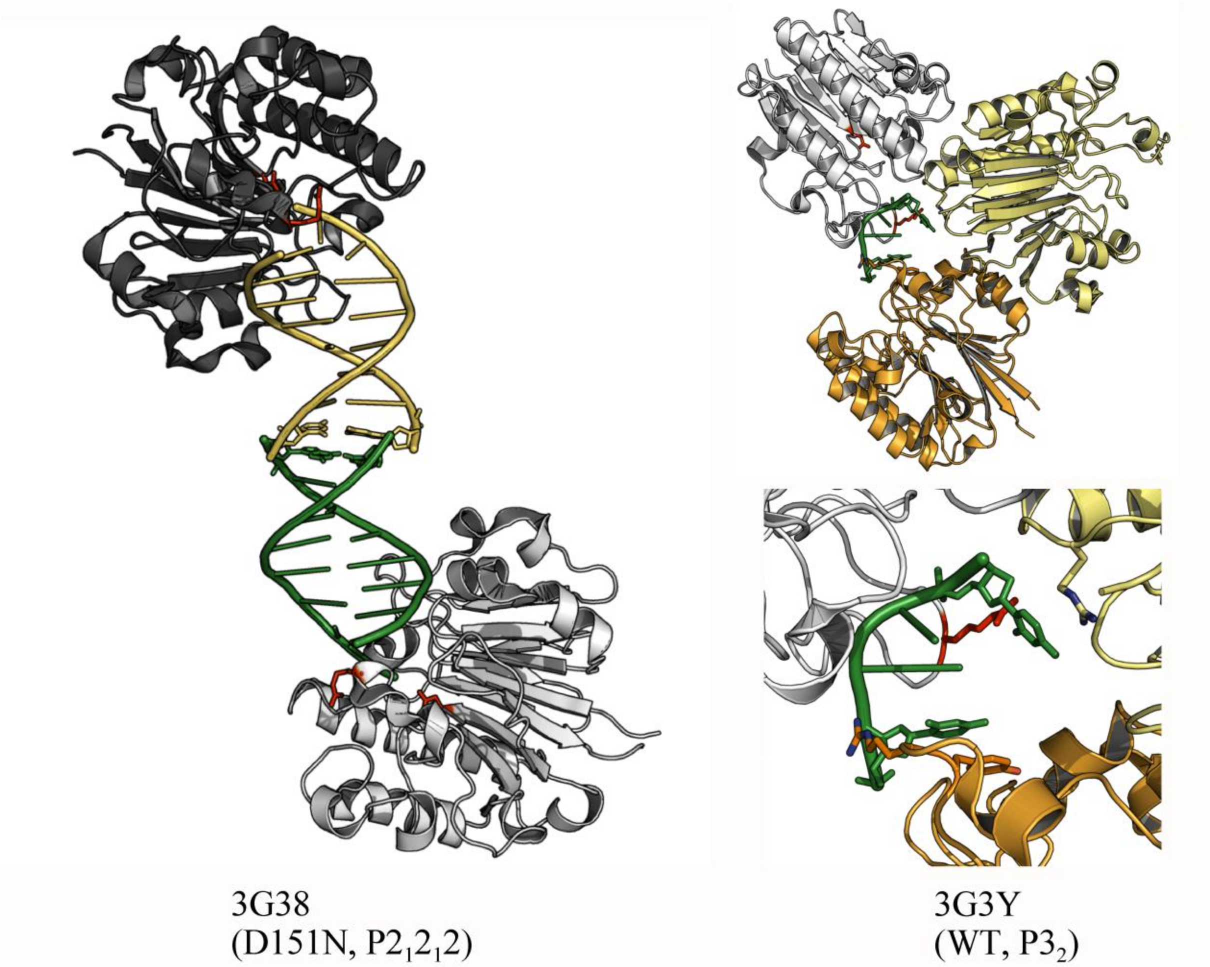


Fig. II-13. Assay of the dU endonuclease activity of Mth0212 under conditions used for protein-DNA complex formation (a) and similar to the final crystallization solution of the Mth0212-DNA complex structure 3GA6 (b), respectively. The represented results were kindly provided by Elena Ciirdaeva. In principle, the assay was performed as described in Georg et al., 2006. The substrate used is shown schematically on top of the readouts. Different central X/Y base pairs were tested as denoted for each experiment on the left side. The numbers on the right indicate the incubation time in minutes.

In most of the Mth0212-DNA complex structures, the residues of the DNA contribute significantly to crystal contact formation. Two examples are shown in Fig. II-14. In the structure 3G38, the terminal base pairs of the duplices bound by each of the two protein molecules in the asymmetric unit form stacking interactions between both base pairs. In contrast, in the Mth0212-ssDNA complex structure 3G3Y, stacking interactions are observed with the terminal base at each end and an arginine and a tyrosine side chain of two different symmetry equivalent protein molecules. Such crystal contacts have to be kept in mind during the discussion about DNA distortion since effects of the DNA conformation by crystal packing and binding to the repair enzyme cannot be distinguished unambiguously. This in particular applies to the nucleotides involved in a crystal contact itself and the residues nearby. In contrast, in regions further away from such interactions their effect can be assumed often to be negligible.



3G38
(D151N, $P2_12_12$)

3G3Y
(WT, $P3_2$)

Fig. II-14. Crystal contacts involving DNA in cartoon representation. For orientation, the side chains of Asp151 which has been varied in mutational studies as well as of the inserting Arg209 are highlighted as red sticks. Residues involved in stacking interactions are represented in stick mode.

Comparison of Mth0212 with other members of the ExoIII family

As discussed in the manuscript the major differences between the structures of Mth0212 and its human homolog Ape1 occur in the DNA binding loops. A detailed analysis of the conservation on the amino acid sequence level which has been omitted from the manuscript is given in the following, the numbering corresponds to Mth0212. In general, the amino acid sequence is well conserved among the homologs (Fig. II-15). When comparing Figs. II-12a and II-12b, the loops vary significantly in their charge which is representative for the whole nuclease domain that contains 39 arginine and lysine residues in Mth0212, whereas Ape1 and ExoIII comprise only 33 and 34 positively charged residues, respectively (Pfeifer et al., 2005). In contrast to Ape1, in Mth0212 the α-helix adjacent to the DNA binding loop II does not start before the non-DNA-contacting Ser118.

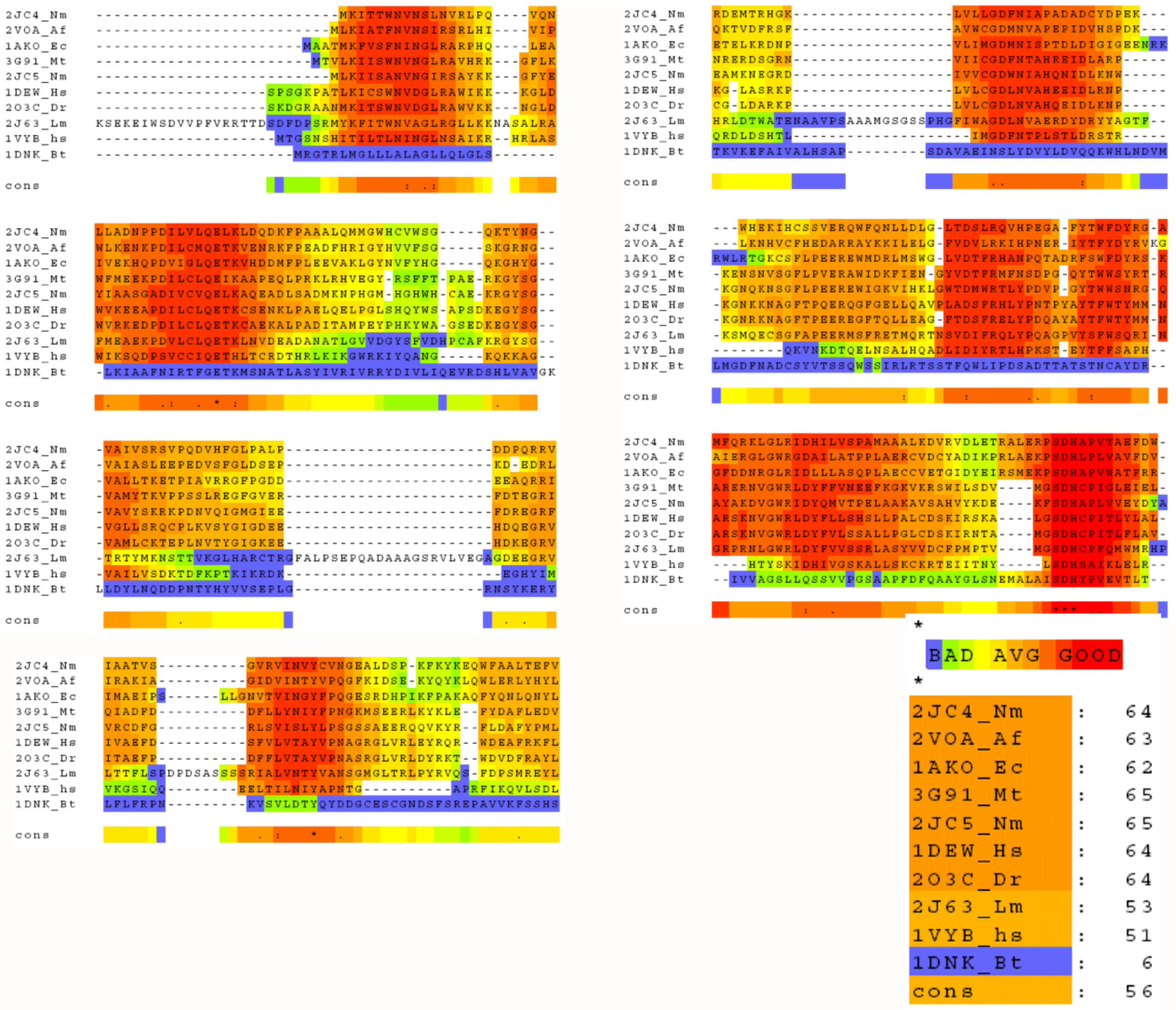


Fig. II-15. Amino acid sequence alignment of Mth0212 and its homologs with ClustalW (v. 2.04) and T-COFFEE CORE (v. 7.71). The letter/number codes of type XXXX_YY specify the respective source organism abbreviated with YY as follows and the PDB code XXXX of one appropriate crystal structure: Nm = *Neisseria meningitides*, Af = *Archaeoglobus fulgidus*, Ec = *Escherichia coli*, Mt = *Methanothermobacter thermoautotrophicus*, Hs = *Homo sapiens*, Dr = *Danio reo*, Lm = *Leishmania major*, Bt = *Bos taurus*.

In Loop I, only one of six common residues (Arg121) and additionally the first residue of loop I in Mth0212 (Asn114) are conserved, whereas three out of eleven amino acids are shared in loop III, namely Trp206, Tyr208 and Arg213. Some substitutions are functionally conservative, e.g. both Ser207 and the equivalent Thr268 of Ape1 provide a hydroxyl group for potential hydrogen bond formation. With three equal residues out of six the degree of conservation is significantly higher in loop II (Pro164, Lys165, Asn166). At the beginning of this loop, Arg163 of Mth0212 contacting a DNA backbone phosphate is replaced by an asparagine in Ape1 (Asn222). The Nδ atom of Asn222 is suitable for such interactions as well (1DE9, 1DEW, but not 1DE8) (Mol et al., 2000). Loop II belongs to the protein area which only participates in DNA binding in 3GA6.

The five sequence motives characteristic of AP endonucleases (Kaneda et al., 2006) shortly mentioned in the manuscript are also found in Mth0212 and resemble that of its homologs significantly. Thr204 of Mth0212 for example which stabilizes the Lewis acid Asp222 fits well with the equivalent residues Thr265 of Ape1 and Ser211 of ExoIII.
However, a more detailed comparison in concert with the Mth0212-DNA interactions observed in the crystal structures revealed especially one striking sequence context in motif III, namely Trp-Trp-Ser-Arg-Tyr 205-209. It is conserved only in some bacteria such as *M. thermoautotrophicus*, *Methanosarcina mazei* and *Bacillus subtilis*. The ExoIII homolog of the latter (ExoA) has been searched for a dU endonuclease activity with negative outcome (laboratory of Prof. Dr. Fritz, unpublished data). Therefore, the motif alone cannot be sufficient to extend the endonuclease substrate specificity to dU. The equivalent of the included Trp205 has been shown to be important for the recognition of an abasic site by its insertion into the AP site pocket (Kaneda et al., 2006), and Arg209 intercalates into the DNA base stack or forms cation-π-stacking interactions with the terminal DNA base (pair) in the majority of the Mth0212-DNA complexes.

For efficient binding of an AP-DNA substrate, Ape1 requires at least four base pairs on the 5`- and three base pairs on the 3` side of the abasic lesion (Wilson et al., 1995). This length and the binding position of the DNA duplex are also in good agreement with the Mth0212-DNA interactions observed in the complexes, in particular in the structure 3GA6 (Fig. II-12). A superposition of the abasic DNA contained in Ape1-DNA complex structures with the DNA bound in 3GA6 shows that an abasic site fits well in the substrate binding pocket of Mth0212 (Fig. II-16). In pure DNA, abasic deoxyribose moieties with β-conformation have been shown by NMR studies (Goljer et al., 1995) to remain intra-helical, while the α-conformer of AP sites flipped out of the base stack as in the Ape1-DNA complex.

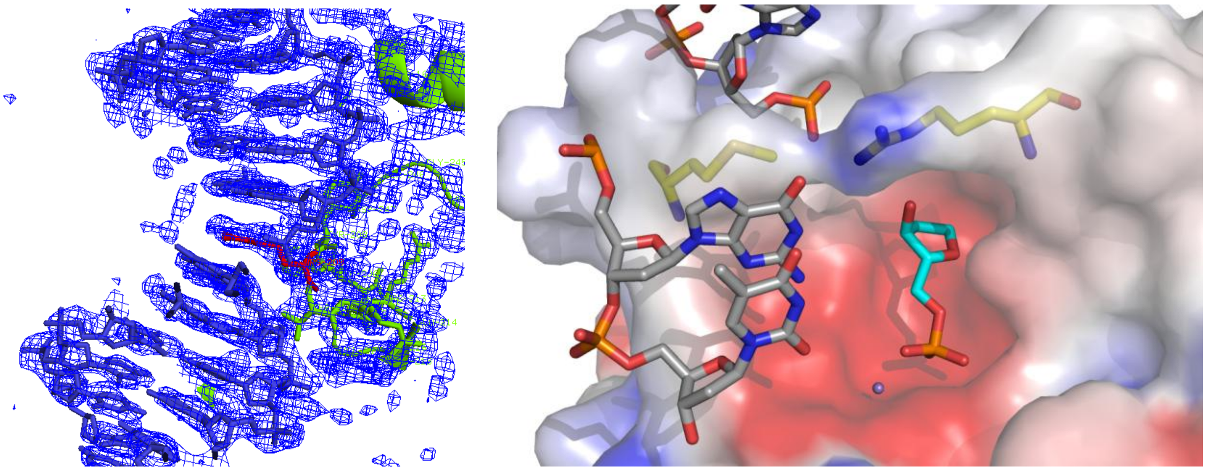

(a) (b)

Fig. II-16. Intercalation of the side chain of Arg209. (a) Electron density map of the Mth0212-DNA complex structure 3GA6. The map is contoured at the 1.0 σ level. DNA and some contacting protein residues are represented in stick mode in blue and green, respectively. Arg209 which inserts between the two duplices shown above and below the guanidinium group is highlighted in red. (b) Model of an abasic site in the substrate binding pocket of Mth0212. Some residues of the represented AP site containing DNA are shown after superposition of the DNA strands comprised in the Ape1-DNA structure 1DE9 onto the DNA bound in 3GA6. They are represented as sticks and coloured by atom with carbon atoms in grey and cyan for general nucleotides and the abasic lesion, respectively. The magnesium ion in the active site is indicated as a purple sphere. The surface of Mth0212 is coloured according to the electrostatic surface potential (negative charges in red, positive charges in blue). The intercalating residues Met117 and Arg209 of Mth0212 are represented as sticks and coloured by atom with carbon atoms in yellow). Oxygen, nitrogen and sulfur atoms are coloured in red, blue and yellow, respectively.

With regard to the ubiquitous dU repair enzymes, Mth0212 is completely unrelated to both structurally defined superfamilies of UDGs known so far on the sequence and structural level. The superfamilies consist of a single family and of four UDG families, respectively (1.: Ung homologs such as human UDG (Mol et al., 1995), human nuclear UNG2 (Krosky et al., 2006) and UDG from *E. coli* (Putnam et al., 1999); 2.: thymine DNA glycosylase (TDG) / mismatch-specific UDG (MUG) family such as MUG (Moe et al., 2006), single-strand-selective monofunctional UDGs such as SMUG1 (Wibley et al., 2003), TthUDG from *Thermus thermophilus* (Hoseki et al., 2003) and TtUDGB (Kosaka et al., 2007).

Mutational studies

Commonly, the Nε atom of Lys116 forms a hydrogen bond either to a free 5`-OH group of a terminal nucleotide (molecule A in 3GA6), water-mediated contacts to atoms of a DNA base (3G0R) or direct or water-mediated hydrogen bonds to other protein atoms (3G00, 3G3Y) or solvent molecules (3G2D) or is not fixed (3G4T, 3G38). In contrast, an interaction with the phosphate group in the active site similar to that of Arg177 is observed in molecule A of the complex 3G3C, in which the side chain of Lys116 delivers a water-mediated hydrogen bond to a phosphate anion coordinated in the active site and enters the DNA from the major groove at about the same level as Arg209 but - due to the absence of an extended helix - without insertion. The recognition of dU could also occur inside the DNA helix, and based on its localization the side chain of Lys116 is a potential determinant of the substrate specificity. Preliminary mutational studies intended to validate such a scenario indeed indicated that in the mutant Lys116Ala the dU and AP endonuclease activities of Mth0212 were decreased to a higher extent than the 3`-5` exonuclease activities. Thus, this mutant was used for crystallization experiments as well. However, in subsequent biochemical experiments, the gradual decrease in nucleolytic activities could not be confirmed.

Comparison of Mth0212 with the ExoIII homolog Mm3148 from Methanosarcina mazei

Recently, in addition to Mth0212, its homolog Mm3148 from the mesophilic archaeon *Methanosarcina mazei* has been found to exhibit a very slight dU endonuclease function as well, although this activity increases to significant, detectable levels only upon the mutation of two residues (unpublished data, Swetlana Ber, laboratory of Prof. Fritz). Therefore, this enzyme has been analyzed by means of X-ray crystallography. The apo structure of Mm3148 has been determined to a resolution of 1.4 Å with R factors of R_{work} = 14.6 % and R_{free} = 18.1 %. It is represented in Fig. II-17 in superposition with the Mth0212 apo structure 3G91. The r.m.s. deviations between Mth0212 (3G91) and Mm3148 amount to only 0.8 Å for 245 aligned residues. The catalytically essential residues fit well, only the position of the coordinated magnesium ion differs slightly. This deviation might be due to the absence of a phosphate in the active site of Mm3148, while in the Mth0212 structure 3G91 a phosphate anion is bound (Fig. II-17b).

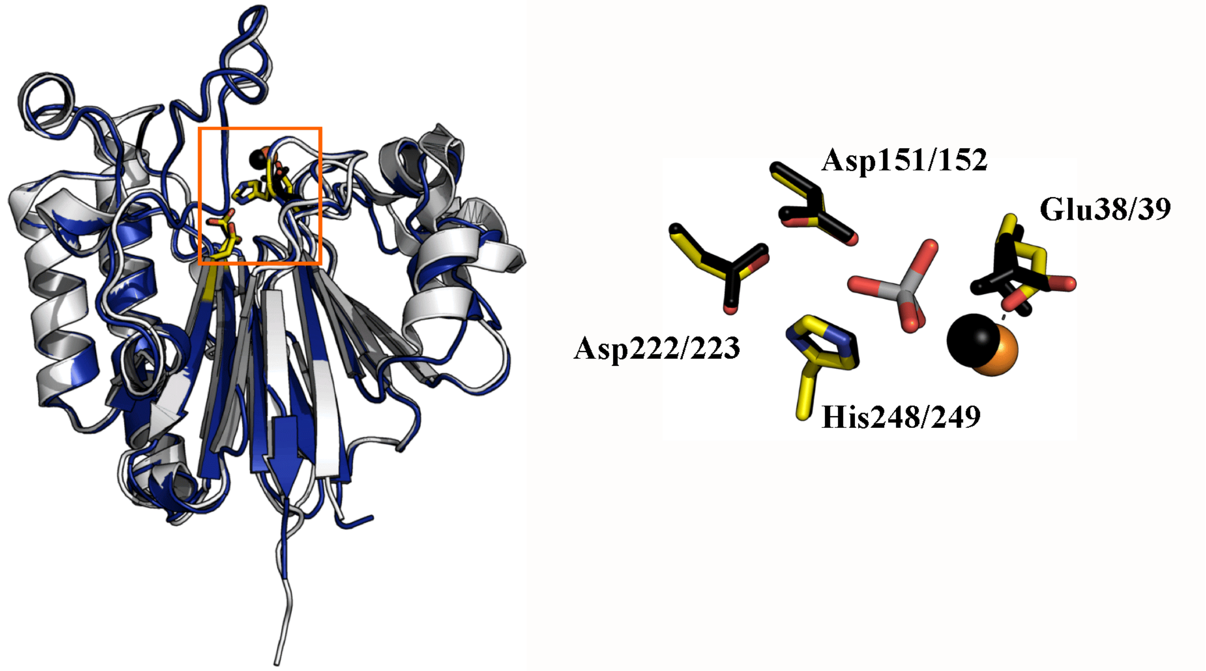


Fig. II-17. Superposition of Mth0212 (3G91) and its homolog Mm3148. (a) Overall structure in cartoon representation. 3G91 is coloured in grey, Mm3148 (rationally designed double mutant) in blue with the catalytically essential residues shown as sticks and coloured by atom (enclosed in the orange box, enlarged view in (b), carbon atoms highlighted in yellow, Mg^{2+} as black sphere). (b) Close up of the active site with the catalytically essential residues of Mth0212 (residue numbers on the left side) and the equivalent amino acids of Mm3148 (residue numbers on the right side) represented in stick mode. The phosphate anion bound in the active site is shown in stick mode and coloured in grey and red, while the magnesium ion of the structures of Mth0212 and Mm3148 is shown as black and orange sphere, respectively. The contact formed between the metal ion and Glu39 of Mm3148 is indicated by broken lines.

Due to the still ongoing mutational studies in the laboratory of Prof. Fritz, the structure has not been deposited in the PDB yet and the structural analysis of Mm3148 is not described and discussed in detail here but will follow in a currently prepared manuscript.

The supplementary data showed that despite of a considerable similarity of all Mth0212-DNA complexes with regard to the overall structure, the complexes exhibit significant differences in the respective DNA binding mode and the allocation of the active site.

The DNA repair enzyme Mth0212 has been shown to comprise the catalytic hallmarks of the ExoIII family of nucleases such as an 3`-5` exonuclease and an AP endonuclease activity and additionally to serve as a 2`-deoxyuridine endonuclease. In contrast to reports by Pfeifer and

co-workers (Pfeifer et al., 2005), no unspecific incisions in undamaged DNA were observed in more recent studies at low ionic strength (Georg et al., 2006; Schomacher et al., 2009).
This work describes structures of Mth0212 in its apo form and in complex with DNA. Although the target dU residue could not be depicted in the active site, the Mth0212-DNA structures shed light onto the dU recognition. Additionally, the coordination of catalytically essential divalent metal ions and the exonucleolytic binding mode could be visualized.
In the following, data presented in the manuscript and in particular results described in the supplemental will be discussed in more detail.

An approach to the structural analysis of Mth012 has already been made based on a 3D-model predicted with SWISS-Model (Pfeifer et al., 2005). The suggested formation of a disulfide bridge does not seem to occur. A connection between Cys35 and Cys149 which are putatively involved in the formation of an S-S bridge (corresponding to Cys93 and Cys208 in Ape1) could not be expected in the structures due to the use of *E. coli* BL21-CodonPlus(DE3)-RIL cells for protein expression. The positions of these cysteines relative to each other provide hints for the absence of a disulfide bridge even under suitable conditions. The respective side chains do not face each other in the central β-sheets, but the sulfur atoms point in opposite directions with a distance of about 3.6 Å (3G91). The occurrence as sulfhydryl groups could not be explained by the addition of DTT to the crystallization drop, since the used amount of DTT does not result in reduced conditions due to its instability at room temperature. Moreover, ExoIII from *E. coli* lacks the two cysteines already suggesting that the disulfide bridge found in Ape1 is not essential for the catalytic hallmarks of the ExoIII family. Since *M. thermoautotrophicus* lives under anaerobic conditions even the existence of a disulfide bond in the crystal structure probably would not play any important physiological role.

When comparing the various Mth0212-DNA complex structures the used type of DNA ends rather than the nucleotide sequence seems to play an important role for the formed complex and the obtained structure. Using AT-rich sequences did not result in complexes different from that obtained with GC-rich sequences. Several complexes have been crystallized in different space groups. The relative orientation of the protein and DNA molecules in the crystallization process could not be predicted.

For example, the same DNA sequence as for 3GA6 except for a 2 nt long 3`-overhang at one helical end was used for the structure 3G0R. The overhang was assumed to prevent exonucleolytic binding of Mth0212 to the nucleic acid. Crystallization in the same crystal form as in 3GA6 seemed unlikely since for 3GA6 the crystal packing seemed to be significantly stabilized by the pseudo-continuous DNA / Arg209 stacking interactions. These contacts should not be feasible anymore with the modified DNA duplex. The three nucleotides of the 13-mer missing in the structure appear to have been degraded since the final nucleotide of the respective strand is bound by Mth0212 with its 3`-hydroxyl group ruling out flexibility of the missing nucleotides.

Removal of 2`-deoxyuridine from single-stranded DNA

The removal of dU from single-stranded (ss) DNA usually is catalyzed by thymine DNA glycosylases (TDGs), SMUG1 or the methyl-CpG-binding protein 4 (Neddermann et al., 1996; Haushalter et al., 1999; Hendrich et al., 1998). In the genome of *M. thermoautotrophicus*, no homologous sequence has been detected. In combination with the double-strand (ds) specificity of Mth0212, the repair of dU residues in ssDNA thus still remains unknown. The ability of Ape1 to excise AP sites also from ssDNA even if with a significantly lower efficiency compared to dsDNA (Marenstein et al., 2004) had been unidentified for a long time despite of extensive biochemical studies. Likewise, although not revealed so far, under special conditions, Mth0212 might also be capable to remove dU residues from ssDNA.

12. CONCLUSION AND FUTURE PERSPECTIVES

Recently, the general statement that any uracil base in DNA is removed by UDG had to be partly corrected. At least in *M. thermoautotrophicus* an alternative strategy has evolved. The enzyme Mth0212 substitutes for the lacking UDG. The respective dU endonuclease function responsible for the essential removal of 2`-deoxyuridine from dsDNA is combined with an AP endonuclease and a 3`-5` exonuclease activity. In order to understand how the different nucleolytic activities of Mth0212 can be accomplished in a single active site and to get a deeper insight into the recognition of dU, structural characterization of Mth0212 was crucial.

The crystallographic analysis of Mth0212 in its apo form and in complex with different types of DNA representing substrates and products has been performed in the scope of this PhD work. The results shed light onto the mode of DNA binding in the active site. Five crystal structures of Mth0212 alone as well as nine structures in complex with different DNA substrates and products were determined to a resolution between 1.2 and 3.1 Å using wild-type or mutant proteins (K116A and D151N). Mth0212-DNA contacts across the whole interaction surface of the enzyme were revealed.

Based on detailed comparison of these structures amongst each other and with other ExoIII homologs a putative recognition mechanism for the unique dU endonuclease activity could be proposed. The observation of characteristic protein-DNA interactions such as an insertion of the arginine side chain 209 into the helical base stack provided the rationale for subsequent mutational studies (unpublished data, Elena Ciirdaeva and Swetlana Ber, laboratory of Prof. Fritz, University of Göttingen). The obtained results in activity assays confirm the structure-based expectations as follows. The side chain of arginine 209 seems not to be required for the 3`-5` exonuclease activity but to be crucial for the dU endonuclease function which is a unique feature of Mth0212. Thus, despite numerous previously performed mutational studies the Arg209Ala variant which was created based on the structural analysis of Mth0212-DNA complexes is the first mutant lacking singularly the dU endonuclease activity. Mutational studies regarding additional residues that are assumed to play an important role during the dU recognition mechanism such as Lys125 have to be carried out in order to further elucidate this unique substrate specificity of Mth0212.

13. REFERENCES

Alkema, W.B., Hensgens, C.M., Kroezinga, E.H., de Vries, E., Floris, R., van der Laan, J.M., Dijkstra, B.W., Janssen, D.B. (2000). *Protein Eng.* 13(12), 857-63.

Ames, B.N., Shigenaga, M.K., Hagen, T.M. (1993). *Proc. Natl. Acad. Sci. U S A*, 90, 7915-7922.

Aravind, L., Koonin, E.V. (2000). *Genome Biol* 1, RESEARCH0007.

Barnes, D.E., Lindahl, T. (2004). *Annu. Rev. Genet.* 38, 445-476.

Barzilay G., Hickson I.D. (1995). BioEssays 17:713-719.

Barzilay, G., Walker, L.J., Robson, C.N., Hickson, I.D. (1995). *Nucl. Acids Res.* 23, 1544-1550.

Beernink, P.T., Segelke, B.W., Hadi, M.Z., Erzberger, J.P., Wilson, D.M., 3rd, Rupp, B. (2001). *J. Mol. Biol.* 307, 1023-1034.

Bekesi, A., Pukancsik, M., Muha, V., Zagyva, I., Leveles, I., Hunyadi-Gulyas, E., Klement, E., Medzihradszky, K.F., Kele, Z., Erdei, A. (2007). *Biochem. Biophys. Res. Commun.* 355, 643-648.

Bernstein, C.B., H. (1991) Aging, Sex and DNA Repair. *Academic Press.*

Demple, B., Harrison, L. (1994). *Annu. Rev Biochem.* 63, 915-948.

Erzberger, J.P., Wilson, D.M. 3rd (1999) *J. Mol. Biol.* 290(2), 447-57.

Fondufe-Mittendorf, Y.N., Harer, C., Kramer, W., Fritz, H.J. (2002). *Nucl. Acids Res.* 30, 614-621.

Friedberg, E.C., Walker, G. C., Siede, W., Shultz, R. A., Ellenberger, T. (2006). *DNA Repair and Mutagenesis.* ASM Press, Washington.

Gerlt, J.A. (1993) Mechanistic principles of enzyme-catalyzed cleavage of phosphodiester bonds, in: Stuart, R.S.L., Linn, M., Roberts, R.J. (Eds.), Nucleases, *Cold Spring Harbor Press*, Plain view, NY, 1–34.

Georg, J., Schomacher, L., Chong, J.P., Majernik, A.I., Raabe, M., Urlaub, H., Muller, S., Ciirdaeva, E., Kramer, W., Fritz, H.J. (2006). *Nucl. Acids Res.* 34, 5325-5336.

Goljer, I., Kumar, S., Bolton, P.H. (1995). *J. Biol. Chem.* 270, 22980-22987.

Haushalter, K.A., Todd Stukenberg, M.W., Kirschner, M.W., Verdine, G.L. (1999). *Curr. Biol.* 9, 174-185.

Hendrich, B., Bird, A. (1998). *Mol. Cell Biol.* 18, 6538-6547.

Hoseki, J., Okamoto, A., Masui, R., Shibata, T., Inoue, Y., Yokoyama, S., Kuramitsu, S. (2003). *J. Mol. Biol.* 333, 515-526.

Hosfield, D.J., Guan, Y., Haas, B.J., Cunningham, R.P., Tainer, J.A. (1999). *Cell* 98, 397-408.

Kaneda, K., Sekiguchi, J., Shida, T. (2006). *Nucl. Acids Res.* 34, 1552-1563.

Kosaka, H., Hoseki, J., Nakagawa, N., Kuramitsu, S., Masui, R. (2007). *J. Mol. Biol.* 373, 839-850.

Krokan, H.E., Drablos, F., Slupphaug, G. (2002). *Oncogene* 21, 8935-8948.

Krokan, H.E., Standal, R., Slupphaug, G. (1997). *Biochem. J.* 325 (Pt 1), 1-16.

Krosky, D.J., Bianchet, M.A., Seiple, L., Chung, S., Amzel, L.M., Stivers, J.T. (2006). *Nucl. Acids Res.* 34, 5872-5879.

Levin, J.D., Demple, B. (1990). *Nucl. Acids Res.* 18, 5069-5075.

Lindahl, T. (1993). *Nature* 362, 709-715.

Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B., Sperens, B. (1977). *J. Biol Chem.* 252, 3286-3294.

Marenstein, D.R., Wilson, D.M., 3rd, Teebor, G.W. (2004). *DNA Repair* 3, 527-533.

Moe, E., Leiros, I., Smalas, A.O., McSweeney, S. (2006). *J. Biol. Chem.* 281, 569-577.

Mol, C.D., Arvai, A.S., Slupphaug, G., Kavli, B., Alseth, I., Krokan, H.E., Tainer, J.A. (1995). *Cell* 80, 869-878.

Mol, C.D., Hosfield, D.J., Tainer, J.A. (2000). *Mutat. Res* 460, 211-229.

Mol, C.D., Izumi, T., Mitra, S., Tainer, J.A. (2000). *Nature* 403, 451-456.

Neddermann, P., Gallinari, P., Lettieri, T., Schmid, D., Truong, O., Hsuan, J.J., Wiebauer, K., Jiricny, J. (1996). *J. Biol. Chem.* 271, 12767-12774.

Parikh, S.S., Walcher, G., Jones, G.D., Slupphaug, G., Krokan, H.E., Blackburn, G.M., Tainer, J.A. (2000). *Pro.c Natl. Acad. Sci. U S A* 97, 5083-5088.

Pearl, L.H. (2000). *Mutat. Res.* 460, 165–181.

Pfeifer, S., Greiner-Stöffele, T. (2005). *DNA Repair* 4, 433-444.

Putnam, C.D., Shroyer, M.J., Lundquist, A.J., Mol, C.D., Arvai, A.S., Mosbaugh, D.W., Tainer, J.A. (1999). *J. Mol. Biol.* 287, 331-346.

Schneider, T.R., Sheldrick, G.M. (2002). *Acta Cryst. D* 58, 1772-1779.

Schomacher, L., Chong, J.P., McDermott, P., Kramer, W., Fritz, H.J. (2009). *Nucl. Acids Res.*, doi:10.1093/nar/gkp102.

Seeberg, E., Eide, L., Bjoras, M. (1995). *Trends Biochem. Sci.* 20, 391–397.

Tye, B.K., Nyman, P.O., Lehman, I.R., Hochhauser, S., Weiss, B. (1977). *Proc. Natl. Acad. Sci. USA* 74, 154-157.

Vidal, A.E., Harkiolaki, M., Gallego, C., Castillo-Acosta, V.M., Ruiz-Perez, L.M., Wilson, K. Gonzalez-Pacanowska, D. (2007). *J. Mol. Biol.* 373, 827-838.

Werner, R.M., Jiang, Y.L., Gordley, R.G., Jagadeesh, G.J., Ladner, J.E., Xiao, G., Tordova, M., Gilliland, G.L., Stivers, J.T. (2000). *Biochemistry* 39, 12585-12594.

Wibley, J.E., Waters, T.R., Haushalter, K., Verdine, G.L., Pearl, L.H. (2003). *Mol. Cell* 11, 1647-1659.

Wilson, D.M., 3rd, Barsky, D. (2001). *Mutat. Res.* 485, 283-307.

Wilson, D.M., 3rd, Takeshita, M., Grollman, A.P., Demple, B. (1995). J. Biol. Chem. 270, 16002-16007.

Wist, E., Unhjem, O., Krokan, H. (1978). *Biochim. Biophys. Acta* 520, 253-270.

Yonekura, S., Nakamura, N., Yonei, S., Zhang-Akiyama, Q.M. (2009). *J. Radiat. Res. (Tokyo)* 50, 19-26.

APPENDIX 1: ABBREVIATIONS

General abbreviaitons:

aa	amino acids
Å	Ångström [1 Å = 10^{+10} m]
AP	apurinic / apyrimidinic (abasic)
BESSY	Berlin`s electron synchroton
B factor	temperature factor [Å^2] (crystallography)
bp	base pairs
C-terminal	carboxy-terminal
Da	Dalton [1 Da = 1 g/mol]
DESY	Deutsches Elektronensynchroton (German electron synchrotron)
DNA	deoxyribonucleic acid
Ds	double-stranded
DTT	dithiothreitol
dU	2`-deoxyuridine
EDTA	N, N, N´, N´ ethylene-diamine-tetraactetate
et al.	*et altera* (lat.: und andere)
F_c, F_o	calculated structure factor, observed structure factor (crystallography)
FFT	Fast Fourier Transform
HEPES	N-(2-hydroxyethyl)-piperazin-N`-2-ethansulfonat
HPLC	High Performance Liquid Chromatography
λ	wavelength [Å]
M	molar
MR	Molecular Replacement (crystallography)
MW	molecular weight [g/mol]

NCBI	National Center for Biotechnology Information
Nt	nucleotide(s)
N-terminal	amino-terminal
PDB (-ID)	Protein Database (identity code)
φ	phi
Φ (u,v,w)	Patterson function (crystallography)
ψ	pseudouridine
R factor	indicator of the quality of X-ray diffraction data and structures, respectively, ranging from 0 to 1 (lower values reflect better quality)
R_{anom}	anomalous R factor
$R_{p.i.m.}$	precision indicating merging R factor
r.m.s.d.	root mean square deviation
SDS-PAGE	sodium dodecylsulphate polyacrylamide gel electrophoresis
ss	single-stranded
S-SAD	sulfur single anomalous dispersion
Tris/HCl	tris(hydroxymethyl)-aminomethane-hydro chloride
v/v	volume per volume [%]
w/v	weight per volume [%]

Abbreviations of enzyme names and organisms:

Please confer tables 2 and S1 of the submitted manuscript (Part I: 66.3 kDa protein, pp. 67/68) as well as tables 1 and S1 of the manuscript in preparation (Part II: Mth0212, pp. 149/179), respectively.

One- and three-letter codes of amino acid residues:

A	Ala	alanine	M	Met	methionine
C	Cys	cysteine	N	Asn	asparagine
D	Asp	aspartate	P	Pro	proline
E	Glu	glutamate	Q	Gln	glutamine
F	Phe	phenylalanine	R	Arg	arginine
G	Gly	glycine	S	Ser	serine
H	His	histidine	T	Thr	threonine
I	Ile	isoleucine	V	Val	valine
K	Lys	lysine	W	Trp	tryptophan
L	Leu	leucine	Y	Tyr	tyrosine

Abbreviation of DNA residues:

abbreviation	base	nucleoside
A	adenine	2`-deoxyadenosine
C	cytosine	2`-deoxycytidine
G	guanine	2`-deoxyguanosine
T	thymine	2`-deoxythymidine
U	uracil	2`-deoxyuridine

APPENDIX 2: CURRICULUM VITAE

Name	**Kristina Lakomek**
Geburtsdatum	27.05.1981
Geburtsort	Göttingen
Staatsangehörigkeit	Deutsch

1987 – 1991	Katholische Grundschule Godehardschule, Abteilung Albrecht-von-Haller-Str., Göttingen
1991 – 1993	Katholische Orientierungsstufe Bonifatiusschule II, Göttingen
1993 – 2000	Max-Planck-Gymnasium, Göttingen
Juni 2000	Abitur
2000 – 2005	Biologiestudium (Diplom) an der Georg-August-Universität Göttingen (Immatrikulation zum WS 2000/01)
Juli 2002	Vordiplom
Juli 2004	Diplom-Prüfung
September 2004 – Juni 2005	Diplomarbeit: Titel: „Kinetische Charakterisierung der 2-Thiouridin-Synthetase MnmA und Co-Kristallisation eines MnmA-RNA-Komplexes", angefertigt in der Abteilung Molekulare Strukturbiologie am Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen unter Anleitung von Herrn Prof. Dr. Ralf Ficner
Juni 2005	Diplom in Biologie (Georg-August-Universität Göttingen)
September 2005	Beginn der experimentellen Arbeiten zur vorliegenden Dissertation in der Abteilung Molekulare Strukturbiologie am Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen unter Anleitung von Herrn Prof. Dr. Ralf Ficner

Summary of the structures deposited in the PDB

PDB-ID	**66.3 kDa protein (Part I)**	**residues (chain)**	**Ser248 – Cys249 ***
3FBX	Crystal structure of the lysosomal 66.3 kDa protein from mouse solved by S-SAD	V63-T238 (A) C249-P592 (B)	-
3FGR	Two chain form of the 66.3 kDa protein at 1.8 Angstroem	V63-T238 (A) G245-S248 (A) C249-P592 (B)	-
3FGT	Two chain form of the 66.3 kDa protein from mouse lacking the linker peptide	P60-N239 (A) C249-P592 (A)	-
3FGW	One chain form of the 66.3 kDa protein	V63-N239 (A) G245-P592 (A)	+

* I) Ser248 - Cys249: connectivity of the electron density between Ser248 and Cys249 (putative autoproteolytic cleavage site for enzyme activation).

PDB-ID	**Mth0212 (Part II)**	**Mth 0212**	**DNA**	**ligands ***
3FZI	1.9 Angstrom structure of the thermophilic exonuclease III homologue Mth0212	WT	-	Mg^{2+}
3G1K	Mth0212 (WT) crystallized in a monoclinic space group	WT	-	Mg^{2+}
3G8V	The rationally designed catalytically inactive mutant Mth0212(D151N)	D151N	-	-
3G91	1.2 Angstrom structure of the exonuclease III homologue Mth0212	K116A	-	Mg^{2+} PO_4^{3-}
3G0A	Mth0212 with two bound manganese ions	WT	-	Mn^{2+} (2)
3G2C	Mth0212 in complex with a short ssDNA (CGTA)	WT	ss	Mg^{2+}, PO_4^{3-}
3G3Y	Mth0212 in complex with ssDNA in space group P32	WT	ss	Mg^{2+}
3G2D	Complex of Mth0212 and a 4 bp dsDNA with 3`-overhang	D151N	ds	PG4 3`-OH
3G3C	Mth0212 (WT) in complex with a 6bp dsDNA containing a single one nucleotide long 3`-overhang	WT	ds	Mg^{2+}, PO_4^{3-} 3`-OH
3G4T	Mth0212 (WT) in complex with a 7bp dsDNA	WT	ds	Mg^{2+}, PO_4^{3-} 3`-OH
3G38	The catalytically inactive mutant Mth0212 (D151N) in complex with an 8 bp dsDNA	D151N	ds	3`-P
3G0R	Complex of Mth0212 and an 8bp dsDNA with distorted ends	D151N	ds	scP
3G00	Mth0212 in complex with a 9bp blunt end dsDNA at 1.7 Angstrom	D151N	ds	PO_4^{3-} 3`-OH
3GA6	Mth0212 in complex with two DNA helices	D151N	ds	3`-P, scP glyc

* II) Ligands: molecules and ions bound in the active site (for complex structures with two protein molecules in the asu: ligands coordinated in either of the two active sites of the asu). PG4 = tetraethylene glycol, 3`-OH = free terminal 3`-hydroxyl group of the DNA, 3`-P = terminal phosphate moiety at the 3`-end of DNA, scP = scissile phosphodiester bond of the DNA, glyc = glycerol.