



KJE-3900

Master's Thesis in Chemistry

Inhibition study of *Vibrio cholerae* Endonuclease I.

Annfrid Sivertsen

May, 2008

Faculty of Science
NorStruct
Department of Chemistry
University of Tromsø

KJE-3900

Master's Thesis in Chemistry

Inhibition study of *Vibrio cholerae* Endonuclease I.

Annfrid Sivertsen

May, 2008

Preface

The work of this project is carried out at NorStruct, The Norwegian Structural Biology Center, Department of Chemistry, Faculty of Science, University of Tromsø.

I would like to thank my supervisor Professor Arne Smalås for keeping a survey of the project, giving me the necessary help and feedback in the writing process, but encouraging to independent work from the start. To the guys that gave the practical tutorial from day to day, Dr Ronny Helland, Dr Bjørn Altermark and Associated Professor Bjørn Olav Brandsdal; I appreciate the time you spent explaining and discussing the upcoming issues during the project. Sharing from your experiences, most of the time far beyond what I ever asked for. I'm grateful for the various inherited scripts from Ronny, and his pedagogical skills that make the world of structure determination a lot easier to maneuver in, and to Bjørn Olav for patience in the learning phase of basic computational skills and dealing with corrupted files. Bjørn put down the hard work along with Dr Laila Niiranen by their characterization and structural determination work of the nuclease, and provided me with purified protein in abundance that gave me a head start in the project.

Thanks to all the present and former members of NorStruct contributing to the overall including working environment of the group. To the members who stepped in when things got stuck in one way or the other, either by providing an extra pair of hands, apprentice Bjarte A. Lund and Dr Solveig Karlsen, or by sharing their personal favorite solution to exploding structures. To friends for frequent coffee breaks, hiking, skiing and hunting expeditions, and in general having a huge tolerance for my ever workaholic tendencies, you're all welcome to the party!

Index

Index	I
Keywords, Definitions & Abbreviations	II
Abstract.....	V
Introduction.....	1
Aim of the study.....	23
Material & Methods.....	25
Activity measurement & IC ₅₀ determination	25
Crystallization, Data Collection & Refinements	27
Docking.....	30
Virtual Screening	37
Results.....	39
Experimental IC ₅₀ values	39
Crystallization & Data Collection.....	44
Structure determination & Refinements	46
Structure validation.....	48
Analysis of changes in the active site of VcEndA.....	53
Docking & Virtual Screening	58
Discussion.....	67
Concluding remarks	73
Further work & Development.....	75
References.....	77
List of appendixes.....	81

Keywords, Definitions & Abbreviations

- Active site - the binding and catalytic site of an enzyme, providing the structural features that recognizes the substrate and hence a competitive inhibitor.
- dsDNA - double stranded DNA
- ssDNA - single stranded DNA
- Docking - computational modeling techniques that predict the binding and conformations of small molecules (ligands) upon binding in a complex with a protein target.
- DPI - dispersion precision index, Cruickshank's DPI for coordinate error is calculated using R-factor, number of reflections, number of parameters and number of observables. Completeness of the data is also taken into account:
$$\text{DPI} = \sqrt{N_{\text{atom}} / (N_{\text{refl}} - N_{\text{param}})} R_{\text{factor}} D_{\text{max}} \text{compl}^{-1/3}$$
$$N_{\text{atom}}$$
, the number of the atoms included in the refinement, N_{refl} , the number of reflections included in the refinement, R_{factor} , the overall R-factor, D_{max} , the maximum resolution of reflections included in the refinement, compl is the completeness of the observed data.
- Fitness - evaluation of a molecular modeling docking experiment, by calculating the total interaction energy between a ligand and a protein target by a given function. An analogue to the term scoring function.
- Force field - a function expressing the energy of a system as a sum of diverse molecular mechanics terms.
- GA - Genetic Algorithm, a search algorithm used in molecular docking that applies biological expressions and basic ideas from Darwinian and Mendelian classical theories of evolution and inheritance.
- IC₅₀ - a measure of drug effectiveness, where the IC₅₀ value is the concentration of a compound that is sufficient to inhibit 50 % of enzymatic activity.
- Inhibitor - a molecule that reduces the effectiveness of a catalyst. The inhibition is classified as competitive, noncompetitive/mixed or uncompetitive, based

upon the interaction between the inhibitor and the substrate and catalyst.

In biological systems the catalyst may be an enzyme.

LGA - Lamarckian Genetic Algorithm, a search algorithm applied in molecular docking procedures. LGA search is a hybrid of a genetic algorithm search followed by a local minima energy search.

MC - Monte Carlo algorithm is a stochastic search method in molecular docking. MC makes random Cartesian moves and rejecting or accepting the results of these moves.

Pose - the orientation and conformation a small molecule obtains upon modeled binding to a protein in molecular modeling docking experiments.

Receptor - the assigned name of a binding site in a protein. The receptor may coincide with the active site but this is not necessarily so.

Rmsd - root-mean-squares deviation, a measure of average distance between the corresponding atoms in different models of proteins.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{i=N} \delta_i^2}$$

where δ the distance between N pairs of equivalent atoms.

Target - the term is assigned to the macromolecular that is the object of the study or experiment.

VcEndA - recombinant endonuclease I from *Vibrio cholerae*. Also referred to in literature as DNase, VcEndA is coded by the gene *dns*.

VS - Virtual Screening is automatic evaluating of a large library of small molecules by distinguishing between active and inactive compounds.

VsEndA - recombinant endonuclease I from *Vibrio salmonicida*.

Vvn - recombinant endonuclease I from *Vibrio vulnificus*.

2g7f - the entry code for the Mg²⁺- containing deposited structure of VcEndA in the Protein Data Bank. This structure is referred to as the deposited 2g7f structure throughout the thesis.

Abstract

The *Vibrio cholerae* bacteria resistance against introduction of new genetic material through transformation is caused mainly by a small extracellular or periplasmic endonuclease of type I, the *VcEndA* of 24.7 kDa coded by the *dns* gene, (Focareta et al, 1987; Focareta et al, 1991; Altermark et al, 2007b). The *VcEndA* homologues in other bacteria, *Serratia marcescens* (Timmins et al, 1973), *Erwinia chrysanthemi* (Moulard et al, 1993), *Aeromonas hydrophila* (Chang et al, 1992), *Vibrio vulnificus* (Wu et al, 2001) and *Vibrio salmonicida* (Altermark et al, 2007b), are identified as the main mechanisms of preventing a successful transformation in these organisms. Of a broader commercial interest is the identification of the EndoI in *Escherichia coli*, that shares 60 % sequence identity with *VcEndA* (Jekel et al, 1995). This project aims to find a lead compound for an inhibitor that is commercially exploitable, and will be applied as an additive in a transformation kit that prevents nuclease activity. An inhibitor would increase the yield in transformation procedures, and delete the step of creating endonuclease type I negative strains prior to transformation experiments. As a starting point, the Hepes molecule known to decrease the activity of *VcEndA* and the homologues *VsEndA* from *Vibrio salmonicida* (Altermark, Ph.D thesis 2006), was used as a template to find more active compounds. In this thesis I report the work and results from an *in vitro* screening of selected compounds with similar structural features as the Hepes molecule, and their activity measured by IC₅₀ values. I also report an X-ray crystallography study with both soaked and co-crystallized approaches, and observed changes in the active site of the catalytic important residues Arg99 and Glu113 upon binding of an inactive compound. Computational modeling experiments with molecular docking, and comparison of the performance of three different docking programs, GOLD, AutoDock and Glide are carried out. To find more novel active compounds, a virtual screening by the program GOLD was performed with two libraries of small molecules. By the activity measurements, three compounds with the consensus feature of an aminoethanesulfonic acid group followed by a hetero- or homo cyclohexane ring were identified. In the structures from data sets collected from soaked crystals, the inactive molecule cacodylate was found bound in the active site. Observation of a change in the conformations of

residue Arg99 and the nearby Glu113 is shown for two data sets compared to an empty site. The formation of a salt bridge between Arg99 and Glu113 shows similarity to the findings of Arg99 conformations in dsDNA-VVn complexes of the close homologous endonuclease type I, Vvn in *Vibrio vulnificus* (Wu et al, 2001; Li et al, 2003). The comparison of the molecular modeling programs GOLD, AutoDock and Glide indicate that GOLD is most suited to perform modeling experiments of the *VcEndA* system. This program is able to differentiate between active and inactive compounds upon assigning fitness scores, as well as consistently treat active compounds by assigning similar docking poses. The program AutoDock is also considered to give satisfactory docking poses, but are penalized for not consistently differentiate between active and inactive compounds when assigning fitness scores. The results from the docking experiments and the virtual screening strengthen the interpretation of the IC₅₀ values, the consensus structural features and the changes in Arg99 upon binding of the inactive compound cacodylate.

Introduction

The gram negative *Vibrio cholerae* bacteria, have through history been, and still is, a feared microorganism due to its pathogenic and contagious properties that have caused cholera pandemics affecting all continents in the world (Kaper et al, 1995). Identified as a severe threat to public health throughout at least two centuries, have lead to extensive characterization of the bacteria. The virulence of *Vibrio cholerae* is mainly caused by the secreted protein cholera enterotoxin, also referred to as CT or ctxA, but also other factors that increase the infection rate have been identified (Singh et al, 2001; Kaper et al, 1995). Infection comes as a result of ingesting contaminated food or water containing pathogenic strains of the bacteria. After passing through the acid barrier of the stomach, the bacteria colonizes the epithelium in the small intestine where the cholera enterotoxin is known to disrupt the ion transport in the intestinal epithelial cells. This causes the characteristic loss of water and electrolytes that gives the cholera diarrhea, and leads to the dehydration that in severe cases may be lethal by causing acidosis and hypovolemic shock (Kaper et al, 1995). Recommendations regarding treatment of cholera infection is given by the WHO, World Health Organization, and implies immediately replacement of lost fluid and electrolytes, and in additional medication with antibiotics (Kaper et al, 1995). The severe contagious effect is connected to poor water supplies and sanitary conditions often known in developing countries with poor infrastructure. The WHO has updated information about the cholera pandemic on its website, and the latest outbreaks are reported in Iraq, Angola, Sudan and countries in the West Africa (21.01.2008).

Vibrio cholerae is also a natural bacterial inhabitant of aquatic environments, and is associated with crustacean copepods and aquatic plants (Singh et al, 2001). The bacteria isolated from the majority of environmental samples exhibit non-pathogenic properties due to lacking the gene for cholera enterotoxin (Singh et al, 2001). The strains that are responsible for pandemics, and that express this virulence factor, are the O1 and O139 strains (suggested as a hybrid of O1 and non-O1 strains) (Kaper et al, 1995; Singh et al, 2001). Of these strains two biotypes, the Classical and the El Tor types are reported, where the main difference is that the Classical type gives a larger portion of severe

infected patients than El Tor (Kaper et al, 1995). As *Vibrio cholerae* bacteria is linked to water, and the pandemics with few exceptions have their origin from the Indian subcontinent and the Ganges delta in Bengal, it is suggested that water acts as a reservoir for the bacteria (Kaper et al, 1995).



Figure 1. Electron microscopy picture of the *Vibrio cholerae* bacteria. Copy from Wikipedia URL: http://en.wikipedia.org/wiki/Vibrio_cholerae 08.05.2008.

The *Vibrio cholerae* bacteria resist genetic manipulation by artificial transformation in laboratories, whereas conjugally transferred DNA (Hochhut et al, 2000) and DNA transferred by transduction by bacteriophages (Jiang et al, 1998, Ichige et al, 1989) can be stably maintained by *Vibrio cholerae* cells. The systems that are required for transferring DNA are not unique for *Vibrio cholerae*, but are common in *Vibrios* and other gram negative bacteria. Miller et al. (2007) have in their study detected that natural transformation in *Vibrio cholerae* between lineages of non-pathogenic strains in the same aquatic habitat is present, and also that the trend is that large DNA fragments are possible to transfer. This is interpreted as a sign of natural competence, and that whole metabolic or biosynthetic pathways may be exchanged. The genome should therefore not be thought upon as static, but changing along with the exposure to evolutionary pressure. The resistance towards artificial genetic manipulation by transformation is identified to two DNases, a DNase with a mature size about 24 kDa (*dns* gene) and a larger DNase of about 100 kDa (*xds* gene) (Focareta et al, 1987; Focareta et al, 1991). In their study Focareta et al. (1991) made single and double site-directed mutations of the *dns* and *xds* gene in the *Vibrio cholerae* genome. The DNase of 24 kDa, from now referred to as

VcEndA, was found to be the most important nuclease in preventing transformation, although the 100 kDa *xds* gene product also showed nuclease properties, but with a minor impact on the overall activity. Focareta et al. (1991) suggested that the *VcEndA in vivo* could be important in the degradation of the DNA-rich mucus in the intestine, and by making the mucus less viscous it would be easier to efficiently colonize, and at the same time provide a nitrogen and carbon source to the bacteria from the degraded DNA. An additional role of *VcEndA* is guarding the genome from foreign and possible harmful DNA. To perform transformation experiments, the gene *dns* may be mutated to make *VcEndA* deficient strains, termed *endA⁻*. In other organisms, *Vibrio vulnificus* (Wu et al, 2001), *Serratia marcescens* (Timmins et al, 1973), and *Escherichia coli* (Jekel et al, 1995), homologues to *VcEndA* are found responsible for the difficulties in introducing new genetic material by transformation methods. The EndoI of *Escherichia coli* shares 60 % sequence identity with *VcEndA* (Jekel et al, 1995). The similarity to the *E. coli* homologue indicates that endonucleases of type I is possible to be commercially exploitable, as this is one of the most used microorganisms in recombinant biotechnology. DNase activity of *VcEndA* is shown in figure 2 as dark lysate zones on a DNase test agar plate, indicating degradation of DNA in the media.

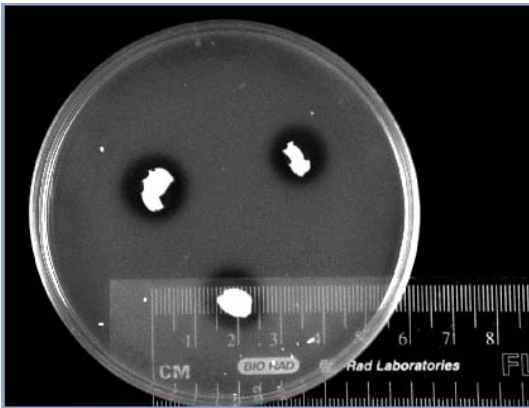


Figure 2. DNase test agar plate with colonies of *Vibrio cholerae*. The agar plate was incubated at 37 °C for 24 hours before the lysate zones were developed by adding 5 ml 1 M HCl for 2-3 minutes. The lysated zones measures four mm from the edges of the colonies. All three colonies are of the same strain.

Focareta et al. (1987) determined in their study that *VcEndA* was localized only in the periplasmic fraction when *E. coli* K-12 was used as the expression host. This is also reported for the homologue NucM from *Erwinia chrysanthemi* that shows 59 % identity

and 88 % similarity with *VcEndA*, and contains all of the catalytic important residues conserved along with the base sequence on both sides (Moulard et al, 1993). Chang et al. (1992) found in a study of *Aeromonas hydrophila* that a number of characterized proteins that were known to be secreted to extracellular space, only got exported to the periplasmic space when expressed in *E. coli*. This was also the case of the endonuclease homologue of *VcEndA* in *Aeromonas hydrophila*. The *VcEndA* homologue was detected extracellular, but to a smaller degree than expected if an efficient transport system was functional. All homologues contain an N-terminal signal sequence that varies in lengths about 20 residues and in their composition between organisms. Based on these results Chang et al. (1992) suggested that the *E. coli* bacteria have a limited capacity to transport proteins through the double membrane, and that a variation in the N-terminal signal sequences between the homologues nucleases are responsible for the difference in secretion destinations. Focareta et al. (1987) proposed that *VcEndA* require accessory genes to be transported to the extracellular space, and that these are presence in the *Vibrio cholerae*, but lacks in the *E. coli* genome. Both group's theories imply differences in secretion systems between the origin and host organisms, which is a plausible explanation of the observed localizations. Altermark et al (2007b) have biochemically and structurally compared the *VcEndA* with a close cold adapted homologue, endonuclease *VsEndA* from the fish pathogen *Vibrio salmonicida*. The results show that the enzymes are optimized to their respective natural environment, which is biochemically dissimilar due to different salinity. The optimization is shown in the differences in stability and activity, when comparing the mesophilic and psychrophilic homologue. This indicates a close contact to the environment, and may seem to favor the theory of an extracellular localization outside the bacteria itself.

The DNA-backbone is built from alternating phosphates and sugar deoxyriboses, linked by phosphoester bonds. Each phosphate is bound to two sugars, and is therefore a phosphodiester. The chemical properties of the phosphate make it suitable as the linking unit group, and responsible for its important role in other parts of biochemistry (Westheimer, 1987). Phosphoric acid, H_3PO_4 , and its ionization constants pK_{a1} 2.15, pK_{a2} 7.21 and pK_{a3} 12.36, allow for a negatively charged backbone at physiological pH,

even when linked to two deoxy-riboses at the same time. This negative charge protects the diester bond from undergoing S_N reactions, by repelling negative charges but also neutral nucleophiles, as the charge do not allow any electron pair to approach (Westheimer, 1987). Phosphate dianions are also poor leaving groups when S_N1 , S_N2 and eliminations reaction are taken into concern. Spontaneous hydrolysis rates are therefore in a negligible order. The stabilizing effect of the phosphate makes it perfect for the use as linkage group between the nucleosides that hold the genetic information. The importance of having stable genetic material is not difficult to argue, as the opposite will be a great misfortune in the ever *survival of the fittest* concerning all organisms. The stability of the DNA backbone does not however prevent enzymatic cleavage of the same diester bond.

Enzymes are biological catalysts that decrease the activation energy for a reaction by interacting with the substrate. This is the general basis for physiological reactions, where the production rates are speeded up to utilize energetically stable compounds as substrate at temperatures that are not harmful for the organism itself. An enzyme may be active toward a group of substrate, or selective toward just one specific compound. Some are even selective toward one specific stereo-isomer of the substrate. The activity of enzymes may be decreased by interaction with other compounds than the natural substrate. An interaction of this type is termed inhibition, and is classified by the different scenario type the alternative compound may bind to the enzyme (Helbæk, 1999). In a competitive inhibition does the alternative to the substrate, the inhibitor, interacting with the enzyme in a similar way as the substrate. The interactions occur in the active site, and the degree of inhibition may be decreased by increasing the substrate concentration. Interaction in another area of the enzyme may be possible and still decrease the activity. Substrate is then able to bind in the active site, but the enzyme will not produce the products as a result of the changes the inhibitor has introduced to the system. This inhibition type is termed noncompetitive or mixed. In noncompetitive inhibition, the decrease in activity is not dependent of the substrate concentration. A third inhibition interaction is possible upon interaction outside the active site, but instead of interacting only with the enzyme,

the inhibitor binds to the enzyme-substrate complex. This is termed uncompetitive inhibition, and the inhibition effect is proportional to the substrate concentration.

Proteins that interact with DNA do so in order to support their function. A rough survey of protein-DNA interactions, characterizes the proteins in three classes with family subgroups (Luscombe et al, 2002). The classes are defined as the non-specific class where the binding is independent on the DNA base sequence, the highly specific class where all members of a family target the same base sequence, and the multi-specific class where the binding is specific, but individual members of a family target different base sequences. The two binding modes in relation to the DNA molecule are interactions with the DNA backbone or base-interaction contacts. The non-specific class has, not surprisingly, a higher backbone to base interaction ratio than the two sequence specific classes, but has also highly conserved residues that are base interacting. Non-specific DNA binding proteins bind without exceptions to the minor groove of the DNA molecule, where the bases have similar van der Waals surfaces. The base-interactions are thought to be important in stabilizing the deformed structures of the DNA, as a widened minor groove when interacting to members of this non-specific class, is observed. The backbone interactions have in general a stabilizing effect on the complexes of all the three classes, and the residues are overall well conserved and located in the DNA-binding motif of each family (Luscombe et al, 2002).

Endonucleases belong to the enzyme family of nucleases, which catalyze the cleavage of the phosphoester bond of nucleic acid molecules by hydrolysis (Mishra, 2002). As they cleave DNA and RNA, they hold a biological important position in the metabolism of nucleic acids and genetic maintenance in organisms. They are also widely used tools in recombination processes in modern biotechnology. The roles of nucleases may be as care-keepers of the organisms own genetic material by removal and repair of damages, a part of the organism defense system against alien DNA or RNA, or by being a part of the apoptosis process, to mention a few (Mishra et al, 2002). A hierarchical classification system based on some consensus criteria is made to keep track of the nucleases. These criteria are the substrate types DNA or RNA, the specificity of nucleolytic attack,

applying the prefix endo- if internal or exo- if the cleavage is at the terminal of the sequence, the nature of the products by the two possible terminations at the 3' or 5' phosphate group, and further by the nature of the hydrolyzed bonds. Additional criteria may be the nature of the DNA-substrate by mismatch of basepairs, damaged or topological DNA, site specificity, structure selectivity, or functional ability to reconstruct DNA molecules (Mishra, 2002). Not all nucleases fit neatly to the consensus classification system, but by using the terms given above the terminology of nucleases are much easier to work with.

The chemical elements needed for an efficient enzymatic catalyzation of a hydrolyzation of the phospho-diester bond, are in general a nucleophilic group that the phosphoryl group can be transferred to, a basic element that is able to activate and position the nucleophile, a general acid that can protonate the leaving group, and one or more positively charged groups that can stabilize the phosphoanion intermediate state (Galburt et al, 2002). Endonucleases are known to have one or more bound metals in their active site, and by directly coordinate water molecules they lower the pK_a 's by acting as a Lewis acid. The water molecules may act as a nucleophile or a general base if left as a hydroxide ion, or if left acidic they may behave as a proton source to the leaving group. If the metal ion is divalent, it will be able to stabilize the -2 transition state of the phosphoanion (Galburt et al, 2002).

The characterization and cloning of *VcEndA* was first described by Focerata et al. (1987), but the nuclease was not structural determined until twenty years later by Altermark et al. (2006b). There are two deposited structures in the Brookhaven Protein Data Bank of *VcEndA*, both solved by X-ray crystallography. The highest resolution structure (PDB entry 2g7e), with a resolution of 1.6 Å, is crystallized at low-pH, and the second deposited structure with a resolution of 1.95 Å (PDB entry 2g7f), crystallized at neutral pH. The major difference between these structures is the lack of the catalytic Mg^{2+} in the active site of the low-pH form, whereas presence in the neutral pH-form. A buried chloride ion approximately 7 Å from the nearest solvent molecule is identified in the structure of *VcEndA* (Altermark et al, 2006b). This chloride is regarded as a structural or

stabilizing ion, and is not a catalytic part of the structure. In the characterization of the *VcEndA* the optimum conditions for catalytic activity is found to be 175 mM NaCl at pH 7.5-8.0, and at a temperature of 50 °C (Altermark et al. 2007b). The calculated molecular mass without the N-terminal signal is 24.7 kDa, and is verified by SDS/PAGE. Altermark et al. (2007b) also tested substrate specificity, and found that *VcEndA* has a very low RNase activity compared to DNase activity at the reported optimum conditions, but were able to efficiently cleave plasmid DNA, dsDNA and ssDNA. They concluded upon these results that DNA is the natural substrate, and *VcEndA* therefore is a DNase at its physiological condition. The product of the catalytic cleavage is a 5' phosphate group and a 3' oxygen leaving group. By chance in the same study it was discovered that the buffer compound Hepes decreased the activity of both *VcEndA* and *VsEndA* (Altermark, Ph.D thesis 2006).

The characterization of endonuclease Vvn, a close homologue of *VcEndA* from *Vibrio vulnificus*, shows that they share 74 % identity and that all catalytic important residues are conserved (Wu et al, 2001; Li et al, 2003). The Vvn's structure is deposited both in its native form (Li et al, 2003, PDB entry 1ouo) and in two complexes with dsDNA of two different lengths as a His80Ala mutant. One complex with a 8 basepair long dsDNA substrate (Li et al, 2003, PDB entry 1oup) and one complex with a 16 basepair dsDNA substrate (Wang et al, 2007, PDB entry 2ivk)). As Li et al. (2003) investigated the features of the structure, they found that the Vvn structure had a novel V shaped fold. The catalytic active site contained a $\beta\beta\alpha$ -metal motif that was similar to the described active site in phage T4 Endo VII endonuclease (Raaijmakers et al, 1999). This motif is also observed in the H-N-H family and His-Cys box family of endonucleases (Galburt et al, 2002). The $\beta\beta\alpha$ -metal motif in Vvn contains a divalent positively charged metal ion, a magnesium ion in 1ouo and a calcium ion in 1oup, coordinated by the residues Glu79 and Asn127. By additionally coordinating four water molecules, the $\beta\beta\alpha$ -metal motif forms an octahedral geometry (Li et al, 2003). The Vvn binds non-specificly to the minor groove of the DNA by bending the backbone 20-40° dependent of the base stacking. The minor groove is widened, and one of the DNA backbones is inserted in the V shaped cleft of the Vvn (Li et al, 2003, Wang et al, 2007). In the two dsDNA-Vvn structures both a

Vvn-substrate and a Vvn-product is found, mimicing the before- and after cleavage state. Based on these data, Li et al. (2003) have suggested a mechanism for the hydrolysis of the scissile phosphodiester bond, see figure 3. In the proposed mechanism, His80 acts as a general base that activates the water molecule W1 into becoming an attacking nucleophile. The backbone carbonyl oxygen of Glu113 strengthens the basic property of His80 by forming a hydrogen bond to Nε. The activated W1 attacks the phosphodiester bond by an in-line substitution reaction, kicking the 3' deoxyribose out as a leaving group after protonating the OH group with an acidic hydrogen from the acidic Mg²⁺-coordinated water molecule W2. To support the negatively charged intermediate state, residue Arg99 changes conformation by stretching down and stabilize the phosphate anion directly coordinated to the catalytic Mg²⁺. Now the phosphate is neutralized by its poor leaving group quality, and the 3'-deoxyribose part is free to wander off. The mechanism explains the product fragments containing 3'-OH and 5'-deoxyribose bound phosphates.

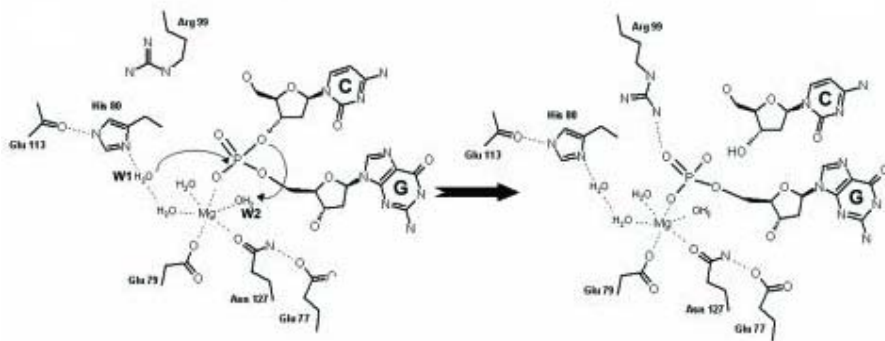


Figure 3. Suggested mechanism of cleavage of the DNA scissile backbone by the periplasmic nuclease Vvn in *Vibrio vulnificus*, by Li et al. (2003). The nucleophilic water molecule is labeled W1 and the water molecule that protonates the leaving group is labeled W2.

The proposed mechanism of Vvn share similar features to the described mechanism of the homing endonuclease I-PpoI of the His-Cys box family (Galburt et al, 1999; Galburt et al, 2002) and the non specific *Serratia* nuclease (Miller et al, 1994; Friedhoff et al, 1996; Miller et al, 1999; Friedhoff et al, 1999), see figure 4. The *Serratia* and I-PpoI endonucleases share no similarity in the sequence or overall fold outside the active site,

whereas the catalytic residues in the active sites are remarkably conserved (Friedhoff et al, 1999). The nature of the substrates for the two endonucleases is also different. Where *Serratia* has an extracellular nuclease that hydrolyzes ssDNA, dsDNA and RNA with little sequence specificity, the *I-PpoI* cleaves only dsDNA with long palindromic recognition sites (Friedhoff et al, 1999). The catalytic binding of the *I-PpoI* is although reported to only have unspecific contacts with its dsDNA target (Galburt et al, 2000). A common mechanism is proposed by Friedhoff et al. (1999), where the catalytic important residues are Arg57, Arg87, His89, Asn119, Glu127 and Arg131 in the *Serratia* endonuclease, and Arg61, His98 and Asn119 in *I-PpoI*. In these structures an identical Mg^{2+} -water cluster, with the amide Asn119 as the only residue from the protein that coordinate the magnesium, and five additional water molecules to define the octahedral geometry is found (Miller et al, 1999). Upon binding to substrate, the Mg^{2+} also coordinates to the 3' oxygen of the (deoxy)ribose and the non-bridging oxygen of the phosphate (Miller et al, 1999).

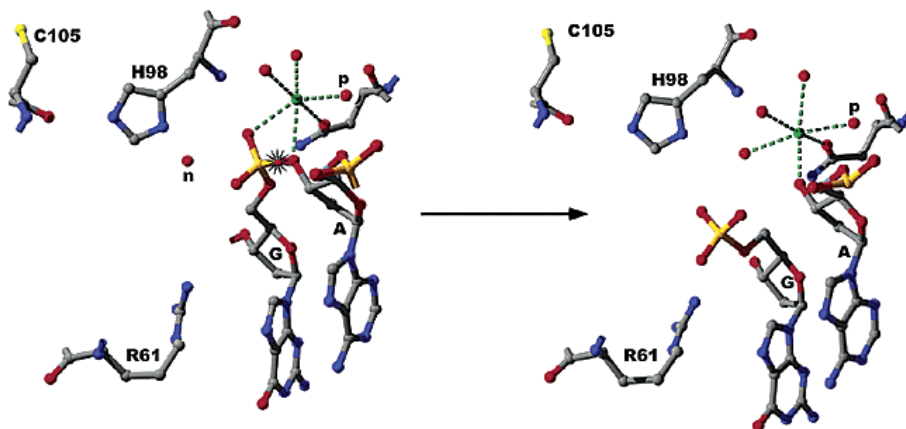


Figure 4. The mechanism of the *I-PpoI* catalyzed hydrolysis of cleavage of the phosphodiester bond. Bound, but uncleaved substrate is shown to the left, and a cleaved product complex to the right. The nucleophilic water molecule is labeled *n* and the acidic water molecule coordinated to Mg^{2+} that protonate the leaving group, is labeled *p*. The mechanism is identical to the suggested common mechanism published in Friedhoff et al. (1999). Copy from Galburt et al. (2002).

To support the proposed identical mechanisms, the catalytic activity profile of hydrolysis of the same substrate was tested for both enzymes. The two enzymes were observed to produce 3'-OH and 5'-phosphoester fragments at the same rates (Friedhoff et al, 1999). The mechanism of *Serratia* endonuclease and *I-PpoI* distinguishes from other

endonuclease mechanisms by the nucleophilic water molecule not being directly coordinated by the metal ion, and by the His98 acting as a general base (Galburt et al, 2002).

In a mutation study of the *Serratia* endonuclease (Friedhoff et al, 1996), the active site was characterized. The conclusions were that the His89 acts as a general base activating a water molecule as a nucleophile, the catalytic magnesium is positioned by the Asn119 and directly binding the substrate, and at the same time acting as a Lewis acid decreasing the pKa of the directly coordinated water molecules. These water molecules become acidic and may provide a proton for the leaving group. The magnesium ion is also thought to partially stabilize the phosphoanion product. In the mechanism of I-*PpoI*, the backbone carbonyl oxygen of a cysteine (Cys105) enhances the basic property of the basic histidine (His98), which in the Vvn and *VcEndA* is the backbone carbonyl oxygen of Glu113. In addition, the magnesium in the *Vibrio* structures is coordinated by Glu79 making the number of coordinating water molecules four instead of five. The active site in I-*PpoI* is found to put a strain on the bound DNA substrate that is relieved in the cleaved product (Galburt et al, 2002). Both the structure of the *Serratia* endonuclease and the structure of the I-*PpoI* have an arginine similar to Arg99 in the structures from the *Vibrio* bacteria (Arg57 *Serratia*, Arg61 I-*PpoI*), that stabilizes the phosphoanion product. In the proposed mechanism of Vvn the Arg99 conformation is shifted in a bend, whereas in the mechanism of *Serratia* and I-*PpoI* the 5' phosphate product is moving.

Most studies of small molecule interactions with proteins are within the field of drug discovery, and the aim may as well be to find a compound that stimulates a response, an agonist, instead of inhibit it, an antagonist. The first stages in drug discovery studies are the determination and validation of a target responsible for the activity of interest. Example of an identified but hard to validate class, are ion channel proteins that are the basis of neural function. The result of a difficult validation process is that this class of reasonable drug targets is not resembled in the large part of on-the-market drugs (Drews, 2000). The highest degree of validation lies in observing the modification of a target, as blocking of a receptor or inhibition of an enzyme, and observed reversion of disease

symptoms in clinical studies. In the identification process of a target, sophisticated microarrays are developed, and may contain the whole human genome (Smith, 2004). In microarray identification of targets, the difference between correlating and causing effect of a disease may be difficult to distinguish. If up- or down-regulated genes are used to in the identification process, protein modification or relocations *in vivo* are not taken into concern and may lead to wrong analysis and interpretations. Different cautions have to be made in inhibitor and drug design, even if the same terms and methods are applied. In drug design toxicity, binding energies that give the physiological effect but minimize undesired side effects, and abruption of multiple signaling pathways other than the one of pathological interest, must be thoroughly considered. For these reasons not all targets are suitable as drug receptors (Triggle, 2005). A number of diseases have multiple pathologies and a number of therapeutically important molecules interact with multiple targets. Two rather different diseases are stroke and schizophrenia, where in the latter case the standard effective drug interacts with a number of amine receptors and transporters in the central nervous system (Triggle, 2005). The stroke pathology concerns ionic imbalance caused by ion-channels, and the lack of effective drugs is likely because the different targets need a multi action drug, or a complex cocktail of one target drugs. In an inhibitor study performed *in vitro*, where the aim is to totally block a certain activity, pathology concerns are of minor or none importance if they do not effect the desired response.

The determination of a protein target structure at molecular and atomic levels is today dominated by the two techniques X-ray crystallography and NMR (Evans, 1995). Both methods have limitations either in steps in the workflow, or by more technical causes. For both X-ray crystallography and NMR structural determination studies, a certain amount of purified protein is needed. In most cases the protein target has to be expressed by recombinant methods, before a purification protocol is established. Recombinant expression and protein purification are experimental processes in the field of biotechnology, and are separate sciences by themselves as well as being bottlenecks in structure determination work. In NMR studies are the size of the protein and its solubility, as well of being thermal stabile over time, the main limitations. The solubility

and stability are properties of the protein target, and the restriction of favor of small size or domains of larger complexes, due to the strength of the magnetic fields possible to apply in experiments. The NMR takes advantage of isotopes with specific magnetic spins, and labeling the proteins with these isotopes make this field of structure determination usable in some cases. As the protein is inspected in solution, information about dynamic and flexible properties of the target is gained. The most utilized method although, outnumbering the cases of NMR determined structures, is the X-ray crystallography technique. This method also holds some restrictions regarding application due to the need of a uniformly ordered quality crystal. This is not always feasible, as some proteins are not stable in a soluble phase long enough to start to crystallize, or it is not possible to find the right experimental conditions within reasonable time and resources. If a quality crystal is obtained, the exposure to a strong X-ray beam produce diffraction patterns of the electron densities within the crystal. The diffraction pattern may be transferred into a structural electron-density model by applying Fourier transformations on the generated wave functions. The measured diffraction pattern is determined by the amplitude of the waves, and hence the phases of the wave functions are lost. This problem is assigned the term *phase problem*. To solve the phase problem, experimental procedures as collecting additional datasets, either on different wavelengths, SAD (single anomalous dispersion) and MAD (multiple anomalous dispersion), or changing the system slightly in heavy metal derivatives by soaking crystals with heavy metals, SIR (single isomorphous replacement) and MIR (multiple isomorphous replacement). An easier way to model phases is to transfer the phases of a similar already solved structure to generate approximate phases. This method of utilizing already known phase information is called molecular replacement. The quality of structural information in crystallization experiments is dependent of the quality of the data set of diffraction pattern, and indeed being able to solve the phase problem. X-ray crystallography determines only rigid systems, and flexibility is observed as invisible or poorly defined parts in the electron density maps, or high temperature factors. The development in technology with the application of precision robotics, opens for the possibility of more efficient administration of both resources and man-power by generating automated high-through-put processes.

The process in drug and pharmacy industry from the target discovery, validation and finding a compound of interest to the problem, until an effective quality drug is out on the market, is a time and cost expensive procedure. The path of *in vitro*, *in vivo* studies and clinical testing phases are long, and at every step the procedure may be terminated if the development does not satisfy the criteria set for the project. These may be poor prospects of solving the problem, difficulties in production steps, to high expenses and so on. To make the search for good drug candidates, known as lead compounds, more efficient, modeling experiment methods like virtual screening, VS, and molecular docking are developed. There are two different strategies to find lead compounds by computational methods, a ligand-based and a receptor based strategy. Ligand-based screening is performed when the targets receptor structure is not available, but one or more ligands are known to bind to this receptor (Jain, 2004). A pharmacophore, which is the structural feature descriptions in sets of steric and electronic parameters of a molecule that is responsible for the recognition and the biological response of a macromolecule, as blocking or triggering the activity, may be the basis for these screens. Hidden similarity of three dimensional pharmacophore patterns may be observed when comparing the activity of apparently unrelated structures concerning connectivity (Barbosa et al, 2004). A similarity principle that neighbouring compounds in the three dimensional structural space will hold the same activity is stated, but the contrary is not valid as different compound pairs may display similar activity values (Barbosa et al, 2004). The receptor-based screens demand a basis of knowledge about the receptor structure, but not necessarily any known active compound or ligand. The ligand-based screening is a strong method for identifying ligands for large classes of proteins that are difficult experimentally to structural determine and gain information about at atomic and molecular level. Examples of such classes are ion-channels and integral membrane proteins that hold important functions in biological pathways (Jain, 2004). Molecular docking is a docking method of the receptor-based type. Compounds are pulled from a library of molecules by the use of search algorithms, and are regarded as hits and ranked by their hypothetical quality. All docking programs seek to solve the docking problem, defined as the prediction of the correct bound interaction between two molecules, when given the atomic coordinates (Halperin et al, 2002).

Molecular docking is computational experiments where the identification of biological macromolecular interactions with small molecules or other macromolecules, are modeled. The macromolecule is in most cases a protein with an experimentally determined structure either by X-ray crystallography or NMR techniques, or a homology model based on high sequence similarity of already solved structures. Various computer programs have been developed to perform molecular docking, using different search algorithms and fitness functions. Three main categorize methods are developed to represent the large atomic system of a protein during a docking procedure; these are the atomic, the surface and the grid-based receptor representations. The methods applying surface representations are most used in protein-protein docking, and the computational expensive and accurate methods with atomic representation in final ranking steps. The grid representation of a receptor is based on energetic contributions that are stored in grid points. The reading of the grid is the only representation of the protein that has to be evaluated when introducing a small molecule to the system. The most basic representation of grids is given by van der Waals and electrostatic potentials. Protein-protein dockings depend largely upon surface representation, and will not be considered or described further in this thesis. These interactions are nonetheless important regarding the prediction of cellular pathways by macromolecular interactions and assemblies, as well as producing basic information for inhibitor design (Halperin et al, 2002).

The search algorithms can be categorized into three main classes based on the overall feature of search methodology; these are the systematic methods, random or stochastic methods and simulation methods (Kitchen et al, 2004). The search algorithms attempt to find the correct pose of the small molecule when bound to the macromolecule. The systematic search methods try to explore all degrees of freedom in a molecule, but face the problem of combinatorial explosion. To deal with this, the small molecules may be built fragment by fragment in the active site, and linked together covalently to the final molecule by the end of the search. Another approach is to divide small molecules into rigid core fragments and flexible side chains defined by rotatable bonds. The flexibility is modeled one bond at a time upon expanding after the core fragment position is determined. Molecular dynamics and energy minimization are the two major simulation

methods. Molecular dynamics have often problems in crossing high energy barriers, and are therefore carried out at different temperatures. Another way of avoiding these barriers is to start a search by placing the small molecule in different positions. Energy minimization is often used together with other methods, as it is only able to handle local minima.

Random and stochastic search algorithms make random changes to either a single or a population of small molecules. The Monte Carlo and the genetic algorithms, explained in short below, are together with the tabu search examples of random search methods. The standard Monte Carlo algorithms, MC, apply random Cartesian moves to the system, and accepting or rejecting these in the next step based on the Maxwell-Boltzmann probability (Halperin et al, 2002). MC does not require derivated prior information, and uses simple energy functions. Genetic Algorithm, GA, adapt biological terms and principles from the Darwinian evolution theory and Mendelian classical genetics to create and develop conformations of small molecules upon binding to macromolecules (Morris et al, 1998; Taylor et al, 2002; Jones et al, 1995). Each ligand-protein complex generated by the GA algorithm is an individual member of the *population*, which is the set of all present complexes. Each *individual* is coded by its unique genetic material, and have a genetic relationship to the rest of the members of the population in terms of orientation, translation and conformation. The search algorithm states that a particular arrangement of a ligand and a protein can be defined by a set of state values describing the translation, orientation and conformation of the ligand with respect to the protein. These state values are the *genotype description* of the system, and each state variable corresponds to a *gene* in a *chromosome* (Morris et al, 1998). The *phenotype description* of the system is based on the genotype, and is the atomic coordinates of the complex. The start of a GA search is generation of a random initial population, followed by genetic operations working iteratively on all future generations. Genetic operations may be random mutation operations, and mating operation based on elitism of the fitness between individuals of the population. The search ends when either the maximum energy evaluation or the maximum generation is reached.

A hybrid of the genetic algorithm and local energy minimization search has been developed to make the GA more robust (Morris et al, 1998). This may only be done to GA with mapping functions between the geno- and phenotype space that are invertible, meaning the phenotype is a one-to-one function of the genotype. In normal GA the mapping function is only allowed to go from the genotype space and into the phenotype space, the state variables gives the atomic coordinates and not the other way around. The wrong assumption that phenotypic characteristics obtained in an individual's lifetime can become inheritably, is the basis of the Lamarckian genetic algorithm, LGA. Not being true in the real genetic world, it does improve the molecular docking search by making a fitter child in the next generation based on the local search energy minimization step. Figure 5 illustrates the different spaces and how LGA and GA performs a search and produce new generations. GA and LGA have a fitness function incorporated in the search algorithm, whereas other may assign a fitness score after the search is performed.

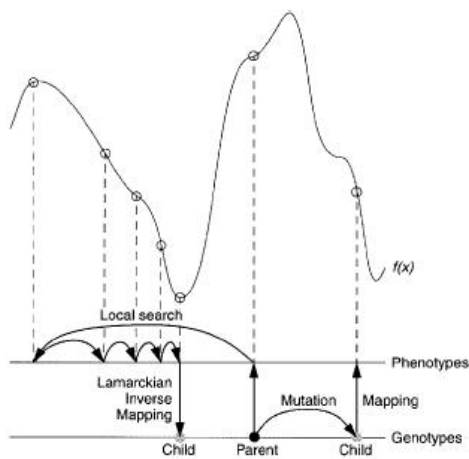


Figure 5. Illustration of genotypic and phenotypic representation, and GA versus LGA search. $f(x)$ is the fitness function evaluating the docking pose in question, the lower line is the genotype and the upper line is the phenotype space. Local search is shown to the left side of the figure, where a local energy minimization search succeeds the GA step. The application of a normal genotypic mutation operation is illustrated to the right. Copy from Morris et al. (1998).

The fitness functions applied are in most cases either force-field based, empirical or knowledge based scoring functions (Kitchen et al, 2004). The force-field fitness functions use molecular mechanics force fields to quantify the unbounded and the bounded state of a system. The various functions are based on different parameter sets, but in general have

a similar form. Interactions are often described by using van der Waals and electrostatic energy terms. Hydrophobic interactions between ligand and protein are normally calculated based on van der Waals interactions modeled by Lennard-Jones potentials, see equation 1 and figure 6. The classical 12-6 function may be altered to get a softer potential of the surface, and improve contact calculations by an annealing that treat hydrophobic and lipophilic interactions better. A softer potential may also make a rigid receptor more yielding and partially make up for lack in allowed flexibility.

$$(1) \quad E_{vdW}(r) = \sum_{j=1}^{N_A} \sum_{i=1}^{N_B} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

In equation 1, N is the number of atoms in molecule A and B, representing the protein and the small molecule respectively, ϵ is the well depth of the potentials and σ is the collision diameter of the respectable atoms i and j . The first part of the equation is the small distance repulsion and the second part the attraction term that approach zero as the distance increase.

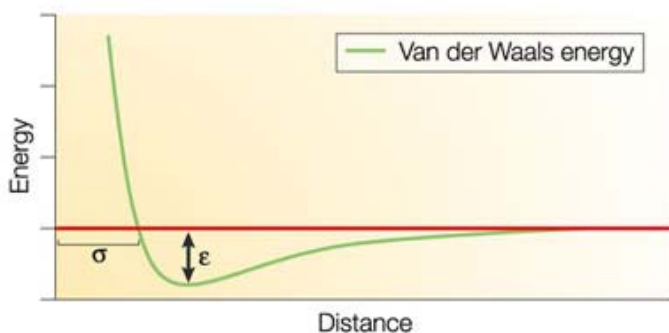


Figure 6. Illustration of the Lennard-Jones potential for the van der Waals interaction in energy, as a function of distances between atom pairs. The depth of the well is marked as ϵ and σ is the collision diameter of any atom pairs. Copy from Kitchen et al. (2004).

The electrostatic potential energy is in most cases represented as a pair-wise summation of Coulombic interactions, given in equation 2.

$$(2) \quad E_{coul}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where r_{ij} is the distance between the atom pair i and j in respectively molecule A and B, q the charge of each atom, and ϵ_0 the electric constant. A negative sign of the summation

indicate a positive interaction, and a positive sign indicate repulsion between the atom pair i and j .

Hydrogen bonds may be represented as Lennard-Jones potentials with a directional component, but some force-field based functions differ between the nature and geometry of hydrogen bond interactions. Empirical fitness functions are fit to reproduce experimental data, regarding conformation and energies upon binding as a sum of parameterized functions. Knowledge based fitness functions are designed to reproduce experimental structures rather than binding energies. Complexes of proteins and ligands are modeled using relatively simple atomic interaction-pair potentials. A limitation of the fitness scoring functions is that they are trained against a limited set of complexes, and are therefore biased towards this particular set. The force-field fitness functions have a more general form, and although validated against sets of complexes, they are regarded as less biased. A new approach to improve the current fitness functions is the combination of different functions to enhance the probability of identifying active compounds. When fitness functions are combined, a possibility of amplifying of errors may occur if the functions are correlated.

A thorough study by Warren et al. (2006) evaluated a selection of ten different docking programs and 37 scoring functions, tested on eight different protein targets. In this study they evaluated the programs performances by three criteria; the binding mode prediction, hence if the search algorithm produced reasonable docking poses, virtual screening for lead identification, and rank-ordering by affinity for lead optimization to evaluate the performance of the fitness function. Also Jain (2004) studied the docking accuracy for eight different docking programs, evaluating search methods and scoring functions capacity of improving the enrichment rate of active over inactive compounds. In both studies the docking pose accuracy were measured by superimposition of the docked pose *versus* the experimental determined state, given by root-mean-squares deviation (rmsd) in Å. Both the best ranked pose and the best coinciding rmsd of all poses were included in the evaluations. The general conclusion in both studies was that in most of the cases, the program search algorithm was able to find the correct experimental determined position

and conformation of an active small molecule upon binding to the receptor. Whereas the fitness scoring functions calculating binding energies, and hence evaluate the docking job by ranking, was not accurate enough to give a define distinction between what may be regarded as active and inactive compounds. An enhancement in the enrichment ratio compared with random screening, were although observed. Warren et al, (2006) concluded that no single program performed well for all of the targets that were tested. The problem of discriminating between correct solutions and false positives, comes from the fact that an active and correctly bound compound do not necessarily have the largest extent of buried surface, contain the largest number of hydrogen bonds or the smallest number of unsatisfied buried polar groups, when compare with other docking solutions (Halperin et al, 2002).

As the docking programs attempt to reflect the real world on molecular and atomic levels within reasonable computational resources available, a common assumption is to allow small molecule flexibility and keep the receptor rigid. This simplification has in-between-zones where parts of a small molecule may be considered rigid and other parts rotatable, as in some systematic search algorithms. The protein receptor may also be considered partially flexible either for specified residue side chains, or a larger area as a hinge changing conformation upon binding. When allowing flexibility in a system, the search is complicated and the need of more computer resources increases. A way of moving around this problem is to soften the surface of both receptor and ligand, making clashes less penalizing. In virtual screening, VS, the aim is to reduce a high number of possible compounds in a library to a manageable amount for experimentally testing. The tradeoff between flexibility and maybe some enhance in precision and the docking speed, may cause a dilemma.

To ensure that the proper tautomers and protonation states are present, both the protein target and the library of small molecule compounds have to be prepared prior to the docking process. This may be done by applying a force field that defines and improves bonds, atomic positions, inter and intra-molecular interactions and dihedral angles. The main differences between the different force fields are the way they treat dihedral angles

and intra-molecular interactions. Chen et al. (2007) investigated how the docking program AutoDock performed on docking experiments with ligand-bound metal ions, evaluating how dependent the program was on the right protonated state of the small molecule as well as placing the ligand correctly in the receptor. They concluded that the common practice of deprotonating charged compounds at neutral pH according to their pKa, was acceptable when using this program.

Aim of the study

The aim of this project was to find a lead compound for an inhibitor of the enzyme Endonuclease I from *Vibrio cholerae*, VcEndA. This lead compound, or a consensus of structural features that inhibit the activity, would have the possibility of acting as a starting point in further development of a more efficient inhibitor. An inhibitor will be useful as an additive in biotechnology transformation kits, by increase the yield of successful transformed organisms, and by deleting the experimental steps *via* endA⁻ strains in transformation experiments with organisms that secrete VcEndA homologues. To complement the study, three approaches of methodology were applied. These were *in vitro* activity measurements, structure determination of small molecule-protein complexes with X-ray crystallography techniques and computational modeling of interactions by docking and virtual screening experiments.

Material & Methods

Standard chemicals used in solutions were purchased from Merck, AppliChem and Sigma. All enzyme used in this section was expressed and purified by Dr Bjørn Altermark after methods described in Altermark et al. (2006).

Activity measurement & IC₅₀ determination

To measure the inhibition effect of the different functional groups of the HEPES molecule scaffold towards *VcEndA*, a selection of compounds with similar structure and functional groups were tested regarding inhibition effect. For this the DNaseAlert™ QC System, High Throughput Fluorometric DNase Detection Assay (Ambion Inc., Austin, TX) was applied. The DNaseAlert™ Kit contains a DNA substrate comprising both a fluorescent reporter and a dark quencher. When the substrate is un-cleaved, the quencher will adsorb emission from the reporter and no signal will be detectable. In the cleaved substrate the reporter and quencher are separated, and the fluorescent part of the molecule is free to send out an emission signal. DNase activity is measured as an increase in fluorescence by a fluorometer. Any inhibition effect would be indicated as loss in fluorescence signal. The inhibition effect of increasing concentrations of an inhibitor is detected as a lower degree of cleavage of the substrate, and hence the initial velocity will have a smaller ascent. The fluorescence was read in real time with a Spectramax Gemini fluorometer (Molecular Devices, Sunnyvale, CA). In SOFTmax Pro (Molecular Devices, Sunnydale, CA), a menu with 23 reads over 3 min, at 8 seconds intervals, was set. The wavelengths for excitation/emission was 535/556 nm as suggested by the DNaseAlert™ Kit protocol, and the plates were autmixed for 1 sec before read without any incubation time. The initial velocities were calculated by reducing the ascent of the curve by the first 3-5 measured points. Minimum three parallels were run for each compound, all at room temperature, 23 ±1 °C.

The set-up of the solutions applied was based on the inhibition trend of the HEPES molecule and the optimal reported enzymatic activity of *VcEndA*. Eight different

concentrations of a potential inhibitor were measured simultaneously to obtain a range of concentration effects. Black microtiterplates with non-binding surfaces (Corning Inc, NY) were used, and each well contained a total volume of 100 μ l. The buffer contained 20 μ l 200 mM Tris/HCl at pH 8.0, 5 μ l 100 mM MgCl₂ and 3.5 μ l 5 M NaCl, yielding a final concentration of 40 mM Tris/HCl at pH 8.0, 5 mM MgCl₂ and 175 mM NaCl. In addition, 10 μ l Substrate from the DNaseAlert™ QC System kit and 0-50 μ l from 50 mM of the possible inhibitor solutions was added, the volumes were adjusted to 90/100 μ l for each well by adding the appropriate amount of nuclease free water from the DNaseAlert™ kit. The reaction was started by adding 10 μ l of enzyme solution to each well with a multi-channel pipette. The fluorescence was immediately measured.

The concentration range of the inhibitor compounds was 0, 0.5, 1.0, 1.5, 2.0, 2.5, 5.0 and 25.0 mM, and the enzyme solution was diluted 1:40 000 to get reads in the interval 0.2-1.5 Rfu/s. The enzyme solution was made by a dilution series of 100x, 50x and 8x, using a stock solution of *VcEndA* at 2.49 mg/ml and a dilution buffer (25 mM Tris/HCl at pH 8.0, 5 mM MgCl₂ and 175 mM NaCl). For each measurement new 50x and 8x dilution series were made, whereas the 100x dilution was replaced after approximately 2 hours due to a decrease in the enzyme activity over time. Protein LoBind tubes from Eppendorf were used in the dilution series, and all of the enzyme solutions were kept on ice during the work.

The initial velocity values were plotted in GraphPad Prism (GraphPad Software Inc., San Diego, CA) to estimate the IC₅₀ which defines the concentration of any inhibitor when 50% of enzymatic activity is lost. In GraphPad, XY data tables were created by plotting logarithmic values of the molar concentration against the Rfu/s values. The plot was reduced using nonlinear regression and one site competition as default, the settings uses competitive inhibition in the active site as inhibition path. As the 0 concentration of the possible inhibition compound correspond to 100 % enzyme activity of *VcEndA*, 10⁶ was chosen as an arbitrary high concentration that would give no enzyme activity for all of the compounds. GraphPad evaluates the IC₅₀ values by calculating a 95 % confidence interval based on the sample set of three parallels. Confidence intervals are intervals that

are generated from a random sample set, so that for all further sample sets, the probability to obtain an interval that includes the true value is defined. Ideally, a confidence interval should be a short interval and have a high degree of confidence.

Crystallization, Data Collection & Refinements

The reported conditions in Altermark et al. (2006b) were used as a starting point for screening of crystallization conditions both for the co-crystallization experiment, and for producing large number of crystals for the soaking experiments.

The crystallization conditions for the co-crystallization experiment were 32 % PEG (poly ethylene glycol) 8000, 0.1 M Hepes at pH 7.75, 0.36 M sodium acetate and 10 mM MgCl₂. The crystallization experiments were put up using a protein concentration of 3.6 mg/ml, a drop size of 2 + 2 µl, and a reservoir volume of 1 ml using the hanging drop vapor diffusion method at room temperature. The protein crystallized after 24 hours, and 48 hours later the crystals had obtained their final size. The crystals were harvested after two weeks by flash-freezing in liquid nitrogen, the cryo conditions applied were equal to the reservoir solution with additional 10 % glycerol, and loops of sizes 0.1-0.2 mm were applied. When tested at the home facilities, the crystals diffracted to 2.6 Å. The final data was collected at BESSY, The Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung, where the resolution was 1.67 Å, and data were collected over 180° with 1° oscillation to a total of 180 images. The data set of the co-crystallization experiment crystals was collected at beamline BL14.2 with a MAR165-CCD detector, and a crystal to detector distance of 130 mm. The data set collected from the co-crystallization experiment was given the name Hepes4mol.

In the soaking approach, optimized crystallization conditions as described in Altermark et al. (2006b) for the neutral form applying sodium cacodylate as buffer, gave diffraction quality crystals. The hanging drop vapor diffusion method was applied with 2 + 2 µl of protein 3.6 mg/ml and reservoir solution on pre-siliconized cover slips. The reservoir solution of 1 ml contained 0.1 M sodium cacodylate at pH 6.6, 0.2 M sodium acetate and

10 mM MgCl₂, using 20-22 % PEG 6000 as precipitant. Crystals grew within a few days, but grew larger before harvested after 5 months. In the first attempts to soak crystals with a small molecule, no additional magnesium was added in the soaking and cryo solutions. The first soak solutions contained 24 % PEG 6000, 0.1 M sodium cacodylate at pH 6.6, 0.2 M sodium acetate in addition to 10 mM of one of the compounds Hepes, Ches or Taurine. The crystals were left in the soak solutions for 2 hours for the compounds Hepes and Taurine, and 24 hours for the compound Ches. The cryo solution had the same composition as the soak solution with additional 8 % glycerol. The data sets collected from these crystals were not of interest as they lacked the catalytic magnesium ion, and further description of data collection and refinements are not reported.

In the next soaking experiment, crystals grown under the same condition as described for the first soak experiment were used. In this experiment, the soak and cryo solutions applied contained 10 mM magnesium chloride, and the soaking was carried out in two steps. In the first step, crystals were soaked for 30 minutes in a solution of composition 23 % PEG 6000, 0.1 M sodium cacodylate at pH 6.6, 0.2 M sodium acetate, 10 mM MgCl₂ and additional 5 mM of one of the compounds Ches or Hepes. In the next step, the crystals were transferred to a soak solution equal in composition as the one applied in the first step, except that the Ches or Hepes concentrations were increased to 10 mM. The crystals were soaked over night in this solution. The cryo solutions applied to protect the crystals upon flash freezing in liquid nitrogen, had the same composition as the second soak solution with additional 8 % glycerol. The diffraction was tested at the home facilities, with a resolution of 3.5 Å. The final data was collected at BESSY, where the resolutions were 2.0 Å for the data set collected from crystals soaked in 5 and 10 mM Hepes solutions, and 1.9 Å for the data set collected from crystals soaked in 5 and 10 mM Ches solutions. The data sets were collected at beamline BL14.1 with a MAR225-MOSAIC CCD detector. The data set collected from the crystal soaked in Ches solutions was collected over a total range of 120° with oscillation of 0.3° between each image to a total of 400 images, and the data set collected from the crystal soaked in Hepes solutions was collected over a range of 60° with oscillation 0.3° between each image to a total of 200 images. The crystal to detector distances was respectively 210 mm and 190 mm with

exposure times of 7 and 10 seconds. The data sets were given the names Ches4 and Hepes2 respectively.

The three data sets described above, Hepes4mol, Ches4 and Hepes2, were collected at a wavelength of 0.91481 Å. The crystals were protected against radiation damage by a liquid nitrogen spray at a temperature of 100-120 K when exposed in the X-ray beam.

The data were processed using XDS (Kabsch, 1993) and the structures refined with the CCP4i program package suite (Collaborative Computational Project, Number 4, 1994). Molecular replacement was performed using the program Phaser that uses maximum likelihood techniques (McCoy et al, 2005) on the co-crystallization Hepes4mol data set, and by MOLREP that applies rotation and translation techniques (Vagin et al, 1997) for the data sets from the soaking experiments, Hepes2 and Ches4. The structure of VcEndA with PDB entry code 2g7e deposited in RCSB Protein Data Bank, was used as template for the molecular replacements. The program REFMAC5 (Murshudov et al, 1997) was used to refine the structures after manual examination and changes introduced in O (Jones et al, 1991) in a reiterated procedure. Default geometric parameters were applied, and water molecules were added in all refinement cycles except for the very last cycle.

The structures were evaluated by the program PROCHECK (Laskowski et al, 1993) and WHAT IF (Vriend, 1990), and superimposing of the structures onto each other and the deposited 2g7f structure, were performed by the program LSQKAB (Kabsch, 1976). The residues in all three data sets that had a rmsd value above 0.3 Å for the main chain, and a rmsd value above 2.0 Å for the side chain when superimposed on the deposited 2g7f structure, were manually inspected in the density maps. A temperature factor analysis was performed with the program BAVERAGE in the CCP4i program suite (Collaborative Computational Project, Number 4, 1994). To assign secondary structure elements, the structures were sent to the DSSP database (Kabsh et al. 1983). Small molecule coordinates were downloaded from the database HIC-Up, Hetero-compound Information Centre -Uppsala.

Docking

Three different docking programs, GOLD, AutoDock and Glide, were compared in their performance in order to find the most suitable program for the *Vc*EndA system prior to a virtual screening. To define the best representation of the binding site, four different scenarios for the receptor were tested for each docking program. These scenarios of the active site were the use of rigid and flexible receptor, and with or without the most exposed water molecule that coordinates directly to the catalytic magnesium ion in the active site. The target for the docking experiments was the deposited structure 2g7f in the Brookhaven Protein Data Bank, and the exposed water molecule is water molecule number 1023 in this entry. The receptor was represented as a grid for all three docking programs. The three docking programs were tested with a library containing twelve out of the thirteen compounds that had experimentally IC_{50} values determined in the activity measurements. Overall, the basic default parameters were applied for each program to avoid sources of bias, and to not introduce sophistications that were not comparable among the programs. The compounds were prepared by the Schrödingers LigPrep program, and the pH adjusted to 8.0 upon the exposure to the force field OPLS_2005. In this list, three compounds were considered active, the Hepes, Ches and MES compound ranked above sulphate and phosphate by their IC_{50} value. For three of the inactive compounds, CAPS, PIPES and POPSO, two protonation states were specified. The performance of each program and setting was assessed by compared the ranking and the ability to differentiate between active and inactive compounds of experimentally tested compounds. The compound cacodylate was omitted from the compound list because the force fields were not able to handle the element arsenic.

GOLD

A trial license was applied and accepted for the docking program GOLD version 3.2 (Jones et al, 1995; GOLD tutorial manual). The program was run both from the graphical interface and from the command line.

GOLD applies a Genetic Algorithm (GA) in the docking searches. A short explanation of key settings is given below. A population is the set of possible solutions, hence the number of docking orientations. Each member of the population is regarded as a chromosome holding information of mapping of hydrogen bond atoms onto complementary hydrogen bonding atoms of the protein. The same is valid for hydrophobic points of the ligand and the match onto hydrophobic areas of the receptor. Conformations about rotatable bonds in flexible ligands, and rotation of–OH bonds in the receptor are also information obtained in the chromosomes.

Each chromosome is given a value by a fitness function, and the whole population is ranked by this function. Chromosomes may mate or mutate, and children with a higher fitness than the worst ranked members of the population, replace these. The population is under a constant selection pressure, describing the ratio of possibility that the fittest member contra an average member of the population, is selected as a *parent* for the next generation. To preserve diversity it is possible to have numerous populations arranged in islands, and each population may contain niches to make sure two members of a population will not hold too similar properties. The numbers of genetic operations, mating, mutation or migration between islands decides how long each docking will run.

The default setting were applied for the docking experiments. These were a population size of 100 and a selection pressure ratio of 1.1. 5 islands were set, producing 5 different populations with a total number of 500 members. The number of genetic operations allowed to run was 100 000 and number of allowed members in a niche was limited to 2. The search was centered at the catalytic Mg²⁺ ion and allowed a radius of 20 Å.

GoldScore was chosen as fitness function, but GOLD also supports the scoring functions ChemScore and ASP (Astex Scoring Potential), in addition to a User Defined Score function. GoldScore is defined as the negative sum of four energy contributions, the external hydrogen bonds between protein and ligand $E_{\text{ext}}(\text{H-bond})$, external van der Waals $E_{\text{ext}}(\text{vdW})$ of the interaction between protein and ligand, internal van der Waals contribution from the ligand $E_{\text{int}}(\text{vdW})$ and internal torsion strain of the ligand $E_{\text{int}}(\text{strain})$. In addition it is possible to add a contribution of internal hydrogen bonds in the ligand $E_{\text{int}}(\text{H-bond})$, and contribution of constraints and covalently bindings between protein and the ligand.

Only the basic GoldScore was applied as the fitness function, and in the final function the external van der Waals contribution was multiplied by 1.375 to encourage hydrophobic interaction in an empirical correction. The GoldScore equation is given in equation 3.

$$(3) \quad E_{\text{total}} = (1.375)E_{\text{ext}}(\text{vdW}) + E_{\text{ext}}(\text{H-bond}) + E_{\text{int}}(\text{vdW}) + E_{\text{int}}(\text{strain})$$

The van der Waals energies were calculated using Lennard Jones potentials, the 12-6 potential for internal van der Waals, and the softer 8-4 potential for the external van der Waals interactions. Hydrogen bond energies and directionalities were empirical and assigned by the parameter file.

To perform a docking with flexible receptor, the configuration file containing the experimental setting was edited before run from the command line. At the end of each file a rotamer library was added, specifying atoms and dihedral angles involved with rotational bonds. The manual advises only to add rotation of bonds experimentally observed to rotate. Arg99 and Glu113 were therefore the only residues which were allowed to have two and one rotatable bonds respectively. The permitted rotation around each bond was determined from the inspection of conformations in crystal structures of these residues. The resulting poses from the docking experiments were inspected by using the graphical interface GoldMine.

AutoDock

The docking program AutoDock 4 (Morris et al, 1998) was run by scripts and by using AutoDock Tools as graphical interface.

The Lamarckian Genetic Algorithm (LGA) was applied as the search algorithm of choice. The receptor was prepared by adding charges and producing a grid for all of the present elements. A grid box of 60x50x60 in Å was centered on the catalytic Mg²⁺ ion.

AutoDock 4 uses semiempirical free energy force field estimations as fitness function (Huey et al, 2007), which includes a full desolvation model with terms for all elements, and intramolecular terms in its energy calculation. The calculation uses pair-wise terms to evaluate interactions between ligand and protein, and an empirical method to estimate the contributions of the surrounding water molecules. The free energy is estimated to be equal the difference between the non binding systems and the bound complex system:

$$(4) \quad \Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf})$$

where the abbreviation L-L stands for internal interactions in the ligand, P-P for the concerning internal interaction in the protein, and P-L for interaction between the protein and the ligand. ΔS_{conf} estimate the conformational entropy lost upon binding. The state $V_{unbound}^{P-L}$ is assumed to be zero, and the same assumption is made for the term $(V_{bound}^{P-P} - V_{unbound}^{P-P})$ in docking experiments with rigid receptor setting. The pairwise atomic terms include evaluation of van der Waals interactions, hydrogen bonding, electrostatic interaction, desolvation and torsional energy, and the whole term is given in equation 5.

$$(5) \quad V = W_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{h-bond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{solv} \sum_{i,j} (S_i V_j + S_j V_i) e^{\left(-r_{ij} / 2\sigma^2 \right)} + W_{tor} N_{tor}$$

The normal van der Waals term is given as a 12-6 Lennard-Jones potential and the hydrogen bond directional term as a 12-10 Lennard-Jones potential. Electrostatic interaction is calculated by a Coulombic term and the torsional term measures the unfavorable entropy due to restriction of degrees of freedom for the ligand due to

binding, and is proportional to the number of sp^3 hybridized bonds in the ligand. The desolvation term has an atomic solvation parameter, S_i , that estimates the energy needed to transfer the atom between a fully hydrated and a fully buried state, and a parameter, V_i , that estimates the amount of desolvation when the ligand is docked. A volume term is added as the exponential part and a sum over all atoms. Each of the contributions is weighted by the constants given as W , which are optimized toward experimentally characterized complexes.

The general setting for the docking experiments was a population size of 150, maximum number of generation set to 27 000, and an elitism rate of 1. The mutation rate was 0.02 and a crossover rate at 0.8 was used. The number of hybrid GA-LS search was set to 20 and the probability of a random LS search at 0.06. The coefficients, W 's, in the free energy calculation from the different contribution in equation 5 were 0.1560 for the normal van der Waals term, 0.0974 for the hydrogen bond terms, 0.1465 for the electrostatic term, 0.1150 for the desolvation term, and 0.2744 for the torsional term.

In the flexible receptor, six residues were defined as allowed to rotate about certain bonds. These residues were Arg72, Glu77, His80, Arg99, Ser131 and Asn132 located in the active site, and were written to a separate file that specified atoms and rotational bonds in these residues. The docking experiments with flexible receptor were run *via* the AutoDock Tools graphical interface. The resulting docking poses were inspected by using AotoDock Tools.

Glide

The docking program Glide, version 4.5 (Schrödinger, LLC, New York, NY, 2007), was applied. The docking experiment was performed using the graphical interface program, although the program may perform the same tasks from the command line.

The search algorithm in Glide is a hierarchical filter that score hydrophobic and polar interactions followed by Monte Carlo sampling (Taylor et al, 2002; Glide manual). The hierarchical filter has three main steps based upon systematic search of a rigid core and flexible side groups as the ligand representation. The first step is a systematic and exhaustive search of all locations and orientations over the active site. The site points are evenly distributed and cover the active site on equally spaced 2 Å grids. The locations and orientations are evaluated by a set of distance criteria concerning the ligand center to surface distance, and the site point to the receptor surface. A match is allowed to continue further down the hierarchical system if the center of the ligand and the site point show consistency. The second step is divided in a diameter test, evaluating allowed number of clashes between ligand and receptor, a subset test where the rotation about the diameter of the ligand is considered. The subset test also considers all atoms capable of forming hydrogen bonds or metal interactions. If the subset test is past, a score step is performed by evaluating the actual position, but also the best possible score by moving ± 1 Å in the x, y and z directions. The poses that obtain the top scores from this scoring is re-scored where the ligand as a rigid body is allowed to move ± 1 Å in the Cartesian directions. The third step in the Glide search is an energy minimization using the calculated grids for the receptor with a softer surface setting. When the conformation is determined internally, an optimization step is performed allowing torsional motion about both core and rotatable side chains. Finally a set of top ranked poses are sampled and searched for alternatively local-minima regarding core and flexible rotatable groups, exploring the possibility to improve the energy score.

GlideScore is based on the fitness function ChemScore, but includes an additional term for steric clashes, and a buried polar term to penalize electrostatic mismatch. For the docking experiment with rigid receptor, GlideScore was applied as fitness function, and

when flexibility is added to the receptor Induced Fit Docking score, IDFScore, was applied. IFDScore is the sum of the GlideScore from the redocking step and 5 % of the Prime energy of the refinement calculation. The energetic contributions to GlideScore is given in equation 6.

$$(6) \textit{GlideScore} = (0.065)E_{vcW} + (0.130)E_{coul} + E_{lipo} + E_{H-bond} + E_{metal} + E_{buryP} + E_{RotB} + E_{site}$$

Default settings were applied in the docking experimnts. The catalytic Mg^{2+} was the center of the search, and in the flexible receptor representation all residue side chains within a radius of 2 Å, in addition to Arg99, were allowed to rotate.

Virtual Screening

Libraries were downloaded from the small molecule database ZINC, version 7 (Irwin et al, 2005), using the search and browse option in the main menu. The libraries were already prepared by Schrödingers LigPrep program by the ZINC database to contain the right protonation state prior to downloading, with pH set between 5 and 9.5 considered as biological active.

Library 1 contained compounds with an aminoethanesulfonic acid group as a structural element. In total the library contained 580 compounds. A negative charge of -1 was set as a constraint. The aminoethanesulfonic acid group was allowed to be a linear terminal element, or an internal or terminal part of more complex linkages or ring systems.

Library 2 contained compounds with either a sulfonic, phosphoric or phosphor mono- or diester group as structural element. In total a number of 1620 compounds. A negative charge of -1 or -2 was set as constraint.

To perform the VS, the docking program GOLD was chosen after evaluation of the docking experiment of the experimental tested compound list. The receptor setting applied contained four Mg^{2+} - coordinated water molecules and flexibility for the residues Arg99 and Glu113, allowing moderately rotation as described in the docking experiment of GOLD with the same flexible receptor settings. The top 100 docking poses of each library were visually inspected in GoldMine.

Results

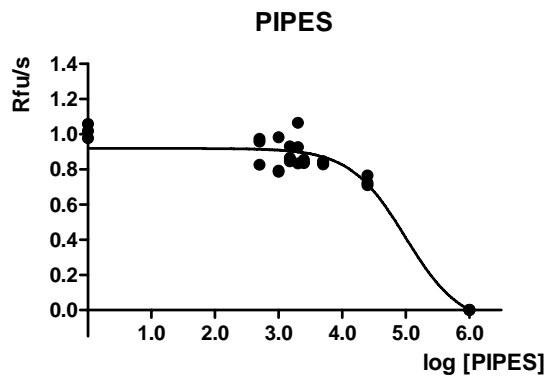
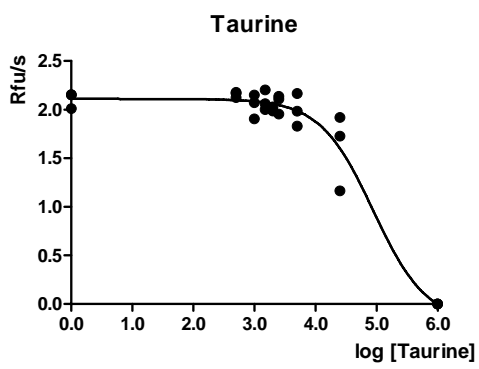
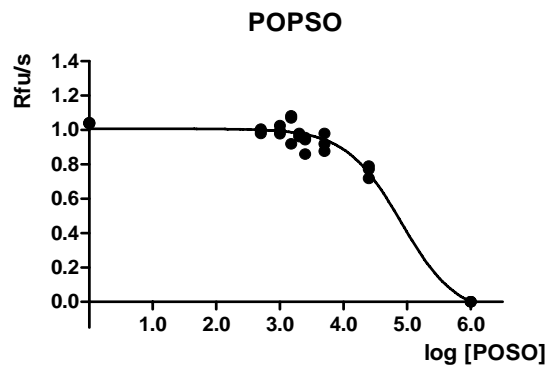
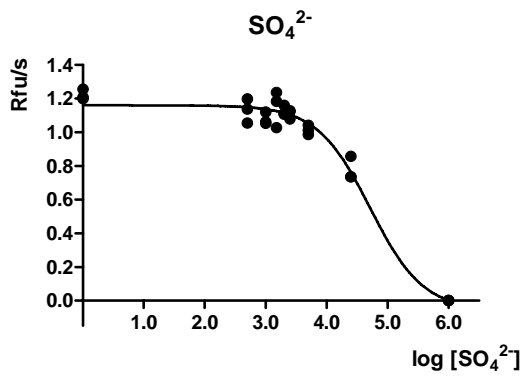
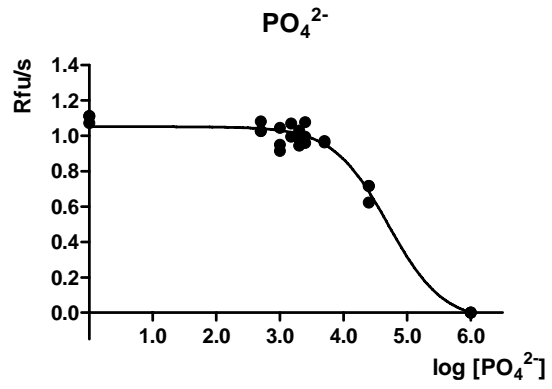
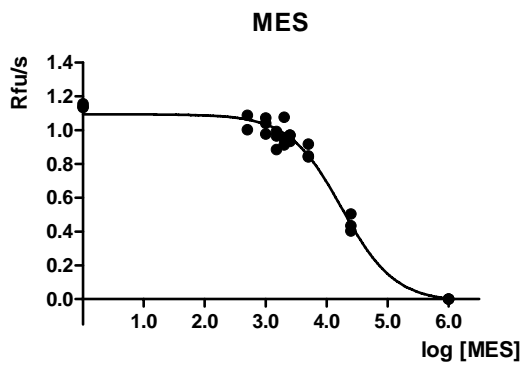
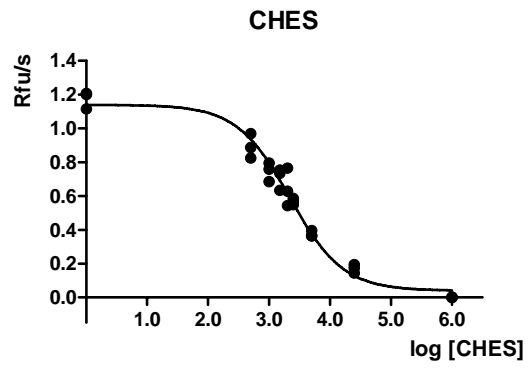
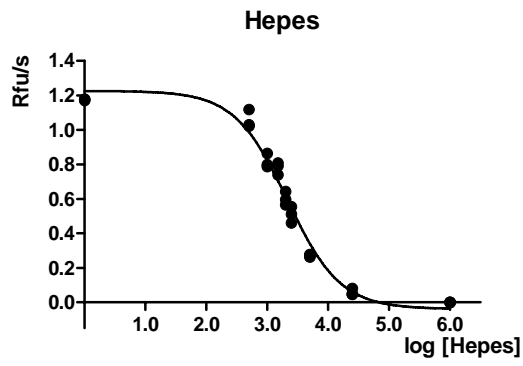
Experimental IC₅₀ values

Table 1. IC₅₀ values of compounds tested for inhibition effect on *VcEndA*, ranked by decreasing inhibition effect.

Compound	IC ₅₀ nM	95 % confidence interval		pKa	Mg ²⁺ - chelating
		Lower limit	Upper limit		
Hepes	2140	1 770	2 587	7.5	no
Ches	2198	1 750	2 759	9.3	
MES	17 672	13 712	22 777	6.1	
HPO ₄ ²⁻	49 994	37 392	66 844	pKa ₁ 2.15 pKa ₂ 7.21 pKa ₃ 12.36	
SO ₄ ²⁻	51 508	37 451	70 842	1.92	
POPSO	80 301	56 542	114 044	7.8	
Taurine	88 068	51 632	150 216	1.5	
PIPES	98 313	49 452	195 451	6.8	no
MOPS	215 033	---	---	7.2	
Cacodylate	574 823	---	---	6.27	
EPPS	1.9x10 ⁶	---	---	8.0	0.0005%
MOPSO	2.3x10 ⁷	---	---	6.9	
CAPS	5.9x10 ⁸	---	---	10.4	0.005%

* The 95 % confidence intervals of MOPS, EPPS, MOPSO and CAPS are very wide and give no reasonable interpretation.

The experimentally determined IC₅₀ values in table 1 show that the two compounds Hepes and Ches, with the values 2140 nM and 2198 nM respectively, reduce the activity of *VcEndA* in an equivalent rate. In addition is the IC₅₀ value of MES, at 17 672 nM, lower than the inactive compounds of the ranked list. These three compounds are considered as active. Lower inhibition effect than for the phosphate and sulfate ions with IC₅₀ values about 50 000 nM are interpreted as no inhibition effect, as the extension beyond the sulfonic acid group does not have a positive effect upon interaction with *VcEndA*. The confidence intervals are omitted in the lower part of the table, as they are too wide to give a reasonable interpretation. The pKa of each compound indicates how likely it is that the sulfonic acid group or phosphate has a negative charge at pH 8.0 in the activity assay. The tested compound CAPS is not likely to be deprotonated at the experimental pH, as it has a pKa of 10.4. The results in table 1 are based on the graphs in figure 7, applying the computer program GraphPad Prism in the analysis.



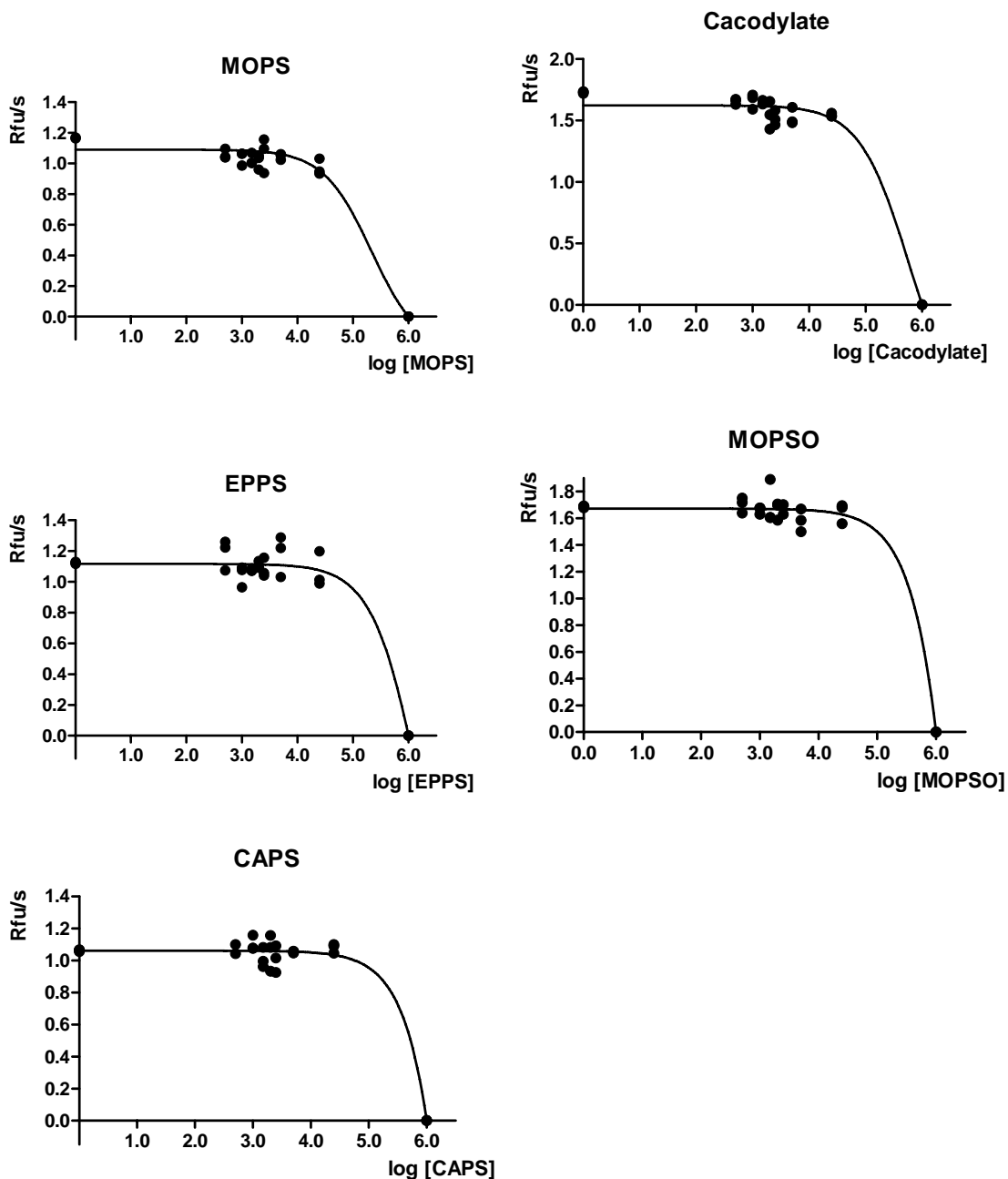


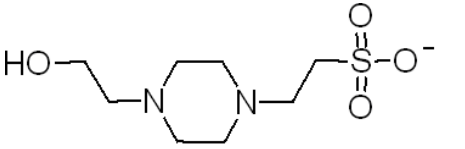
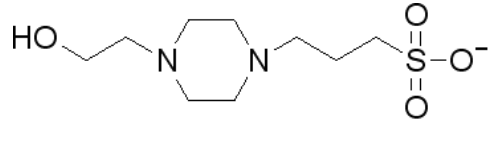
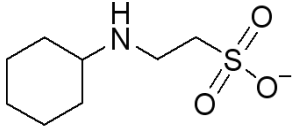
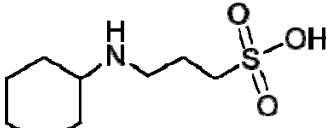
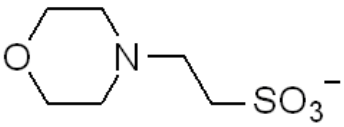
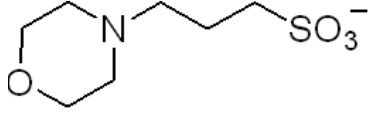
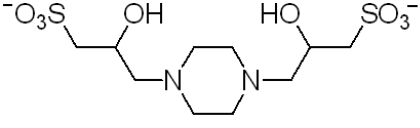
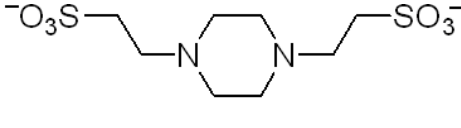
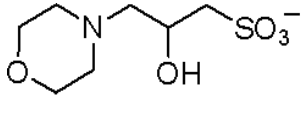
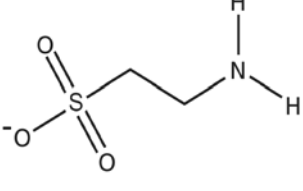
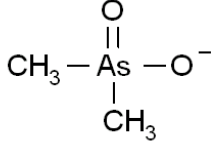
Figure 7. The logarithm of the tested compounds concentrations in nM plotted against Rfu/s by the program GraphPad Prism, each black dot indicates an initial velocity of one parallel. The graphs form the datasets which the IC_{50} values are calculated from.

Figure 7 shows the initial velocity of the cleavage of fluorescence substrate in Rfu/s as a function of the logarithm of tested compounds concentrations in nM. The graphs are sorted in the same order as their IC_{50} values in table 1, and all parallels applied are marked as black dots. The graphs of Hepes, Ches and MES show a decrease in the initial

velocity with an increase in the concentration from 0 to 25.0 nM. Even though all off the graphs decline down to 0 Rfu/s at 6.0, this is a fixed point chosen from the trends of the decrease in activity for Hepes and Ches.

The molecular structures shown in table 2 are aligned with corresponding similar compounds side by side, with a variation in the length of the carbon chain by two and three carbons. The structures of PO_4^{2-} and SO_4^{2-} are not included in table 2. Note that the IC_{50} values in table 1 of these corresponding structures with two *versus* three carbon long chains are not comparable in size. For instance Hepes with an IC_{50} value of 2.1×10^3 with its corresponding three carbons long relative EPPS with an IC_{50} value of 1.9×10^6 . The same is valid when comparing the IC_{50} values of MES, IC_{50} at 1.7×10^4 , and MOPS, IC_{50} at 2.1×10^5 . The large difference between the IC_{50} values of Ches and CAPS, 2.1×10^3 and 5.9×10^8 , are likely caused by the protonated state of the sulfonic acid in CAPS.

Table 2. Structures of the experimental tested compounds in the activity assay.

 <p>HEPES</p>	 <p>EPPS</p>
 <p>CHES</p>	 <p>CAPS</p>
 <p>MES</p>	 <p>MOPS</p>
 <p>PIPSO</p>	 <p>PIPES</p>
 <p>MOPSO</p>	 <p>Taurine</p>
 <p>Cacodylate</p>	

*CAPS is not likely to be deprotonated at pH 8.0, as the pKa is 10.4.

Crystallization & Data Collection

The *VcEndA* crystallized readily, but optimization was necessary to obtain quality crystals that diffracted to high resolutions. With the reported crystallization conditions, *VcEndA* formed multiple crystal forms in a mixture of needles, 2D plates and 3D crystals, see figure 8. The data sets collected from the crystals applied in the first soak experiment did not contain the catalytic magnesium ion in the Mg^{2+} -water cluster, and the statistic of data collections and structure determinations are not further described.

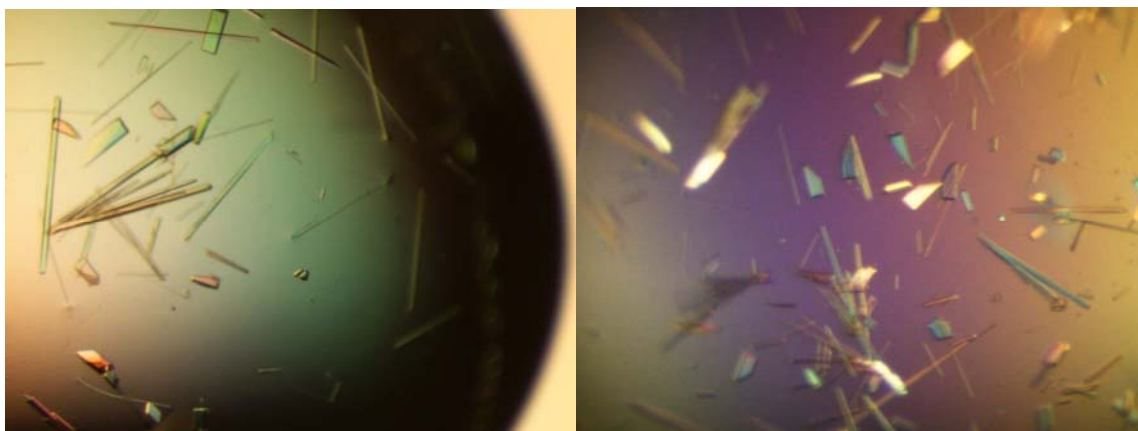


Figure 8. To the left crystals soaked in solutions containing Ches (5 and 10 mM), to the right crystals soaked in solutions containing Hepes (5 and 10 mM).

The space group for the crystals used in the soak experiments, Hepes2 and Ches4, was the orthorhombic $P2_12_12_1$ with one molecule in the asymmetric unit, as previously reported by Altermark et al. (2006) where similar crystallization conditions were applied. The data set of the co-crystallization experiment, Hepes4mol, gave the new monoclinic spacegroup $P2_1$ with four molecules in each asymmetric unit. None of the obtained data sets contained the desired small molecule, but the data sets collected from crystals of the soak experiments containing the catalytic Mg^{2+} , Hepes2 and Ches4, had the buffer molecule cacodylate bound in the active site. This interaction is described in more details in subsequent sections. Statistics of the data collections are reported in table 3.

Table 3. Parameters and statistics of the data collection of the three data sets Hepes2, Ches4 and Hepes4mol. Values for the outer shells are given in parenthesis, 2.11-2.00 Å for Hepes2, 2.01-1.90 Å for Ches4 and 1.76-1.67 Å for Hepes4mol.

	Hepes2	Ches4	Hepes4mol
Data collection			
Diffraction limit	2.0	1.9	1.67
Unit-cell parameters (Å)	a = 40.63 b = 64.37 c = 75.43	a = 40.56 b = 64.05 c = 74.96	a = 63.92 b = 88.86 c = 74.43 $\alpha = 90$ $\beta = 92.705$ $\gamma = 90$
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_1$
Wavelength (Å)	0.91841 (BL14.1 BESSY)	0.91841 (BL14.1 BESSY)	0.91481 (BL14.2 BESSY)
Total No. of reflection	56 088 (8 174)	126 906 (18 088)	354 753 (40 387)
No. of unique reflections	13 970 (2 004)	15 895 (2258)	95 084 (12 893)
Completeness (%)	100 (100)	99.9 (99.9)	98.8 (92.2)
Anomalous completeness (%)	99.5 (99.3)	99.8 (98.7)	---*
$I/\sigma(I)$	6.1 (1.8)	6.2 (1.9)	8.5 (2.3)
R_{merge} (%)	11.9 (42.0)	8.7 (41.6)	4.6 (30.6)
Multiplicity	4.0 (4.1)	8.0 (8.0)	3.7 (3.1)
Wilson B factor (Å ²)	16.4	20.6	18.4

* The anomalous signals were merged for Hepes4mol.

When comparing the data sets Hepes2 and Ches4, the Ches4 statistics come better off even though the crystals are assumed of equal quality. This is a result of the strategies applied in the data collections, as twice as much data were collected for Ches4. This is reflected in numbers of reflections collected, the signal to noise ratio, R_{merge} and twice as high multiplicity, and in the slightly higher resolution of 1.9 Å compared to 2.0 Å. Hence the Wilson B factor is higher for Ches4 than Hepes2, by 4.2 Å². This is most likely not a coincidence, but a result of harsher exploitation in the beam. The collection of the Hepes4mol data set has an overall better statistics than the other two, and the crystal from where it was collected had a higher quality as the resolution of 1.67 indicates. This data set has four molecules in the asymmetric unit, and the packing is rather tight, given by a Matthew coefficient of 2.27 per molecule and a calculated solvent content of 46 %. The Matthew coefficient is even lower for the Hepes2 at 2.05 and 40 % solvent content, and

Ches4 at 2.03 and 39 % solvent content, all indicating a tight packing structure in the crystals.

Structure determination & Refinements

The solution of one molecule in the asymmetric unit of the data sets Hepes2 and Ches4, were identified by being the only solution with a significant Rf/sigma value of about 11.9 in the log file generated by the molecular replacement program MOLREP. All other solutions were flatted out at values below four. The four molecules in the asymmetric unit were identified as the correct solution by the cell content analysis of Matthews coefficients, by the molecular replacement program Phaser. This solution had a relative frequency of 0.922, whereas the next best solution with three molecules per asymmetric unit, gave a relative frequency of 0.293. Only 3 peaks were identified higher than the threshold of 75 %, with Z-scores as high as 19-21. The boundary value is 8 for a clear solution, and the fourth peak may not have been identified by Phaser due to the highest score biasing the 75 % threshold unrealistically high. The solution with four molecules in each asymmetric unit, was identified as the correct one.

The refinements was performed by applying the program REFMAC5 and manual inspection and introduction of changes by the program O, in an iterative process as reported in Material & Methods. Hepes4mol had a few residues with two conformations, and needed more manual work as the automatic procedure did not manage to trace the whole sequence through the densities. The default settings of 3.0 for mF_o-DF_c and 1.0 for $2mF_o-DF_c$ were used in correcting the models. Parameters and statistics of the refinements are given in table 4.

Table 4. Parameters and statistics of the refinements of the three data sets Hepes2, Ches4 and Hepes4mol.

	Hepes2	Ches4	Hepes4mol
Refinement			
R _{work} (%)	16.98	18.58	18.02
R _{free} (%)	26.01	24.95	22.01
Average B factors (Å ²)			
Main chain	15.56	20.671	15.45
Side chain	18.88	24.045	17.89
Water molecules	22.32	22.77	27.69
Total	17.66	22.14	17.88
No. protein atoms	1697	1696	6784
No. of solvent molecules (including Mg ²⁺ and Cl ⁻)	138†	87†	941‡
R.m.s. deviations			
Bond lengths (Å)	0.023	0.022	0.014
Bond angles (°)	1.901	1.793	1.329
DPI (Cruickshank)	0.1931	0.1711	0.1124
Ramachandran plot , residues in %			
Most favoured regions	90.6	93.9	94.5
Additionally allowed regions	8.3	5.5	5.0
Generously allowed regions	1.1	0.6	0.6
Disallowed regions	0	0	0

† Contain a cacodylate molecule in addition to the Mg²⁺ and Cl⁻.

‡ Contain 4 Mg²⁺ and 4 Cl⁻ ions.

The starting and final R factors for the data sets were 29.33 and 16.98 % for Hepes2, 30.42 and 18.58 % for Ches4, 22.71 and 18.02 % for Hepes4mol. The final R_{free} values were 26.01 % for Hepes2, 24.95 % for Ches4 and 22.01 % for Hepes4mol respectively. Ramachandran plots are attached in appendix I, figure AI-1.

Structure validation

Superimposition of the structures, and calculation of rmsd values were carried out to study deviating features between the *VcEndA* molecules in the three data sets. Residues with significant differences were inspected in the density maps with caution to B-factors, the lack of a uniform density and the nature of residue, together with its localization regarding an internal or external position. Main chain deviations were considered more significant than a variation in side chain rotamers. The superimposition of the main chains of the molecules onto the deposited structure 2g7f is shown in figure 9.

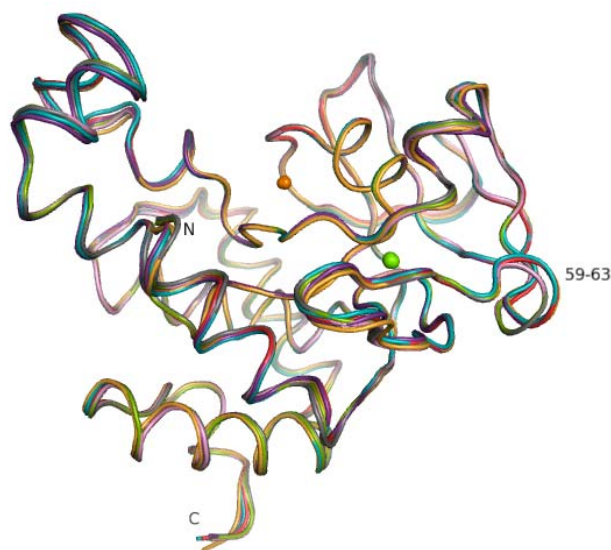


Figure 9. Superimposition of the main chains of the structures *Ches4*, *Hepes2* and molecule A-D of *Hepes4mol* onto the deposited structure 2g7f. The catalytic magnesium ion is shown in orange, the structural chlorine in green, *Ches4* in pink, *Hepes2* in green, *Hepes4mol*; molecule A in purple, molecule B in red, molecule C in blue, molecule D in yellow, and 2g7f in grey. The N-terminal is labeled N and the C-terminal labeled C. The short helix 59-63 show the only significant deviation between the main chain of 2g7f, *Ches4*, *Hepes2*, and the main chain of molecule A-D of *Hepes4mol*.

The only significant deviation is observed in the helix labeled 59-63, assigned by DSSP to 59-62, where the backbones of molecules A, B, C and D of *Hepes4mol* have shifted up to 2.06 Å. When inspected in the electron density maps, the regions of the structure containing the 59-63 helix are not in close contact with neighbouring molecules for any of the molecules A-D. This surface region is on the contrary given more space than in the *Hepes2* and *Ches4* electron density maps, when comparing neighbour molecule interaction as a plausible cause of the deviation. If given more space, this part of the

structure is possibly able to be flexible. The short helix 59-63 is not close to the active site, and when compared to the deposited 2g7f structure, the shift of the helix does not effect the positions of the catalytic important residues, see figure 14 in next section.

The average rmsd values of the main chains, the side chains and for all atoms of the superimposition of Hepes2, Ches4, and molecule A-D of Hepes4mol onto the deposited 2g7f structure are listed in table 5.

Table 5. Root-mean-square deviation values of the superimposition of the structures of Hepes2, Ches4, and molecule A-D of Hepes4mol onto the deposited 2g7f structure. Rmsd values for the residues Arg99 and Glu113 are listed in the four last columns.

	Main chain	Side chain	All atoms	Arg99, main chain	Arg99, side chain	Glu113, main chain	Glu113, side chain
Hepes2	0.151	0.455	0.729	0.220	1.965	0.170	1.069
Ches4	0.184	0.479	0.737	0.298	1.962	0.156	1.410
Mol A	0.385	0.759	1.087	0.438	0.454	0.238	0.342
Mol B	0.320	0.688	1.024	0.551	0.583	0.239	0.389
Mol C	0.371	0.724	1.028	0.546	0.559	0.208	0.257
Mol D	0.377	0.758	1.068	0.220	0.311	0.196	0.258

The structures of Hepes2 and Ches4 show a higher similarity to the deposited structure 2g7f than Hepes4mol, but this may be an effect of the crystallization conditions and different packing as two different space groups indicates. The rmsd values of residue Arg99, which is known to be flexible, and the nearby Glu113, are shown in the two last columns of table 5. The average values of the main chain rmsd of Arg99 in the structures of Hepes2 and Ches4 are slightly higher than the average values of the whole structures. In comparison is the side chain rmsd values much higher than the average, indicating that the side chain has changed conformation. The inspection of the rmsd values for molecule A-D of Hepes4mol, shows a smaller deviation for residue Arg99 of molecule D than for the other molecules, compared to the 2g7f structure. The rmsd values of Glu113 indicate a similar conformer for molecule A-D of Hepes4mol as the conformer of Glu113 in 2g7f. The B values of Arg99 and Glu113 are inspected for all of the Arg99 residues in the respectively pdb files, and show no indication of a poorly defined location. The average rmsd values for the structures, molecules A-D, indicate that the Hepes4mol deviates more from the deposited 2g7f structure than Ches4 and Hepes2. This is illustrated in figure 10 by the superimposition of the main chain rmsd values of the structures of Hepes2, Ches4

and molecule A-D of Hepes4mol onto the 2g7f structure. The rmsd values concerning the superimposition of molecule A-D of Hepes4mol onto 2g7f, shows that molecule A and C deviates most from the deposited structure. Overall the largest deviations, with main chain rmsd > 0.6 and side chain rmsd > 2.0 , are caused by poorly defined floppy surface residues. But some of the large deviations arise from parts of the amino acid sequence that DSSP assign no regular secondary structure elements, hence interpreted as loops.

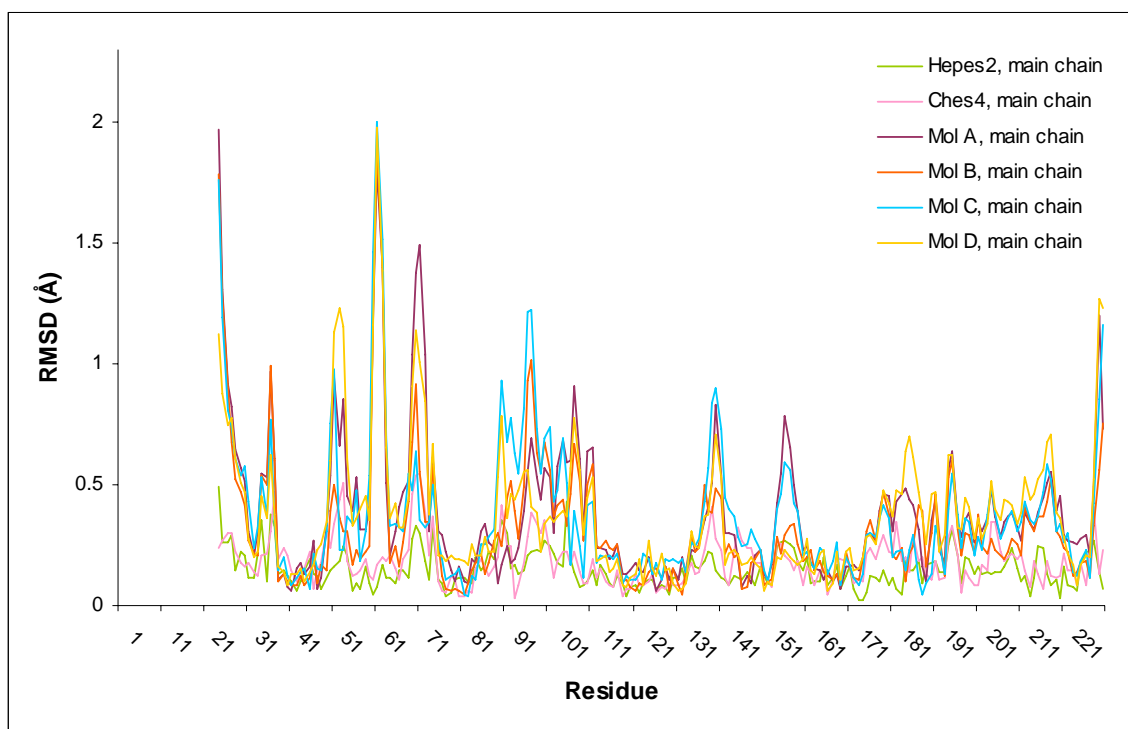


Figure 10. *Rmsd values in Å calculated from the superimposition of the main chain atoms in the structures Hepes2, Ches4 and molecule A-D of the Hepes4mol structure onto the deposited 2g7f structure of VcEndA in the RCSB Protein Data Bank. Hepes2 is shown in green, Ches4 in pink, Hepes4mol; molecule A in purple, molecule B in red, molecule C in blue and molecule D in yellow. Molecule A and C deviates most from the 2g7f structure.*

The majority of high rmsd values of the main chain (>0.3) and side chain (>2.0) of Hepes2 arises from badly defined residues on the surface. This is also the case for the large side chain values of Ches4, but interestingly, the majority of high main chain deviation values are from residues that are well defined. When comparing the structures of Hepes2 and Ches4 to each other, the deviations between these structures are expected to be small as the crystals are grown under the same crystallization conditions. The rmsd values from superimposing of the main chains and the side chains of these two structures

onto each other are illustrated in figure 11. The average rmsd value of the main chain is 0.130 Å, and the average value for all of the atoms is 0.729 Å. The largest RMSD values from the superimposition of Hepes2 and Ches4 onto each other, come from poorly defined surface residues.

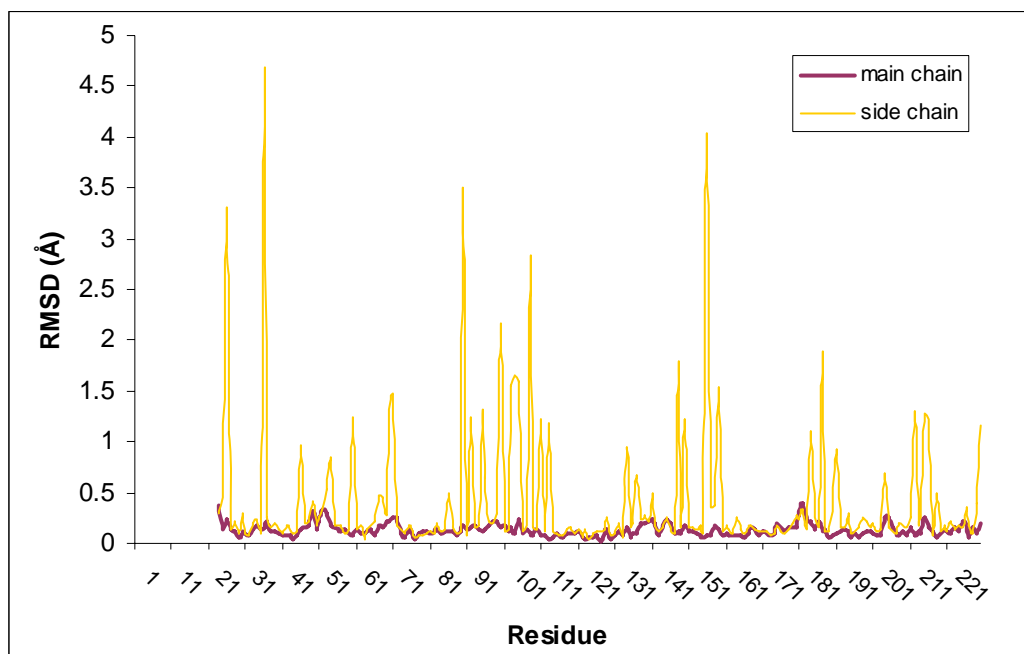


Figure 11. Rmsd values in Å calculated from the superimposition of the structure Hepes2 onto the structure of Ches4. The side chain rmsd values are colored yellow and the main chain values by a thick purple line.

The rmsd values of the superimposed structures of molecule A-D in Hepes4mol onto each other, are listed in table 6, and graphs are attached in appendix II, figure AII-2 to AII-6. The rmsd values indicate that molecule B and C share the highest similarity with a rmsd of 0.199 Å for the main chain atoms and 0.681 Å for all atoms. Molecule C and D differ the most in the Hepes4mol structure with an average rmsd value of 0.332 Å for the main chain atoms, and 1.006 Å for all the atoms in the molecules.

Table 6. Rmsd values in Å for superimposition of molecule A-D in the Hepes4mol data set. Rmsd values for main chain atoms are listed above the diagonal, and values for all atoms are listed below.

	Mol A	Mol B	Mol C	Mol D
Mol A		0.217	0.243	0.258
Mol B	0.984		0.199	0.282
Mol C	0.963	0.681		0.332
Mol D	0.928	0.927	1.006	

The Ramachandran plots do not report any deviations from normal behavior of valid structures, see appendix I figure AI-1 a), b) and c) for plots. The Ser131 residue is in the generously allowed areas in all of the plots. Ser131 is located at the rim of the active site, although not assigned as catalytic important, and is well defined in all of the maps. The unusual phi/psi angle of this residue is most likely caused by constraints by the active site. One additional residue, Lys68, is located in the generously allowed areas in the Hepes2 dataset. Upon inspection, Lys68 has a badly defined density and is located on the surface of the structure. In the data sets Ches4 and Hepes4mol, Lys68 shows the same tendency of being poorly defined, although not alarmed by the Ramachandran plots. DSSP assigns, with minor deviations, the same secondary structure for all molecules in the three structures Hepes2, Ches4 and Hepes4mol. The remarks in the validation reports by WHAT IF are of minor importance, and are mostly addressing poorly defined residues localized on the surfaces. For the Hepes4mol structure, a minority of the asparagine and glutamine residues are flipped in the wrong conformation. These are inspected and are of small concern for the overall quality of the structure.

Analysis of changes in the active site of VcEndA

In the structure of the data sets Hepes2 and Ches4, the buffer molecule cacodylate was found in the active site of both structures with an occupancy of 0.5, see figure 13. A superimposition of the structures Hepes2 and Ches4 onto the 2g7f structure represented by secondary structure elements, are shown in figure 12. The active site is framed by a red square, and catalytic residues are shown in sticks and balls. Cacodylate has a tetragonal structure with an arsenic in the middle that is covalently bound to two methyl groups, a *carbonyl* double bond to oxygen and a hydroxyl group. The *carbonyl*- and the hydroxyl oxygen atoms may not be distinguished visually from each other by their densities in the maps. The cacodylate molecules in the structures of Hepes2 and Ches4 are identically bound, and have replaced the nucleophilic activated water molecule W1 of the active site of the deposited 2g7f structure of VcEndA, see figure 13 for details.

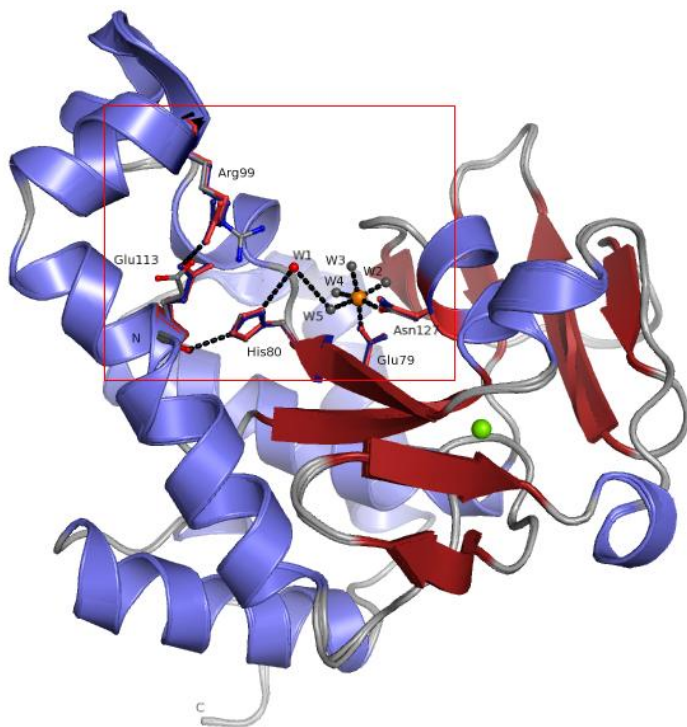


Figure 12. The superimposition of the structures of Hepes2 and Ches4 (the cacodylate molecules are not showed in this figure) onto the structure of 2g7f. The red square frames the active site.

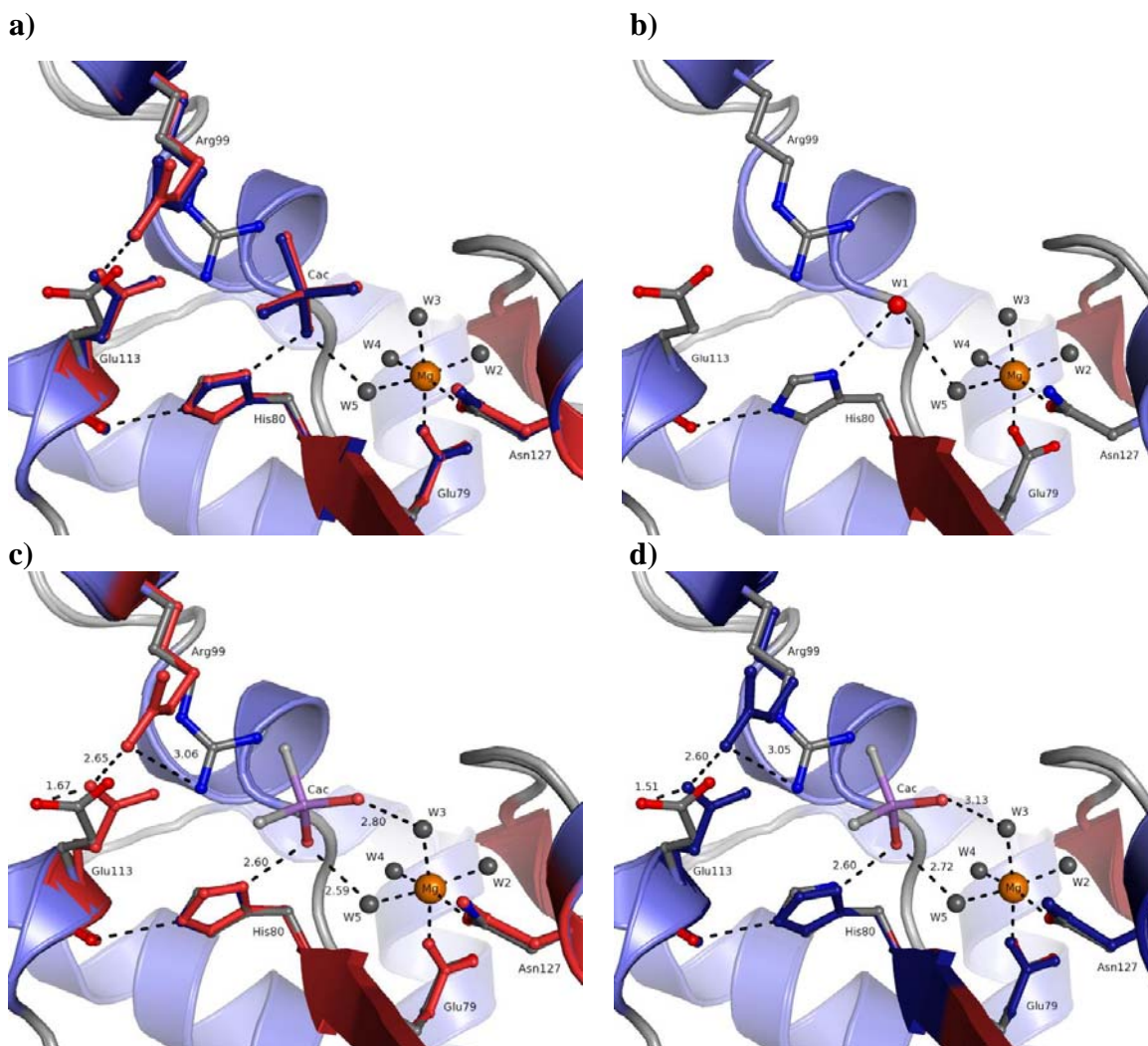


Figure 13. The superimposition of the active sites of the structures *Hepes2* in red and *Ches4* in blue, with a cacodylate molecule labeled *Cac* bound, onto the *2g7f* deposited structure colored by elements. Figure a) shows the superimposition of both *Hepes2* and *Ches4* onto *2g7f*, b) shows the active site of *2g7f* with the nucleophilic water molecule *W1*, c) shows the superimposition of *Hepes2* in red onto *2g7f* with interactions marked as dotted and labeled lines, and figure d) shows the superimposition of *Ches4* in blue onto the structure of *2g7f* with interactions labeled and marked as dotted lines. The residues *Arg99* and *Glu113* have changed conformation in *Hepes2* and *Ches4* compared to the *2g7f* structure, and forms a salt bridge of length 2.65 Å in *Hepes2* and 2.60 Å in *Ches4*.

Figure 13 shows the detailed comparison of the changes of the active site, upon the binding of the inactive compound cacodylate. The overall hydrogen bond pattern of *W1* in *2g7f* appears to be maintained by the *O1* in cacodylate in both of the structures *Hepes2* and *Ches4*. *O1* is hydrogen bonded to the *Nδ1* of the catalytic residue *His80*, with a bond length of 2.60 Å in both datasets, and to the Mg^{2+} coordinated water molecule labeled

W5, 2.59 Å in Hepes2 and 2.72 Å in Ches4 respectively. The O2 of the cacodylate molecule interacts with the Mg²⁺ coordinated water molecule W3, with distances of 2.80 Å in Hepes2 and 3.13 Å in Ches4. To make room for the cacodylate molecule in the active site, the side chains of Arg99 and Glu113 have shifted compared to the coordinates in the 2g7f structure, and formed a salt bridge between the Oε1 of Glu113 and one of the NH₂ group in the guanidinium in Arg99. The measured distances of this salt bridge is 2.62 Å for Hepes2 and 2.60 Å for the Ches4 structure. The changed conformation and the distances reported above are illustrated in figure 13 c) for Hepes2 and in figure 13 d) for Ches4. A pair-wise comparison of the distances of equivalent atoms in the two residues Arg99 and Glu113 are listed in table AIV-1 in appendix AIV.

When the active sites of Hepes2, Ches4 and Hepes4mol are investigated, the interactions according to the published literature are valid. The Glu113 backbone carbonyl oxygen is hydrogen bonded to the Nε2 of His80, and Glu77 supports the position of Asn127 with a hydrogen bond. For the Mg²⁺-water cluster, three out of the four coordinated water molecules have hydrogen bonds to the backbone carbonyl oxygens of residues in the active site. W5 is bound to the backbone of Trp78, W2 to the backbone carbonyl oxygen of Ser131 and to the directly Mg²⁺-coordinated residue Asn127's backbone carbonyl oxygen, and water molecule W4 is hydrogen bonded to the backbone carbonyl oxygen of His80. The fourth water molecule, W3, is not bound to any backbone as it is the one removed upon substrate binding. The W5 is making a water bridge through the nucleophilic W1 to the His80 in molecule A-D of Hepes4mol. This is illustrated in figure 14 by the superimposition of molecule A-D of Hepes4mol onto 2g7f. Interestingly, this water molecule also has a water bridge through another water molecule to the second conformation of Asn132 in molecule A, see figure 14. As it is only in molecule A of Hepes4mol Asn132 has two conformations, this is the only place this water bridging network is observed. The catalytic important residues in Hepes4mol coincide otherwise with the residues in 2g7f for all of the four molecules A-D.

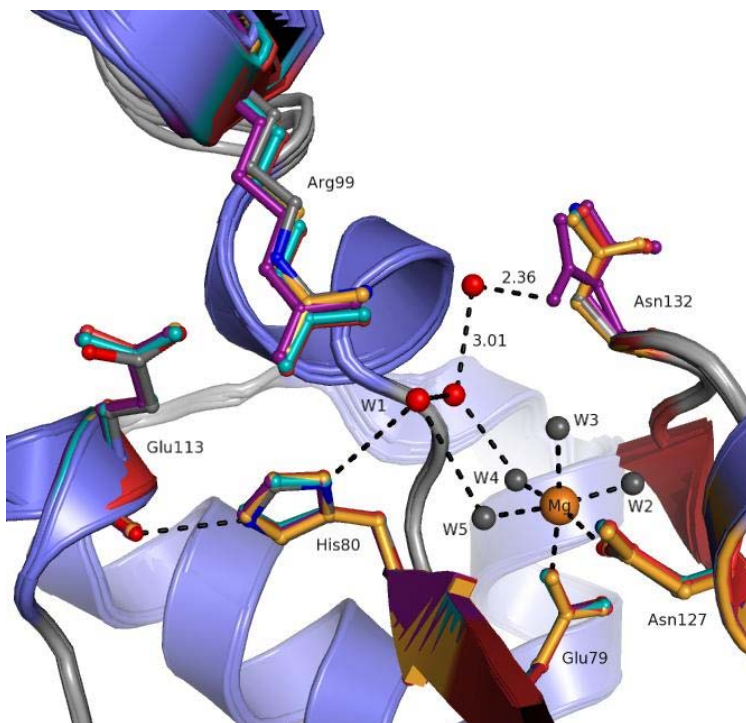


Figure 14. The superimposition of molecule A-D in the Heps4mol structure onto the structure of 2g7f. Molecule A is shown in purple with Asn132 in two conformations, molecule B in red, molecule C in turquoise, molecule D in yellow, and the structure of 2g7f is colored by elements.

When comparing the Heps2 and Ches4 structures with the two mutant His80Ala dsDNA-Vvn complexes (PDB entry code 1oup and 2ivk), by superimposition of the active site residues onto 2g7f, residue Arg99 and Glu113 of Vvn show some flexibility as seen in figure 15. Figure 15 shows the deposited 2g7f structure of VcEndA colored by elements, the active site of Heps2 in red and Ches4 in blue of VcEndA with the small molecule cacodylate bound, and two structures with un-cleaved DNA substrate bound in the His80Ala mutant of Vvn, molecule B of 1oup in orange and A of 2ivk in green, and two structures of cleaved DNA product, molecule A of 1oup in turquoise and molecule C of 2ivk in yellow. Arg99 of the Vvn-substrate structures, in orange and green, have a similar conformation as the bent Arg99 in the VcEndA structures containing a cacodylate molecule, although to a smaller degree. Whereas in the Vvn-product structures, shown in turquoise and yellow, residue Arg99 have a similar conformation as in the 2g7f structure. The position of the cacodylate molecule is nearby the DNA-backbone phosphate that interacts with the catalytic Mg^{2+} in both substrate and product complexes. The distances

from the tetrahedral arsenic molecule in cacodylate, to the phosphate phosphorous in bound DNA-product VvnA is 2.14 Å, and to the DNA-substrate phosphate phosphorous 2.64 Å.

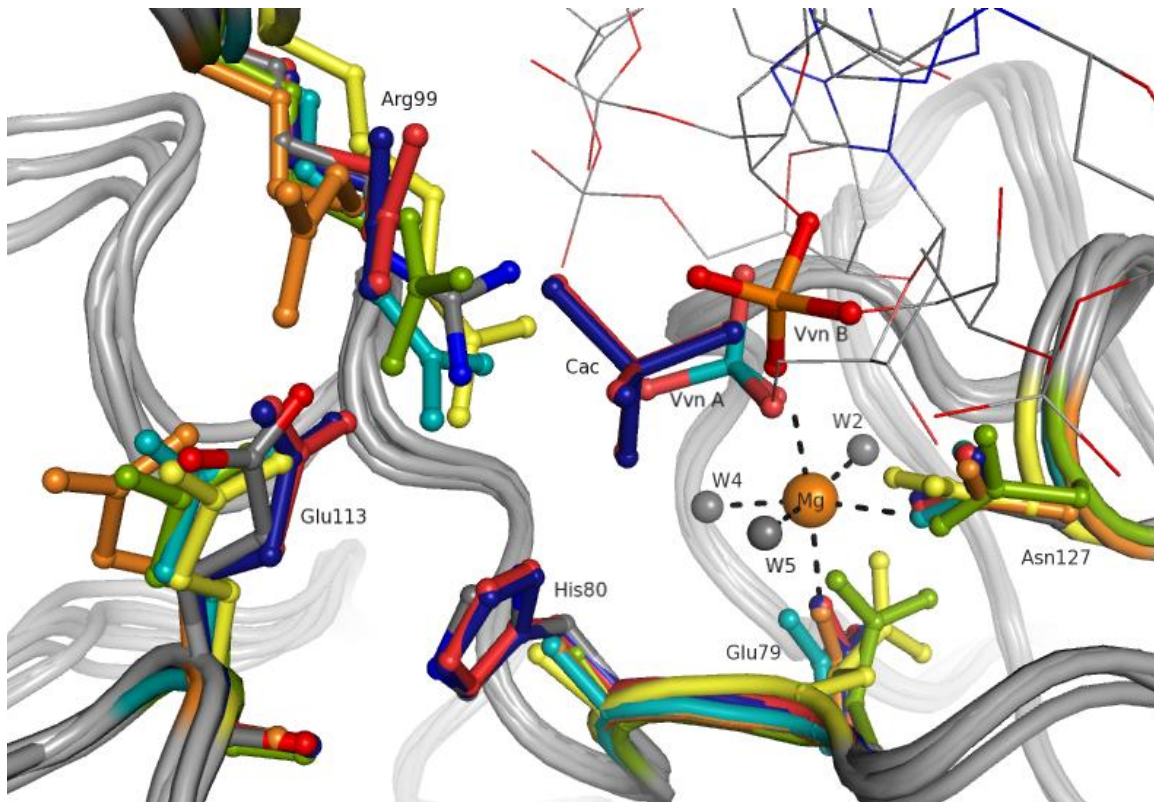


Figure 15. The superimposition of the active sites of the two deposited His80Ala mutant dsDNA-Vvn complexes of the Vvn homologue of VcEndA (entry code 1oup and 2ivk), and the Heps2 and Ches4 structure onto 2g7f. Residues of Ches4 is shown in blue, Heps2 in red, molecule A of 1oup in turquoise (cleaved DNA product), molecule B of 1oup in orange (un-cleaved DNA substrate), molecule A of 2ivk in green (substrate), and molecule C of 2ivk in yellow (product). The structure of 2g7f is colored by elements. Molecule C of 2ivk is equivalent to molecule B, and molecule D of 2ivk is equivalent to molecule A. The phosphate of the DNA backbone in the cleaved product-Vvn complex (1oup) that is covalently bound to the catalytic Mg^{2+} is labeled VvnA and shown in sticks. The phosphate of the DNA backbone in the un-cleaved substrate-Vvn complex (1oup) is labeled VvnB and shown in sticks. The rest of the DNA-substrate is shown in lines.

Docking & Virtual Screening

Docking experiments of the library containing experimentally IC_{50} determined compounds in table 1, by the three docking programs AutoDock, Glide and GOLD, and a virtual screening of the best evaluated system with a larger library, are reported in this section. All of the four different scenarios regarding the receptor state are reported in table 7, 8 and 9 for the three molecular docking programs GOLD, AutoDock and Glide. The four reasonable states of the active site, the receptor, were the assumption of a rigid or a flexible receptor, and with three or four coordinated water molecules in the Mg^{2+} -water cluster. Only the top ranked poses of each compound with experimentally determined IC_{50} values, assigned by the respective programs fitness scores, are evaluated.

Table 7. The ranking of experimentally tested compounds with the docking program GOLD. The compounds are ranked by the assigned GoldScore fitness score. Compounds with more than one protonation state have their charge given in parenthesis.

GOLD							
The receptor contains 4 Mg^{2+} - coordinated water molecules				The receptor contains 3 Mg^{2+} - coordinated water molecules			
Rigid receptor		Flexible receptor		Rigid receptor		Flexible receptor	
Compound	GoldScore	Compound	GoldScore	Compound	GoldScore	Compound	GoldScore
POPSO (-2)	66.12	POPSO (-2)	64.13	POPSO (-2)	66.00	POPSO (-1)	63.97
POPSO (-1)	57.92	POPSO (-1)	61.61	PIPES (-2)	58.87	PIPES (-2)	60.05
PIPES (-2)	56.92	PIPES (-2)	53.22	POPSO (-1)	56.87	PIPES (-1)	57.85
PIPES (-1)	55.70	PIPES (-1)	52.85	PIPES (-1)	56.29	POPSO (-2)	57.11
EPPS	50.48	CAPS (-1)	49.08	MOPSO	51.31	MOPSO	53.90
Hepes	48.86	EPPS	49.00	MOPS	50.34	MOPS	52.90
CAPS (-1)	48.51	MOPS	48.87	EPPS	50.16	MES	51.14
MOPSO	48.08	Hepes	47.91	MES	50.01	EPPS	49.76
MOPS	45.81	Ches	45.70	CAPS (-1)	47.87	Hepes	49.44
MES	45.13	MES	45.32	Hepes	46.21	CAPS (-1)	48.22
Ches	44.95	MOPSO	44.77	Ches	45.77	Taurine	46.46
CAPS (0)	40.90	Taurine	42.35	Taurine	45.47	Ches	44.64
PO_4^{2-}	40.02	PO_4^{2-}	40.40	CAPS (0)	41.03	CAPS (0)	38.13
Taurine	38.90	CAPS (0)	38.39	SO_4^{2-}	37.53	SO_4^{2-}	37.61
SO_4^{2-}	36.43	SO_4^{2-}	35.72	PO_4^{2-}	36.90	PO_4^{2-}	37.33

The assigned fitness of the top ranked docking poses of the three active compounds Hepes, Ches and MES by the docking program GOLD, indicate that GOLD manage to discriminate between active and inactive compounds. Although the active compounds are not at the top of the ranking list, the program shows indications of being able to treat the three compounds by similar criteria. The difference in fitness between four and three

water molecules in the active site shows no indication that the active compounds coordinate directly to the catalytic magnesium ion. GOLD benefits by being allowed moderately flexibility for bonds in residue Arg99 and Glu113, by converging the active compounds to a total of difference of 2.59 by GoldScore. The best ranked docking poses by GOLD were obtained with the setting of flexible receptor and four Mg²⁺-coordinating water molecules. The docking poses show that the program docks all of the compounds in a pocket between the positively charged residues Lys28, Arg72, Arg75 and Arg99 in the active site, see figure 16 a) and figure 17 a). When comparing the best ranked poses of Hepes, Ches and MES, the program have assigned very similar poses with the same localization for the sulfonic acid group, see figure 18 a).

Table 8. The ranking of the experimentally tested compounds with the docking program AutoDock. The compounds are ranked by the assigned energy in kcal mol⁻¹. Compounds with more than one protonation state have their charge given in parenthesis.

AutoDock							
The receptor contain 4 Mg ²⁺ -coordinated water molecules				The receptor contain 3 Mg ²⁺ -coordinated water molecules			
Rigid receptor		Flexible receptor		Rigid receptor		Flexible receptor	
Compound	kcal mol ⁻¹	Compound	kcal mol ⁻¹	Compound	kcal mol ⁻¹	Compound	kcal mol ⁻¹
CAPS (-1)	-4.22	PO ₄ ²⁻	-13.61	SO ₄ ²⁻	-6.05	SO ₄ ²⁻	-17.34
SO ₄ ²⁻	-4.17	SO ₄ ²⁻	-12.94	CAPS (-1)	-5.69	PO ₄ ²⁻	-17.12
MOPS	-3.82	Ches	-12.28	MOPS	-5.21	Ches	-13.98
MOPSO	-3.80	MOPSO	-12.24	Ches	-5.01	Taurine	-13.62
Ches	-3.79	Taurine	-12.23	MES	-4.91	MOPS	-13.07
EPPS	-3.66	PIPES (-2)	-11.87	PIPES (-2)	-4.84	POPSO (-2)	-12.77
MES	-3.60	EPPS	-11.86	EPPS	-4.67	PIPES (-1)	-12.61
Hepes	-3.50	MOPS	-11.86	MOPSO	-4.52	MOPSO	-12.41
PIPES (-2)	-3.13	PIPES (-1)	-11.38	Taurine	-4.20	EPPS	-12.38
CAPS (0)	-2.94	Hepes	-11.26	PIPES (-1)	-4.14	MES	-12.33
POPSO (-2)	-2.88	POPSO (-2)	-11.10	Hepes	-4.04	PIPES (-2)	-11.87
Taurine	-2.64	CAPS (-1)	-10.34	POPSO (-2)	-3.75	CAPS (0)	-11.40
PIPES (-1)	-2.40	MES	-10.08	CAPS (0)	-3.43	POPSO (-1)	-11.19
POPSO (-1)	-0.97	CAPS (0)	-10.03	POPSO (-1)	-2.61	Hepes	-11.17
PO ₄ ²⁻	---*	POPSO (-1)	-9.53	PO ₄ ²⁻	---*	CAPS (-1)	-11.09

* The program was not able to dock this compound.

The docking program AutoDock was able to process the three active compounds Hepes, Ches and MES in a similar way, with a difference in energy of only 0.29 kcal mol⁻¹ for the rigid receptor, four Mg²⁺-coordinated water molecule-setting, see table 8. The other three settings assigned free energy for these compounds in a non-systematic order, and have therefore a lower probability of being able to differ between active and inactive

compounds. The lack of improvement in assigning free energy by adding flexibility to a set of residues in the active site, indicates that the program does not have any advantages from this in the docking process. As the difference in assigned energies from the setting with four water molecules compared with three in the active site are small, directly coordination toward the catalytic magnesium ion are not a likely scenario. When inspecting the docking poses of the rigid receptor, four Mg^{2+} -coordinated water molecule docking experiment by Autodock, see figure 16 b) and figure 17 b), all of the compounds were placed in the same pocket in the active site as reported for GOLD. The three active compounds are shown in figure 18 b) by their highest ranked conformations, and have very similar docking poses, with the sulfonic acid group placed in the same position.

Table 9. The ranking of experimentally tested compounds with the docking program Glide. The compounds are ranked by the fitness functions GlideScore for rigid receptor setting, and IFDScore for the flexible receptor setting. Compounds with more than one protonation state have their charge given in parenthesis.

Glide							
The receptor contain 4 Mg^{2+} -coordinated water molecules				The receptor contain 3 Mg^{2+} -coordinated water molecules			
Rigid receptor		Flexible receptor		Rigid receptor		Flexible receptor	
Compound	Glide-Score	Compound	IFD-Score	Compound	Glide-Score	Compound	IFD-Score
SO_4^{2-}	-6.85	PO_4^{2-}	-631.05	SO_4^{2-}	-8.64	PO_4^{2-}	-630.24
PO_4^{2-}	-6.14	SO_4^{2-}	-624.78	PO_4^{2-}	-8.53	SO_4^{2-}	-626.65
MOPS	-5.68	POPSO (-2)	-620.49	POPSO (-2)	-8.45	CAPS (-1)	-618.69
MES	-5.52	PIPES (-2)	-618.09	Taurine	-8.10	PIPES (-2)	-618.67
MOPSO	-5.49	CAPS (-1)	-617.61	MES	-8.03	POPSO (-2)	-618.65
Taurine	-5.29	MOPS	-616.77	PIPES (-2)	-7.66	EPPS	-617.06
PIPES (-2)	-5.21	POPSO (-1)	-616.63	MOPS	-7.46	Ches	-616.61
EPPS	-4.75	Ches	-616.40	MOPSO	-7.44	Taurine	-616.54
POPSO (-2)	-4.65	MOPSO	-615.93	Hepes	-7.35	MOPS	-616.14
Hepes	-4.36	Taurine	-615.73	POPSO (-1)	-7.25	POPSO (-1)	-615.81
POPSO (-1)	-4.30	EPPS	-615.25	CAPS (-1)	-7.00	MOPSO	-615.29
Ches	-4.26	MES	-614.77	EPPS	-6.73	MES	-614.47
CAPS (-1)	-3.91	Hepes	-614.37	Ches	-6.45	Hepes	-613.87
PIPES (-1)	-3.58	PIPES (-1)	-612.92	PIPES (-1)	-5.94	PIPES (-1)	-613.87
CAPS (0)	-2.68	CAPS (0)	-610.73	CAPS (0)	-2.45	CAPS (0)	-610.02

The docking program Glide assigns non-systematic fitness for the three active compounds in all of the four settings, see table 9. The difference between four and three Mg^{2+} -coordinated water molecules do not indicate that the docked compounds coordinate directly to the catalytic magnesium. The rigid receptor setting shows the highest

probability to discriminate between active and inactive compounds, as it assigns similar fitness score to the top ranked docking of Hepes and Ches, although the low overall position does not give a positive impression. When the side chains in the active site are allowed flexibility, the active compounds fitness converges, but to lower positions in the overall ranking. The lack of being able to converge the docking of Ches and Hepes, as these are considered to obtain equal activity from the experimentally determined IC₅₀ values, together with the low position, penalize the flexible receptor setting. When inspected the best ranked docking poses of the rigid, four water molecules setting, see figure 16 c) and figure 17 c), all of the compounds are docked into the active site pocket. The active compounds have different poses in this pocket, and these are not interpretable as similar poses, see figure 18 c). The sulfonic acid group is not uniformly positioned within the active compounds.

Table 7, 8 and 9 give an overall consensus of guidelines for a protocol of docking experiments for *VcEndA* and small molecules. None of the three programs indicate that the molecules are directly coordinated to the catalytic magnesium, as the differences in energies upon comparison are not considered large enough to energetically favor the loss of water molecule W3. Four Mg²⁺-coordinated water molecules are therefore the appropriate setting. The docking experiments do not seem to gain from an excess of flexibility, but the rotation of Arg99 should be allowed in a given set of bonds, as the performance by the program GOLD improved upon allowing Arg99 and Glu113 flexibility. When visually inspected, the side chain of Arg99 does rotate upon in the modeled bindings of active compounds in GOLD, although not as extreme as observed in the experimentally determined structures of Hepes2 and Ches4 containing a cacodylate molecule.

When comparing the top ranked poses docked by the three programs GOLD, AutoDock and Glide, all of the programs place the small molecule compounds in the same pocket in the active site, see figure 16. The majority of compounds are docked between the positively charged residues Lys28, Arg72, Arg75 and Arg99, see figure 17.

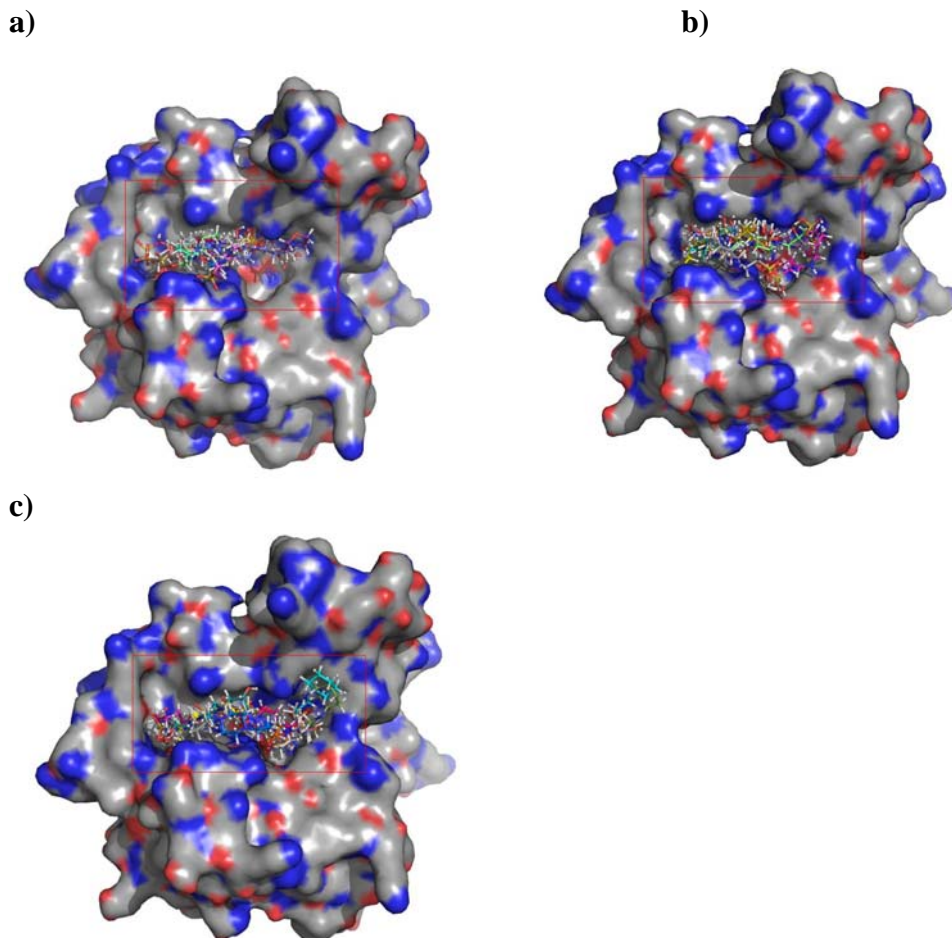


Figure 16. The top ranked docking poses of all compounds with experimentally determined IC_{50} values with the a) flexible receptor setting by GOLD, b) the rigid receptor setting by Autodock, and c) the rigid receptor setting by Glide. All containing four water molecules coordinated to the catalytic magnesium ion. The red squares frame the section of docked compounds in the active site shown in detail in figure 17 and 18. The surfaces are generated from the deposited 2g7f structure and are colored by elements.

When inspected in detail, GOLD docks the compounds more uniformly in the pocket, whereas AutoDock and Glide utilize a larger area and cover the cavity where the catalytic Mg^{2+} -water cluster is localized, see figure 17. In GOLD, only the inactive compound EPPS have its sulfonic acid group in another position than the rest of the docked compounds. AutoDock have a larger portion of four inactive compounds with a deviating sulfonic acid group placement, and the docking program Glide assigns the sulfonic acid group for the docked compounds in two major clusters, with active and inactive

compounds presence in both. In addition one of the inactive compounds has a third deviating placement for its sulfonic acid groups.

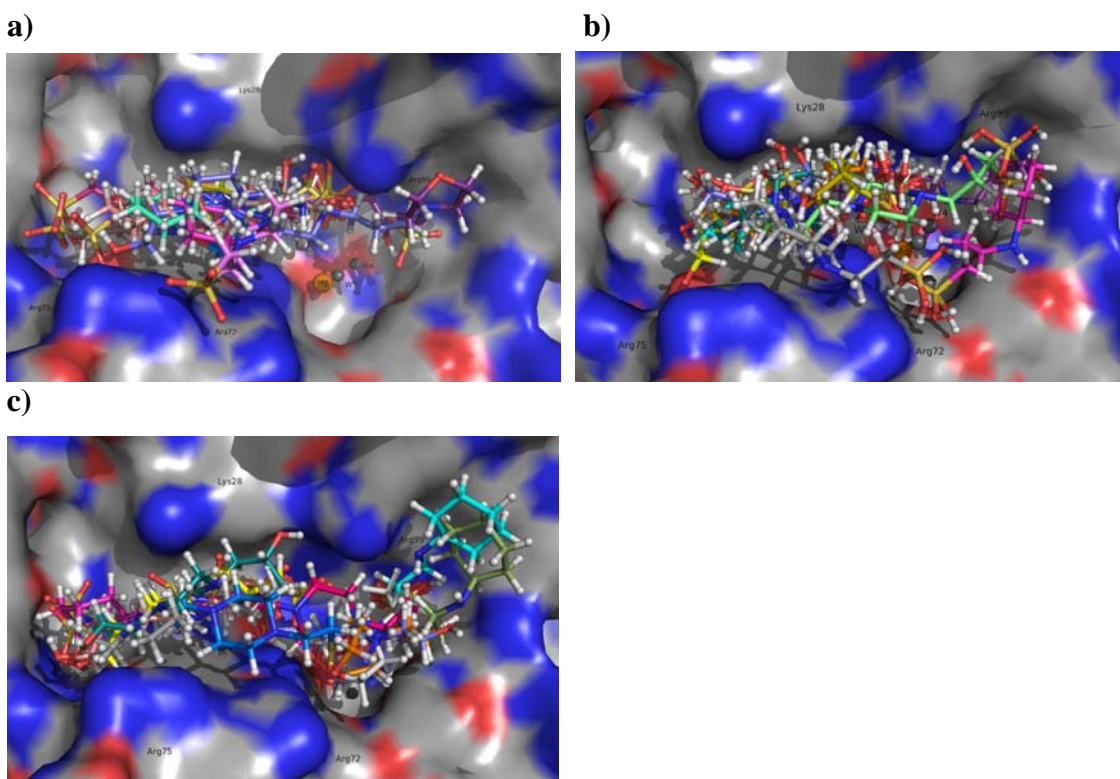


Figure 17. The active site with dockings of the IC_{50} determined compounds by a) GOLD, flexible receptor, b) AutoDock, rigid receptor setting and c) Glide, rigid receptor setting. Four water molecules were coordinated to the catalytic magnesium ion.

The best ranked poses of the active compounds Hepes, Ches and MES are similar for the programs GOLD and AutoDock, see figure 18 a) and b). The docking program Glide docks the top ranked poses in different positions, with the Hepes top ranked pose as the only one that resembles the poses in GOLD and AutoDock, see figure 18 c). The docking programs position of the inactive corresponding compounds with three carbon long side chains and the two carbon long chain, are assigned similar positions when superimposing the poses. The three pairs of active-inactive compounds are Hepes and EPPS, Ches and CAPS (in a deprotonated state) and MES and MOPS, see table 2 for small molecule structure comparison. The only exception of similar positions is observed for the compounds Hepes and EPPS docked by GOLD. The three programs seem to adapt one out of two strategies to solve the differences in length of the carbon chain containing the

sulfonic acid group (figures not shown). AutoDock consistently superimpose the sulfonic acid, the amine and the hetero- or homoatomic cyclohexane groups of the best ranked poses, and forces by this the three carbon chain to curve. The small molecule will as a consequence experience some extra strain. The other strategy is to allow the three carbons long chain to adapt a relaxed straight form, but displace the amine group by one bond length. As a succeeding result the hydrophobic cyclohexane is also displaced. This scenario of either strategy will change the interactions between the small molecule and the protein.

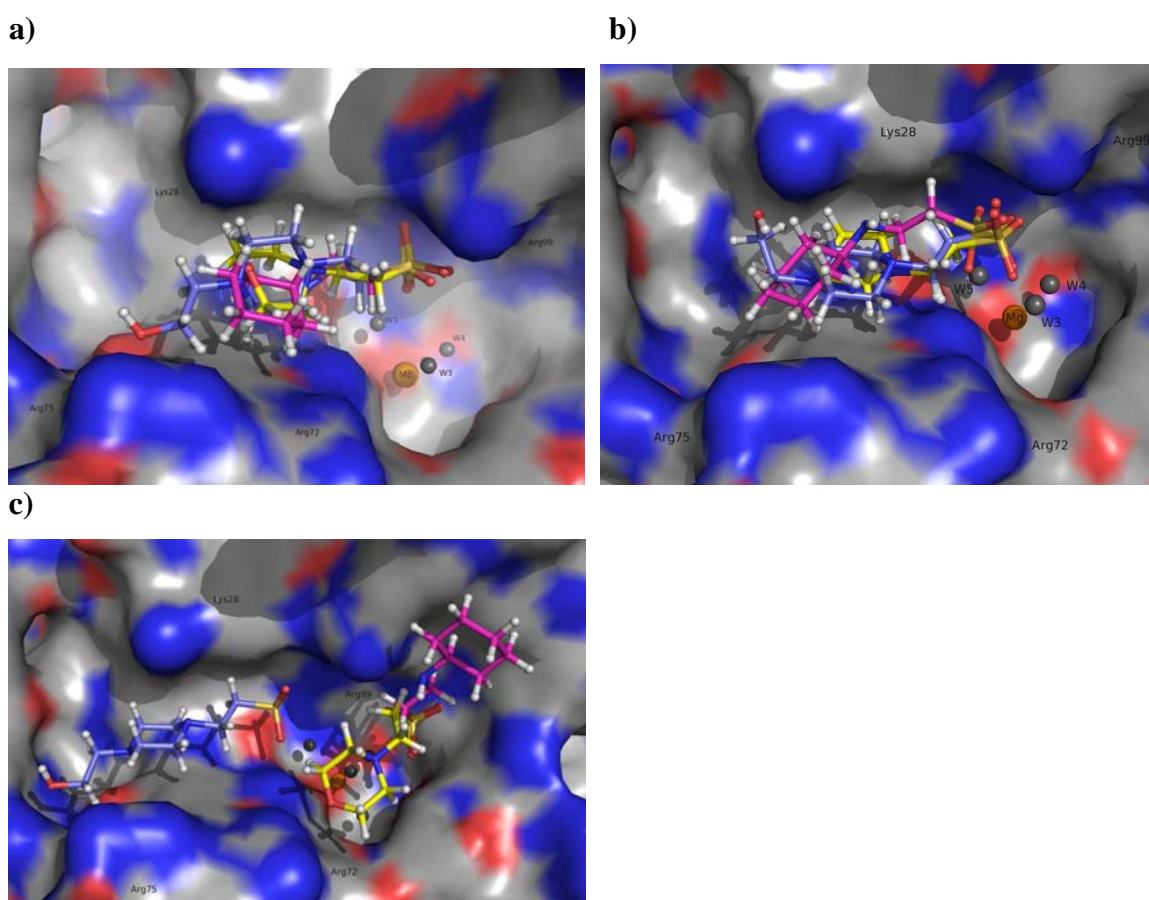


Figure 18. The active site containing the top ranked docking poses of the active compounds Hepes in blue, Ches in magenta and MES in yellow by a) GOLD, flexible receptor setting, b) AutoDock rigid receptor setting and c) Glide rigid receptor setting. All of the settings contain four water molecules coordinating to the catalytic magnesium ion.

A comparison of the experimental position of the cacodylate molecule with the best scored Hepes poses by each docking program is shown in figure 19, where the superimposition of the best ranked docked poses onto the experimental structure Hepes2 containing the cacodylate molecule are illustrated. Note that the surface is modeled from the Hepes2 structure and not from 2g7f as for figure 16-18. Hepes2 and Ches4 have an identical binding of the cacodylate molecule as seen in figure 13 a). The change of conformation observed in the side chain of Arg99 make the necessary space for the cacodylate molecule, as it otherwise would collide with the protein surface. The sulfonic acid group of the Hepes molecule is positioned close to the cacodylate molecule for the docking programs GOLD and AutoDock, where the distances from the sulphur to the arsenic ion is respectively 1.36 Å and 1.14 Å. For the sulfonic acid group in the best ranked docked pose of Hepes by Glide, the distance is in comparison 2.65 Å.

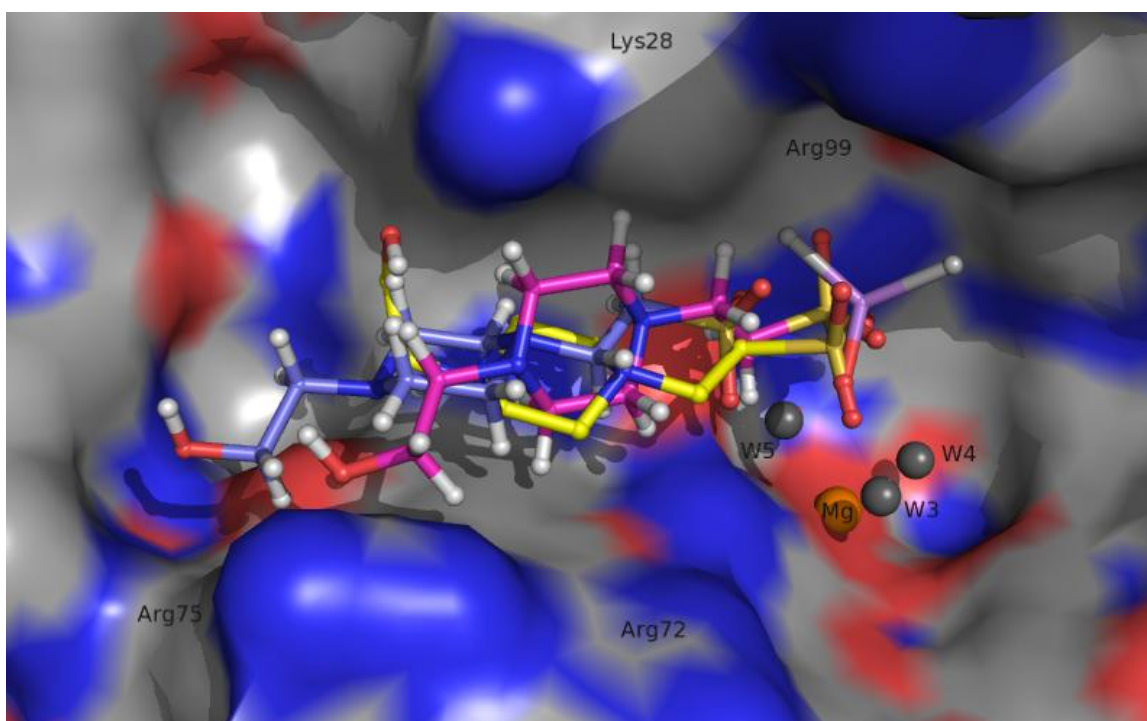


Figure 19. The superimposition of the best ranked docking poses of Hepes by the three programs GOLD in magenta, AutoDock in yellow and Glide in blue, onto the experimental structure of Hepes2 containing a bound cacodylate molecule colored by elements. Note that the surface is generated from the Hepes2 structure and not from the 2g7f as in figure 16-18.

The results from the VS, applying the best setting for the docking program GOLD, were analyzed based on the 100 highest ranked poses. For the library containing compounds with an aminoethanesulfonic acid group as a structural feature, the highest ranked compounds within top 10 were flexible systems containing a linear aminoethanesulfonic acid feature, and one or more aromatic ring with rotatable bonds in between. The flexibility and large surfaces of these compounds are most likely causing an unrealistic high fitness score. The best ranked 10 poses resembled the highest ranked Hepes pose docked with the same receptor setting. Within the top 100 poses, the vast majority of compounds had a close to identical position as the best ranked Hepes molecule pose, regarding similar superimposed features. Common for the compounds was the expansion of a heterocyclic penta- or hexane ring with a rotatable bond on the opposite end of the ring as the aminoethanesulfonic acid group. Overall the further expanding included an aromatic group, and a linear linkage was predominant. Residue Arg99 was observed as moderately flexible within the inspected top 100 poses. An overall superimposition of all of the 100 top ranked poses of library 1, reveals that the compounds are neatly docked into the DNA binding cleft of *VcEndA*.

For the virtual screening of small molecule library 2, the 100 best ranked poses were more scattered over a larger area of the active site, than observed for the 100 inspected poses of library 1 when visually inspected. The same tendencies of high ranked compounds consisting of aromatic fragments and rotatable bonds as for the VS with library 1 were observed. A handful of compounds were docked neatly in very similar poses as the best ranked Hepes pose, and are most likely the compounds that will have similar properties as the Hepes molecule.

Discussion

It seems that one of the structural features of Hepes that is responsible for the decrease in activity of *VcEndA*, is the aminoethanesulfonic acid group. The sulfonic acid part of this functional group may as well be replaced by a phosphate-group, as the experimental IC_{50} values for sulfate and phosphate are rather similar. This deduction is based on the fact that the aminoethanesulfonic acid group is present in the three active compounds Hepes, Ches and MES, that show additional positive interactions beyond the sulfonic acid group. A clear indication that strengthens this assumption is that the corresponding structures with three carbons between the sulfonic acid and amino group show some of the lowest inhibition effects observed by having very high IC_{50} values. Although these IC_{50} values have a high degree of uncertainty as the confidence intervals indicate, they show no further interactions that may be interpreted as a decrease in the activity of *VcEndA* whatsoever. The IC_{50} of Taurine, the compound only consisting of the aminoethanesulfonic acid group, is higher than the IC_{50} of the phosphate and sulfate and hence is interpreted as an inactive compound. This indicates that additional interaction with *VcEndA* beyond electrostatic interaction and the aminoethanesulfonic acid group is necessary to obtain a strong binding that may compete with the natural substrate. A second or third functional group is likely to be observed together with a strong inhibition effect. The three active compounds also have a hetero- or homoatomic cyclohexane as the next feature in their structures, and this hydrophobic feature may be the second necessary fragment in a possible lead compound. So far none of the compounds tested in the *in vitro* activity measurements achieved a better inhibition effect than the Hepes molecule with an IC_{50} value at 2140 nM. The IC_{50} value of Ches at 2198 nM is considered equal to Hepes, as the confidence intervals are broad enough to cover both values. The inhibition effect is not interpretable as more than a weak inhibition, but the binding suggests that there are additional interactions worth to study. An enhancement of this possible second, and maybe a third fragment properties, may have interactions that lead to a more promising lead compound. The third compound considered as active, is the common MES molecule with an IC_{50} value of 17 672 nM. This is a higher value than for Hepes and Ches, but as it indicates further positive interactions beyond the sulfonic acid group and is interpreted as

active. MES will also be applied as a validation compound if Hepes and Ches differ in their behavior.

The attempts to experimentally find the binding of the small molecules Hepes and Ches by crystallization and structure determination, was not succeeded. The co-crystallization approach using Hepes as buffer, gave crystals with the new space group, $P2_1$ containing four molecules in the asymmetric unit. No molecule was found bound in the active site of this data set, but a ring shaped density was observed in molecule C nearby residue Asn132. Asn132 is located in the outer area of the active site. A screenshot of this density is shown in appendix III, figure AIII-1. The density is not strong enough to give an uniform interpretation, but plausible molecules are either a non-specific bound Hepes molecule, glycerol or traces of PEG, as these were in the solutions applied in the experiment. The concentration of the buffer was 0.1 M, and should be more than sufficient to at least partly occupy the binding site if uniform binding occurs. The IC_{50} value of 2140 nM suggests that the interaction is strong enough for binding, although considered causing a weak inhibition effect. The lack of a bound Hepes in this structure may be a result of excluding of the Hepes molecule as the protein crystallizes. Another, but less reasonable, interpretation is that the Hepes molecule may not give a uniform binding mode if one considers the possible rotation around the bonds of the two rotatable carbon chains. If not uniformly bound, and if extremely disordered, the Hepes molecule would become *invisible* in the electron density maps. But if this was the case, at least some traces of densities of the hetero-1,4-diaminhexan ring should be observed in the maps. Most likely if following this scenario, also the trace density of the sulfonic acid group should be observed. The catalytic magnesium ion in the co-crystallized structure gives an interpretation of an active enzyme, and the active site of *VcEndA* is assumed to be the binding site of Hepes as the inhibition effect is competitive.

In the soaking experiments where 10 mM magnesium chloride was added in the soak and cryo solutions, the cacodylate molecule in the structures had a occupancy of 50 % in both the datasets Hepes2 and Ches4. The cacodylate was identified by its anomalous signal from hkl and $-|hkl|$, by not merging these in the scaling process. The anomalous signal of

arsenic is significant at the wavelength 0.91841 Å where the data sets are collected, whereas the signal from the possible other elements chloride and sulfur, are weak at this wavelength, see figure 20.

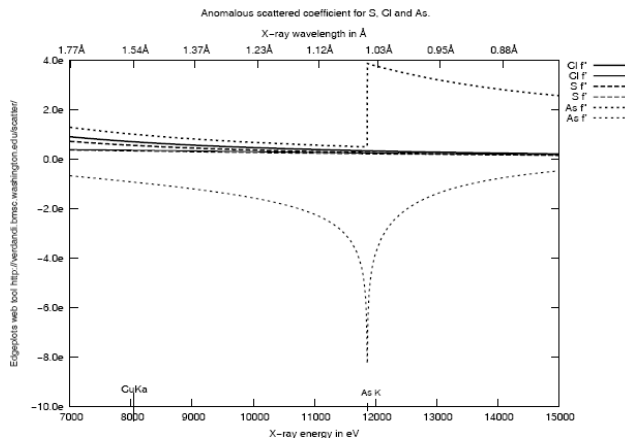


Figure 20. Calculated anomalous scattered coefficients for the elements sulfur, chloride and arsenic. Data were collected at a wavelength of 0.91814 Å, where arsenic is the only element that has a significant coefficient. The website used to calculate the coefficients was http://skuld.bmsc.washington.edu/scatter/AS_form.html 14.04.2008 .

When comparing the signal of the excess density of the observed small molecule with the sulfur in methionine and cysteine residues in the anomalous maps, the signal is significantly stronger than the sulfur signals. The possibility of a sulphate or phosphate instead of a cacodylate is not likely, due to that care had been taken not to expose the protein to particularly these compounds. Cacodylate is likely to have some of the same properties as phosphate, as arsenic is in the same group, just one period beneath phosphorous in the periodic table. If cacodylate is mimicking a phosphate or a sulfonic acid group, it is reasonable to believe that phosphorous and sulfur has a higher affinity and stronger interaction towards the catalytic magnesium ion by their higher electro-negative property. This is a speculative assumption, and the presence of a bound cacodylate may just be an artifact of the crystallization experiment, as buffer molecules are known to be seen in electron density maps. In the data sets where the cacodylate is found in the active site, the side chain of residue Arg99 is rotated away from the catalytic magnesium ion by making a bend by the Cδ. The conformation of Arg99 is identical in the electron density maps for both data sets, when compared with the structure of

VcEndA deposited in the PDB data bank as 2g7f by Altermark et al. (2006b), see figure 13. The change in conformation in this residue may not be caused by any other reason than lack of space and by Arg99 flexible property, as it is known to obtain different conformations (Li et al, 2003). Interestingly, it is the structures of *VcEndA* that holds both outer limits of determined conformations of Arg99.

The concentration of cacodylate at 0.1 M in the crystallization conditions of the soaking experiment outdo by 10 folds the Ches and Hepes concentrations at 10 mM. The activity measurement of cacodylate *versus* Hepes and Ches, shows a difference in the IC₅₀ values by 260 folds (2200:570 000). This indicates that the presence of the cacodylate molecule, and the lack of a Hepes or Ches molecule, is simply a competition won by outnumbering rather than a stronger affinity. The fact that the occupancy of a cacodylate molecule in the electron density maps is no higher than 0.5, also supports the hypothesis of a weak interaction. None the less is the cacodylate molecule found in the active sites, and shows a uniformly binding mode toward *VcEndA*.

The differences in number of water molecules in the crystal structures of Ches4, Hepes2 and Hepes4mol may be explained by the high resolution of 1.67 Å of the co-crystallized data set, and by radiation damage of the crystal soaked in Ches solution. The resolution the respective crystals diffracted to, would indicate that the Hepes4mol should contain the highest number of structural determined water molecules, as it have the highest resolution. Hepes2 have the lowest resolution, but obtain a significantly higher number of water molecules than in the comparable Ches4 structure. This may be an effect caused by a higher thermal disorder, as the crystal the Ches4 data set was collected from was exposed to twice the amount of radiation compared to the crystal the Hepes2 data set was collected from. It is possible that the crystal was beginning to suffer from radiation damages in the end of the data collection, and had started on the dishonorable path of a dying crystal.

The nature of the active site in *VcEndA* is highly hydrophilic, and is characterized by the catalytic Mg²⁺-water cluster providing a metal ion interacting with four water molecules,

and a network of several bridging water molecules. A high density of positively charged residues makes electrostatic interactions important, when binding modes are searched for. The genetic search algorithm in GOLD and the Lamarchian genetic algorithm in AutoDock, found similar top ranked poses for the active compounds, and mutually support each others results. As GOLD was able to assign similar fitness score to the active compounds, it outranks AutoDock in performance. The programs were given two different protonation states for three of the inactive compounds that were experimentally tested. GOLD treated two of these compounds, POPSO and PIPES, see table 2 for small molecule structures and table 7 for fitness scores, with protonation state -2 and -1 very similarly. POPSO and PIPES were given the best fitness for all of the four different scenarios given for the active site. GOLD also ranked the small compounds phosphate, sulphate and taurine together with the protonated state of CAPS, with the overall lowest fitness, see table 7. Obviously GOLD penalizes the lack of additional interactions, and neither exaggerate or understate the electrostatic contributions in the fitness function. This consistency is not observed either in AutoDock or Glide, see table 8 and 9. The docking program Glide performed poorest in processing the three active compounds by the criteria of similarity. As Glide, as well as GOLD, is given good reviews in comparing studies (Warren et al, 2006; Jain 2004) this was not expected. A reason for the poor performance may be that by attempting to avoid biasing of the docking experiment, and therefore applying a basic setting, the program needed more specific definitions than were given. In many cases information about binding modes is not available for a given system, and docking programs are therefore partly blindfolded in the docking search. A well behaved docking program should therefore be able to introduce more information than given as input. The more sophisticated options available in Glide may be an advantage for systems where large structural rearrangements occur upon binding, and more complex systems with known information available.

The precision levels in the docking experiments, together with a high probability that the best ranked pose is not the correctly identified position of the small molecules interaction *in vitro* and *in vivo*, make it difficult to draw any conclusive remarks regarding additional interactions than electrostatic contributions. Possible plausible interactions are between

the amine in the aminoethanesulfonic acid group and the guanidinium group of Arg72 and, or interaction with the amine in Lys28, or more unlikely the amide group in the Mg^{2+} -coordinating residue Asn127. The docking programs GOLD and AutoDock, positions the sulfonic acid group of Hepes close to the position of cacodylate molecule bound in the active site of the experimental structures Hepes2 and Ches4, see figure 19. The sulfonic acid group in the active compounds is likely to have some of the same interactions as the cacodylate molecule shown in figure 13 c) and d). These are a hydrogen bond to the N δ 1 of His80 and the two Mg^{2+} -coordinated water molecules labeled W3 and W5. In addition a possible interaction is between the sulfonic acid group and the guanidinium group of Arg99, but this interaction is dependent on how the flexibility of this residue acts out. If maximum observed flexibility is allowed, there are likely no interactions between the sulfonic acid group and Arg99, whereas if the binding mode of Arg99 has a low or moderately flexibility the possibility of this interaction will occur. In the docking experiment by GOLD using a flexible receptor setting, only a small degree in the shifting of Arg99 was observed compared to the conformation in the structures of Hepes2 and Ches4. This shows a very well behavior from AutoDock and GOLD, and credits the softening of the surface potential functions.

The virtual screening experiments indicate that compounds with linear arranged structures containing the aminoethanesulfonic acid group may be active. To evaluate the VS results, a selection of these compounds should be validated by *in vitro* activity measurements. Intuitively would the compounds that were docked in very similar poses as the Hepes molecule, be the best starting point in a possible next step of the project. Overall, more accurate analysis and docking of these compounds would benefit the project.

Concluding remarks

By experimentally determine IC_{50} values, three out of thirteen compounds were found to decrease the activity of *VcEndA*. The common features of these active compounds were the aminoethanesulfonic acid group, followed by a hydrophobic homo- or heteroatomic cyclohexane ring. None of the IC_{50} values were considered to indicate more than a weak inhibition effect. Although it did not succeeded to determine an experimental structure of a complex with an active compound bound, information about the behavior of the active site was gained. Two data sets collected from crystals in soak experiments contained the inactive cacodylate molecule in the active site. The binding of the cacodylate molecules were identical in the two structures, and provoked the same response by bending the out-reached conformation of residue Arg99 and making a small shift in Glu113, when compared with the native structures of the deposited 2g7f structure in the Protein Data Bank. By changing their conformation, Arg99 and Glu113 formed a salt bridge. A similar conformation of Arg99 is observed in the mutated His80Ala homologous Vvn-substrate complexes (Li et al, 2003; Wang et al 2007), but not to this extend. The new space group $P2_1$ was found when new crystallization conditions were applied in the attempts to co-crystallize the active compound Hepes with *VcEndA* by using the compound as crystallization buffer. These crystals contained four molecules in each asymmetric unit, and had no small compounds bound in the active site. The conformation of residue Arg99 is identical for all four molecules, and resembles the conformation in the deposited structure 2g7f. The change in Arg99 may indicate that by trapping the Arg99 residue in a substrate binding mode, inhibition is possible by a compound with higher affinity toward *VcEndA*. An interesting feature in the two structures containing a cacodylate molecule, is the replacement of the activated nucleophilic water molecule proposed to have an important catalytic property in the suggested mechanism by Li et al. (2003). As cacodylate do not have any observed decreasing effect in the activity on *VcEndA*, no final conclusion can be drawn at this point regarding if this is the mechanism of inhibition by the binding of an active compound. The docking program GOLD and AutoDock performs the most reliable modeling of the *VcEndA* system, as they were able to dock the active compounds into similar poses in the active site pocket. GOLD outranks

AutoDock in performance by also being able to differ between active and inactive compounds by its fitness function. The virtual screening of the library containing compounds with an aminoethanesulfonic acid feature, strengthens the consensus results of the experimental IC_{50} determination, by the observation of predominantly linear aminoethanesulfonic acid groups presence in the highest ranked 100 poses. Of the top 100 poses from the VS of the smallest library, the vast majority resembles the best ranked Hepes pose upon superimposing. The binding mode in an experimentally Hepes-*Vc*EndA complex is therefore likely to be similar to the best ranked Hepes pose.

Further work & Development

The overall aim of this study was to find a lead compound for an inhibitor that is usable in a commercial kit. This kit should be designed to block a variety of extracellular and periplasmic nucleases from different organisms. If this project is further developed to reach this aim, further experiments would have to be performed to verify or invalidate the findings in this thesis. An experimental binding between an active compound and *VcEndA* is necessary to address the correct interactions, and to validate computational docking experiments. As both soaking experiment and co-crystallization under the reported conditions did not succeed, new crystallization conditions should be screened for using crystallization robotics. One of the observed properties of the crystal structures solved in this thesis, were the dense packing in the crystals. If crystals with a more loose packing structure are found, soaking with Hepes or Ches may result in an active small molecule-*VcEndA* complex. To verify the docking programs assigned affinity for the active compounds, ITC (Isothermal Titration Calorimetry) may be performed to find experimental binding energies. This will pinpoint any systematic over- or underestimation of binding energy, and be a guide in designing a better setting for more accurate experiments. More sophisticated docking experiments may be performed by adding information and tuning the computational settings. This would likely demand more computer resources than applied in the reported experiments. The results from the VS would be a starting point in the exploration of a better and maybe more novel lead compound than Hepes and Ches. A selection of compounds from these screenings should be tested in the *in vitro* activity assay. If a promising lead compound is found, the inhibition effect should be tested at other endonucleases of type I, to confirm or invalidate that it inhibits a numerous nucleases within a certain range of homology.

References

- Altermark, B., Moe, E., Willassen, N.P. and Smalås, A.O. (2006a): Comparative studies of endonuclease I from *Vibrio cholerae* and *Vibrio salmonicida*. Ph.D. Thesis, University of Tromsø.
- Altermark, B., Willassen, N.P., Smalås, A.O and Moe, E. (2007b): Comparative studies of endonuclease I from cold-adapted *Vibrio salmonicida* and mesophilic *Vibrio cholerae*. FEBS Journal, **274**, 252-263.
- Altermark, B., Helland, R., Moe, E., Willasesn, N.P., and Smalås, A.O. (2006) IV: Environmental adaption of endonuclease I from *Vibrio salmonicida*. Ph.D Thesis University of Tromsø, Norway.
- Altermark, B., Smalås, A.O, Willassen, N.P., and Helland, R. (2006b): The structure of *Vibrio cholerae* extracellular endonuclease I reveals the presence of a buried chloride ion. Acta Crystallography D Biological Crystallography, **62**, 1387-1391.
- Barbosa, F. and Horvath, D. (2004): Molecular similarity and property similarity. Current Topics in Medicinal Chemistry, **Vol 4**, 589-600.
- Chang, M. C., Chang, S. Y., Chen, S. L. and Chuang, S. M. (1992): Cloning and expression in *Escherichia coli* of the gene encoding an extracellular deoxyribonuclease (DNase) from *Aeromonas hydrophila*. Gene, **122**, 175-180.
- Chen, D., Menche, G., Power, T.D., Sower, L., Peterson, J.W. and Schein, C.H. (2007): Accounting for ligand-bound metal ions in docking small molecules on adenyl cyclase toxins. PROTEINS, **Vol 67**, 593-605.
- Collaborative Computational Project, Number 4. (1994): The CCP4 Suite: Programs of protein crystallography. Acta Crystallography D Biological Crystallography, **50**, 760-763.
- Evans, J.N.S. (1995): Biomolecular NMR spectroscopy. Book, Oxford University Press, chapter 4, p147-204.
- Focareta, T. and Manning, P. A. (1987): Extracellular proteins of *Vibrio cholerae*: molecular cloning, nucleotide sequence and characterization of the deoxyribonuclease (DNase) together with its periplasmic localization on *Escherichia coli* K-12. Gene, **53**, 31-40.
- Focareta, T. and Manning, P. A. (1991): Distinguishing between the extracellular DNases of *Vibrio cholerae* and development of a transformation system. Molecular Microbiology, **Vol 5**, No 10, 2547-2555.
- Friedhoff, P., Franke, I., Krause, K.L. and Pingoud, A. (1999): Cleavage experiments with deoxythymidine 3',5'-bis-(*p*-nitrophenyl phosphate) suggest that the homing endonuclease I-PpoI follows the same mechanism of phosphodiester bond hydrolysis as the non-specific *Serratia* nuclease. FEBS Letters, **443**, 209-214.
- Friedhoff, P., Kolmes, B., Gimadutdinov, O., Wende, W., Krause, K.L. and Pingoud, A.. (1996): Analysis of the mechanism of the *Serratia* nuclease using site-directed mutagenesis. Nucleic Acids Research, **Vol 24**, No 14, 2632-2639.
- Galburt, E.A., Chevalier, B., Tang, W., Jurica, M.S., Flick, K.E., Monnat, R.J. Jr, and Stoddard, B.L. (1999): A novel endonuclease mechanism directly visualized for I-PpoI. Nature structural biology, **Vol 6**, No 12, 1096-1099.
- Galburt, E.A. and Stoddard, B.L. (2002): Catalytic mechanisms of restriction and homing endonucleases. Biochemistry, **Vol 41**, No 47, 13851-13860.

Halperin, I., Ma, Buyong, Wolfson, H. and Nussinov, R. (2002): Principles of docking: An overview of search algorithms and a guide to scoring functions. *PROTEINS:Structure, Functions, and Genetics*, **Vol 47**, 409-443.

Helbæk, M. (1999): Fysikalsk kjemi. Book, Fagbokforlaget, chapter 11.8, 579-598.

Hochhut, B., Marrero, J. and Waldor, M.K. (2000): Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a constin found in *vibrio cholerae* O139. *Journal of Bacteriology*, April, **Vol 182**, No 7, 2043-2047.

Huey, R., Morris, G.M., Olson, A.J. and Goodsell, D.S. (2007): Software news and update A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry*, **Vol 28**, 1145-1152.

Ichige, A., Matsutani, S., Oishi, K. and Mizushima, S. (1989): Establishment of gene transfer systems for and construction of the genetic map of a marine *Vibrio* strain. *Journal of Bacteriology*, April, **Vol 171**, No 4, 1825-1834.

Irwin, J.J. and Shoichet, B.K. (2005): A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modelling*. **Vol 45**, 177-182.

Jain, A.N. (2004): Virtual screening in lead discovery and optimization. *Current Opinion in Drug Discovery*, Vol 7, No 4, 396-403.

Jiang, S.C. and Paul, J.P. (1998): Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology*, August, **Vol 64**, No 8, 2780-2787.

Jekel, M. and Wackernagel, W. (1995): The periplasmic endonuclease I of *Esterichia coli* has amino-acid sequence homology to the extracellular DNases of *Vibrio cholerae* and *Aeromonas hydrophila*. *Gene*, **154**, 55-59.

Jones, G., Willet, P. and Glen, R.C. (1995): Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*. **Vol 245**, 43-53.

Jones, T.A., Zou, J.-Y. and Cowan, S.W. (1991): Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallography Section A*, **A47**, 110-119.

Kabsch, W. (1976): A solution for the best rotation to relate two sets of vectors. *Acta Crystallography Section A*, **A32**, 922-923.

Kabsch, W. and Sander, C. (1983): Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. December **Vol 22**, Issue 12, 2577-2637.

Kabsch, W. (1993): Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography*, **26**, 795-800.

Kitchen, D.B., Decornez, H., Furr, J.R. and Bajorath, J. (2004): Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, **Vol 3**, Nov, 935-949.

Kaper, J. B., Morris Jr., J. G. and Levine, M. L. (1995): Cholera. *Clinical Microbiology Reviews*, Jan. **Vol 8**, Nr 1, 48-86.

Laskowski, R.A, MacArthur, M.W., Moss, D. S. and Thornton, J.M. (1993): PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **Vol 26**, 283-291.

Li, C.-L., Hor, L.-I., Chang, Z.-F., Tsai, L.-C., Yang, W.-Z. and Yuan, H.S. (2003): DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. *The EMBO Journal*, **22**, No 15, 4014-4025.

Luscombe, N.M. and Thornton, J. (2002): Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*, **Vol 320**, 991-1009.

Lyne, P. D. (2002): Structure-based virtual screening: an overview. *Drug Discovery Today*, **Vol 7**, No 20 October, 1047-1055.

McCoy, A., Grosse-Kunstleve., Storoni, L.C. and Read, R.J. (2005): Likelihood-enhanced fast translation function. *Acta Crystallographica Section D*, **Vol 61**, 458-464.

Miller, M.D., Cai, J. and Krause, K.L. (1999): The active site of *Serratia* endonuclease contains a conserved magnesium-water cluster. *Journal of Molecular Biology*, **Vol 288**, 975-987.

Miller, M.C., Keymer, D. P., Avelar, A., Boehm, A. B. and Schoolnik, G. K. (2007): Detection and transformation of genome segments that differ within a coastal population of *Vibrio cholerae* strains. *Applied and Environmental Microbiology*, June, **Vol 73**, No 11, 3695-3704.

Miller, M.D., Tanner, J., Alpaugh, M., Benedik, M.J. and Krause, K.L. (1994): 2.1 Å structure of *Serratia* endonuclease suggests a mechanism for binding to double-stranded DNA. *Nature Structural Biology*, **Vol 1**, No7, 461-468.

Mishra, N.C. (2002): Nucleases- Molecular biology and application. Book, John Wiley & Sons, Inc., Hoboken, New Jersey, chapter 1, 1-26, chapter 3, 53-6.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. (1998): Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, **Vol 19**, No 14, 1639-1662.

Moulard, M., Condemine, G. and Robert-Baudouy, J. (1993): Characterization of the *nucM* gene coding for a nuclease of the phytopathogenic bacteria *Erwinia chrysanthemi*. *Molecular Microbiology*, **Vol 8**, No 4, 685-695.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997): Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D*, **Vol 53**, 240-255.

Panda, D. K., Dasgupta, U. and Das, J. (1991): Transformation of *Vibrio cholerae* by plasmid DNA. *Gene*, **Vol 105**, 107-111.

Raaijmakers, H., Vix, O., Törö, I., Golz, S., Kemper, B. and Suck, D. (1999): X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapping dimer architecture. *The EMBO Journal*, **Vol 18**, 1447-1458.

Singh, D. V., Matte, M. H., Matte, G.R., Jiang, S., Sabeena, F., Shukla, B.N., Sanyal, S.C., Huq, A. and Colwell, R.R. (2001): Molecular analysis of *Vibrio cholerae* O1, O139, non-O1, and non O139 strains: Clonal Relationships between clinical and environmental isolates. *Applied and Environmental Microbiology*, Feb. **Vol 67**, No 2, 910-921.

Taylor, R.D., Jewsbury, P.J. and Essex, J.W. (2002): A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, **Vol 16**, 151-166.

Timmins, K. and Winkler, U. (1973): Isolation of covalently closed circular deoxyribonucleic acid from bacteria which produce exocellular nuclease. *Journal of Bacteriology*, Jan, **Vol 113**, No 1, 508-509.

- URL: <http://www.who.int/topics/cholera/en/> 21.01.2008. The World Health Organization, WHO, website for cholera pandemics and epidemics.
- URL: <http://www.ambion.com/catalog/ProdGrp.html?fkApp=12&fkProdGrp=232> 31.05.2007. Manual and description of DNaseAlert™ QC System, Ambition USA
- URL: http://skuld.bmsc.washington.edu/scatter/AS_form.html 26.03.2008. The web site for calculation of anomalous scattering coefficients for elements of choice.
- URL: <http://www.ccp4.ac.uk/dist/html/INDEX.html> 26.03.2008 Program documentation site of the programs available in the CCP4i package.
- URL: <http://zinc.docking.org> 29.04.2008 The web site of the ZINC small molecule database, version 7.
- URL: <http://xray.bmc.uu.se/hicup/> 29.04.2008 The web site of the Hetero-compound Information Centre-Uppsala, release 12.1.
- URL: http://en.wikipedia.org/wiki/Vibrio_cholerae 08.05.2008 The web site of the free encyclopedia Wikipedia for *Vibrio cholerae*.
- Vagin, A. and Teplyakov, A. (1997): MOLREP: An automated program for molecular replacement. Journal of Applied Crystallography, **Vol 30**, 1022-1025.
- Vriend, G. (1990): WHAT IF: A molecular modeling and drug design program. Journal of Molecular Graphics, **Vol 8**, march, 52-56.
- Wang, Y.-T., Yang, W.-J, Li, C.-L., Doudeva, L. G. and Yuan, H. S. (2007): Structural basis for sequence-dependent DNA cleavage by nonspecific endonucleases. Nucleic Acids Research, **Vol 35**, No 2, 584-594.
- Warren, G.L, Andrews, C.W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E. and Head, M.S. (2006): A critical assessment of docking programs and scoring functions. Journal of Medical Chemistry, **Vol 49**, 5912-5931.
- Westheimer, F.H. (1987): Why nature chose phosphates. Science, **Vol 235**, March 6, 1173-1178.
- Wu, S.-I., Lo, S.-K, Shao, C.-P., Tsai, H.-W. and Hor, L.-I. (2001): Cloning and characterization of a periplasmic nuclease of *Vibrio vulnificus* and its role in preventing uptake of foreign DNA. Applied and Environmental Microbiology, January, **Vol 67**, No 1, 82-88.

List of appendixes

Appendix I: Ramachandran plots from PROCHECK for the structures of Hepes2, Ches4 and Hepes4mol.

Appendix II: Root-mean-square deviations of superimposition of the structures of Hepes2, Ches4 and Hepes4mol generated by using the LSQKAB program.

Appendix III: Picture of unexplained ring shaped density in molecule C of Hepes4mol.

Appendix IV: Pair wise comparison of distances of equivalent atom coordinates of residue Arg99 and Glu113 in the structures of Hepes2 and Ches4.

Appendix I: Ramachandran plots from PROCHECK for the structures of Hepes2, Ches4 and Hepes4mol.

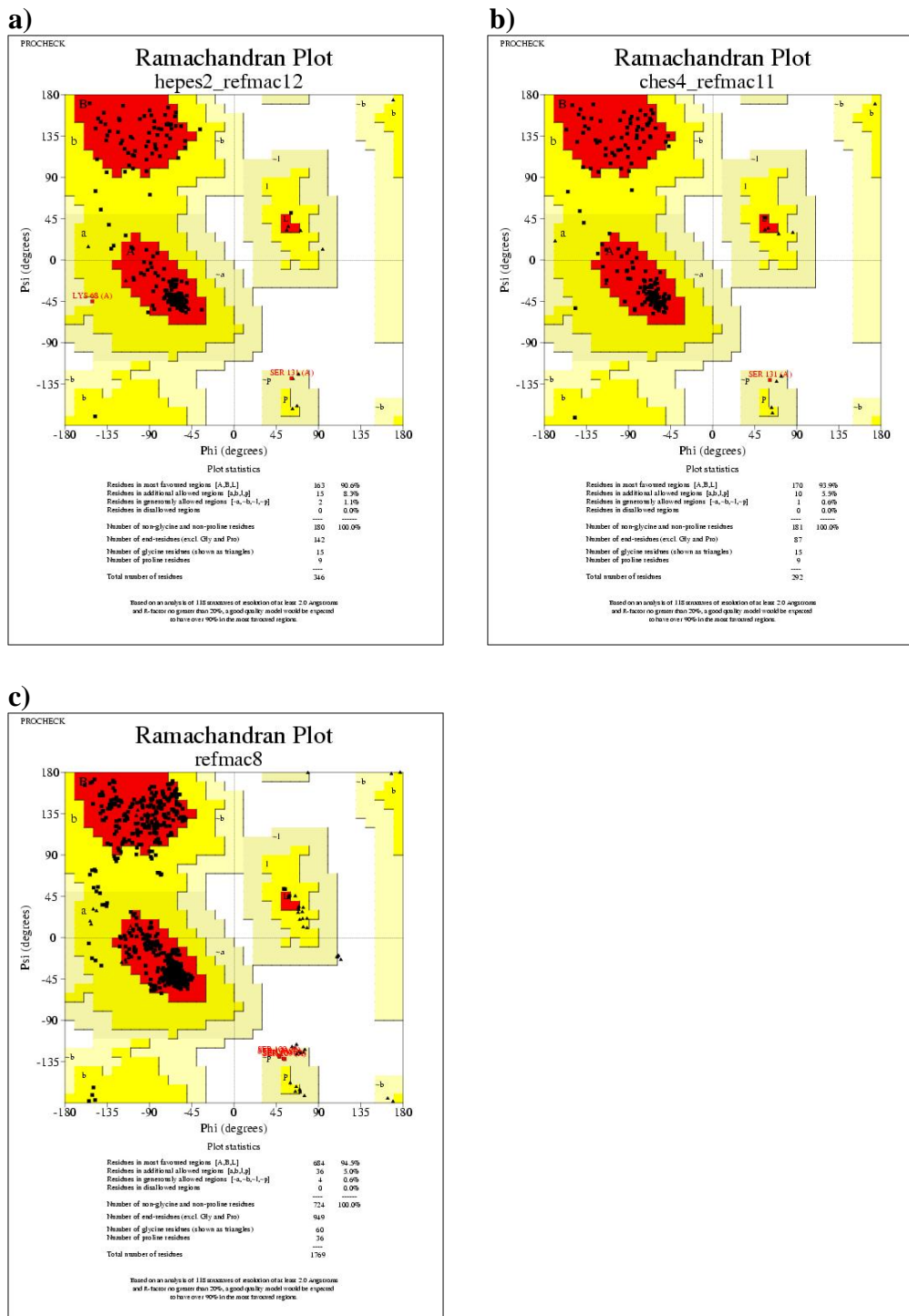


Figure AI-1. The Ramachandran plots for the experimental determined structures Hepes2 a), Ches4 b) and Hepes4mol c) generated by using the program PROCHECK.

Appendix II: Root-mean-square deviations of superimposition of the structures of Hepes2, Ches4 and Hepes4mol generated by using the LSQKAB program.

Hepes2 on Ches4.

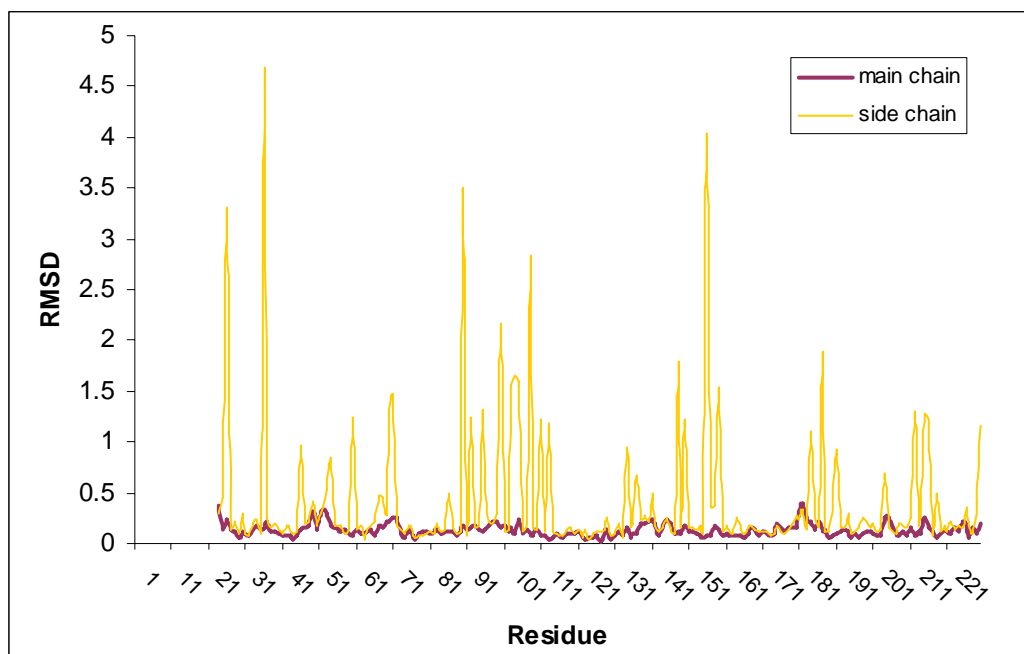
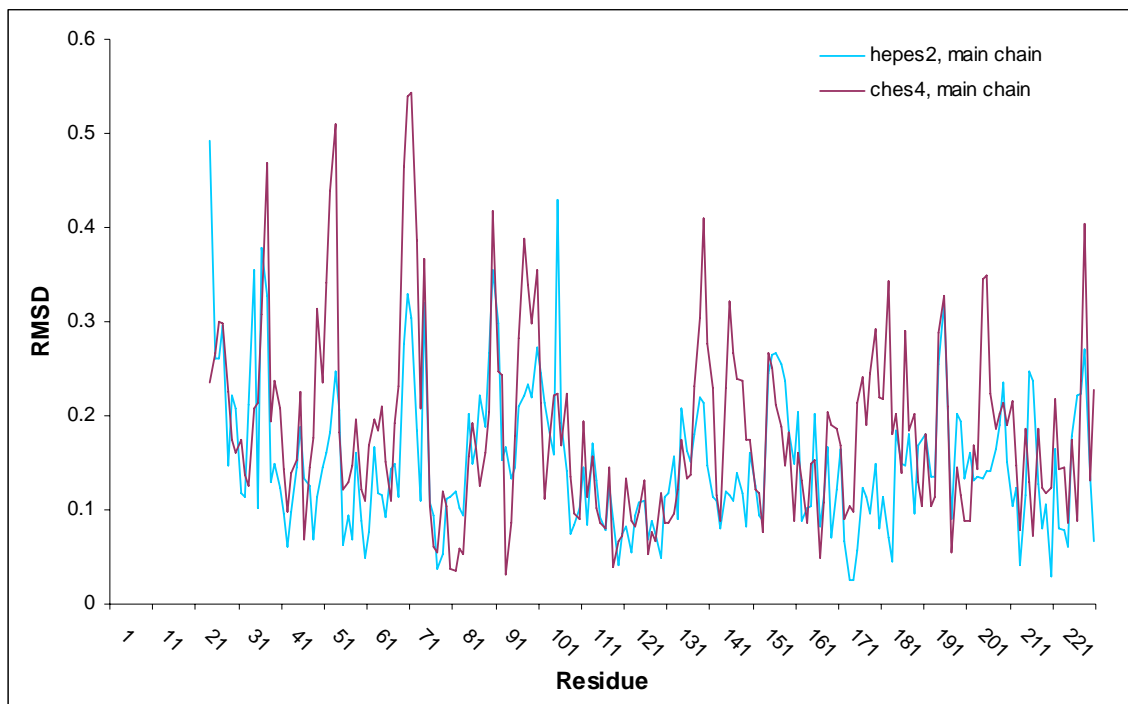


Figure AII-1. Rmsd values calculated from the superimposition of the structure of Hepes2 onto the structure of Ches4. The side chains rmsd values in Å are marked by a yellow line, the main chain rmsd values by a thick purple line.

Hepes2 and Ches4 on 2G7F

a)



b)

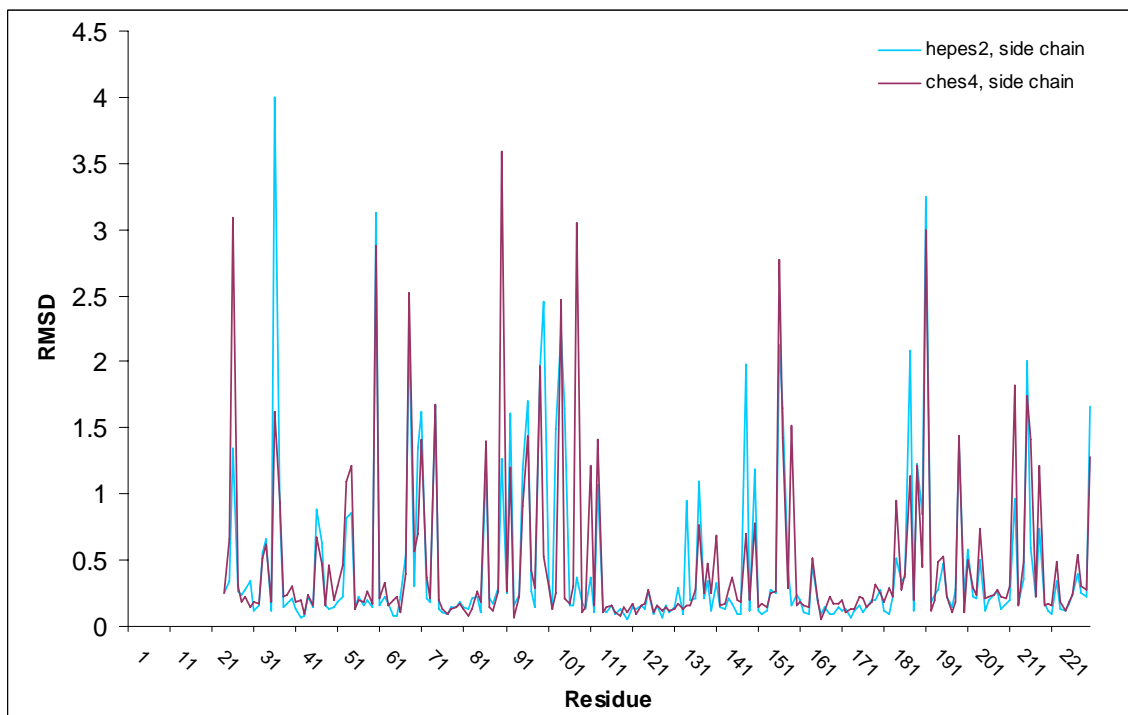
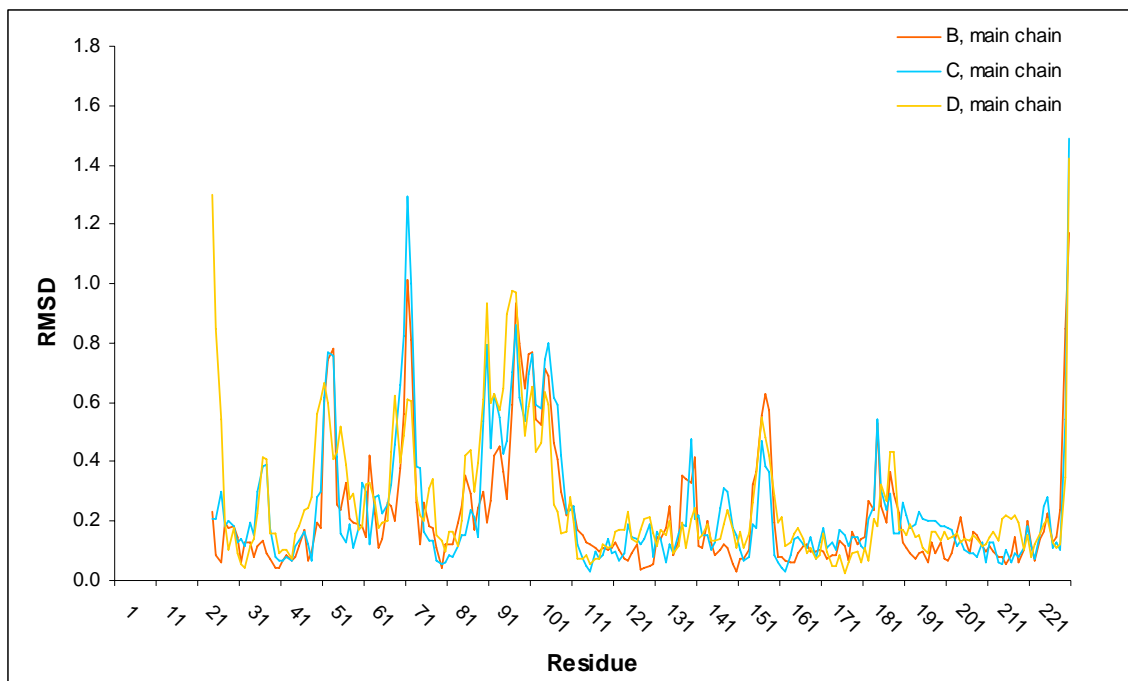


Figure AII-2. Rmsd values calculated from the superimposition of the structure of Hepes2 in blue and Ches4 in purple, onto the deposited structure 2g7f. Figure a) show the rmsd values of the main chain and figure b) the rmsd values of the side chains.

Hepes4mol

Molecule A

a)



b)

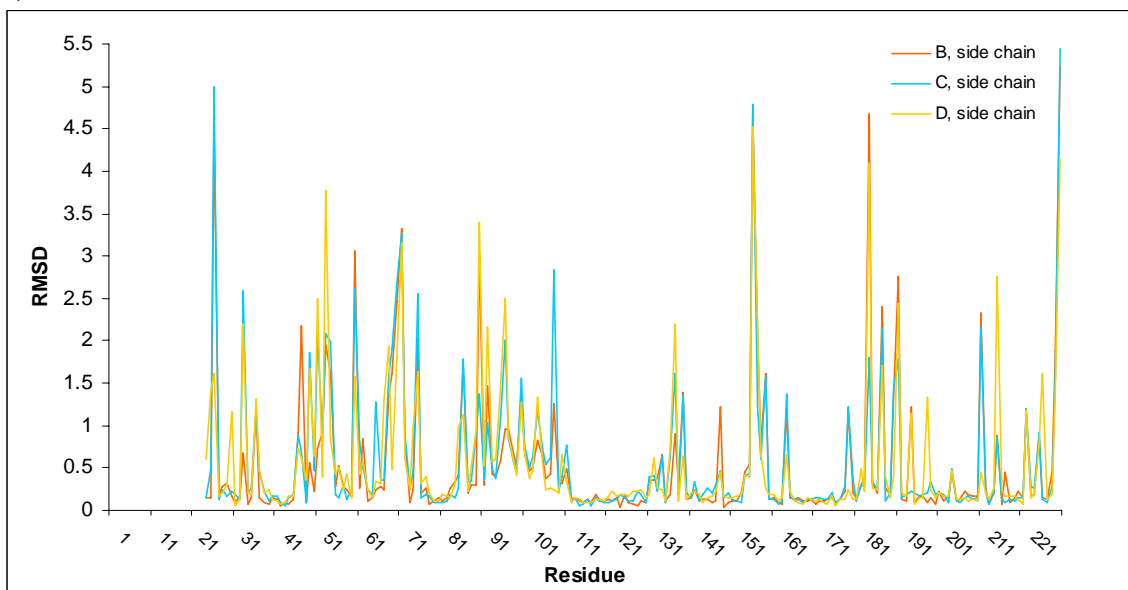
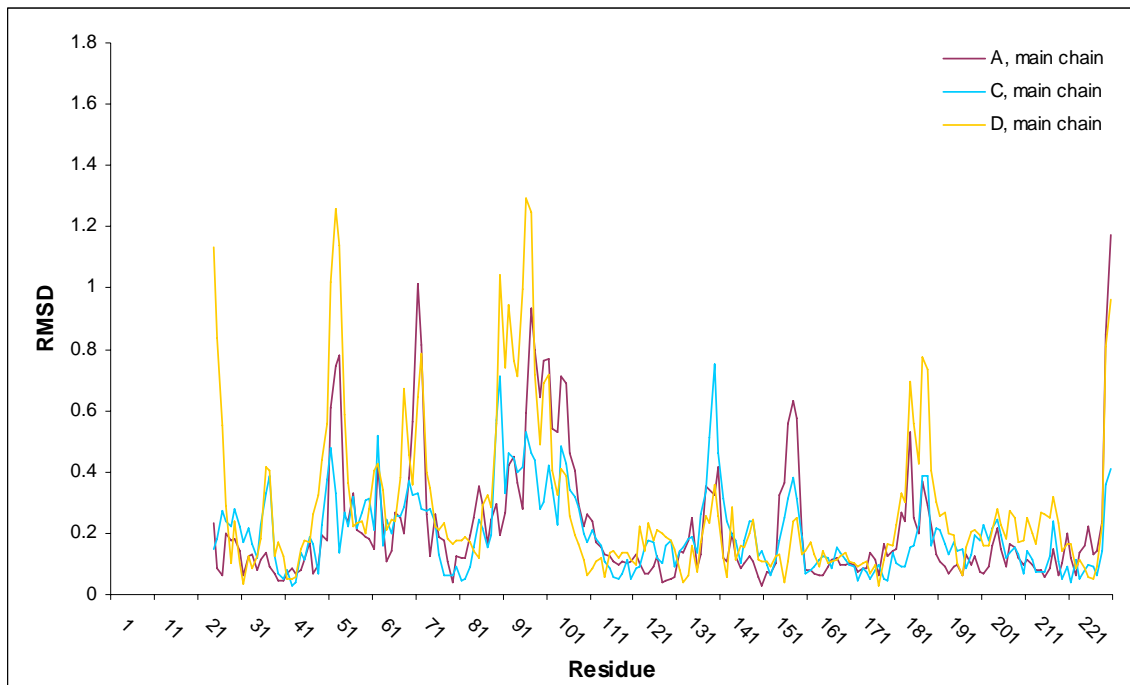


Figure AII-3. Rmsd values calculated from the superimposition of molecule B in red, C in blue and D in yellow, onto molecule A of the Hepes4mol structure. The rmsd values for the main chains are given in figure a) and the side chains rmsd values in figure b).

Molecule B

a)



b)

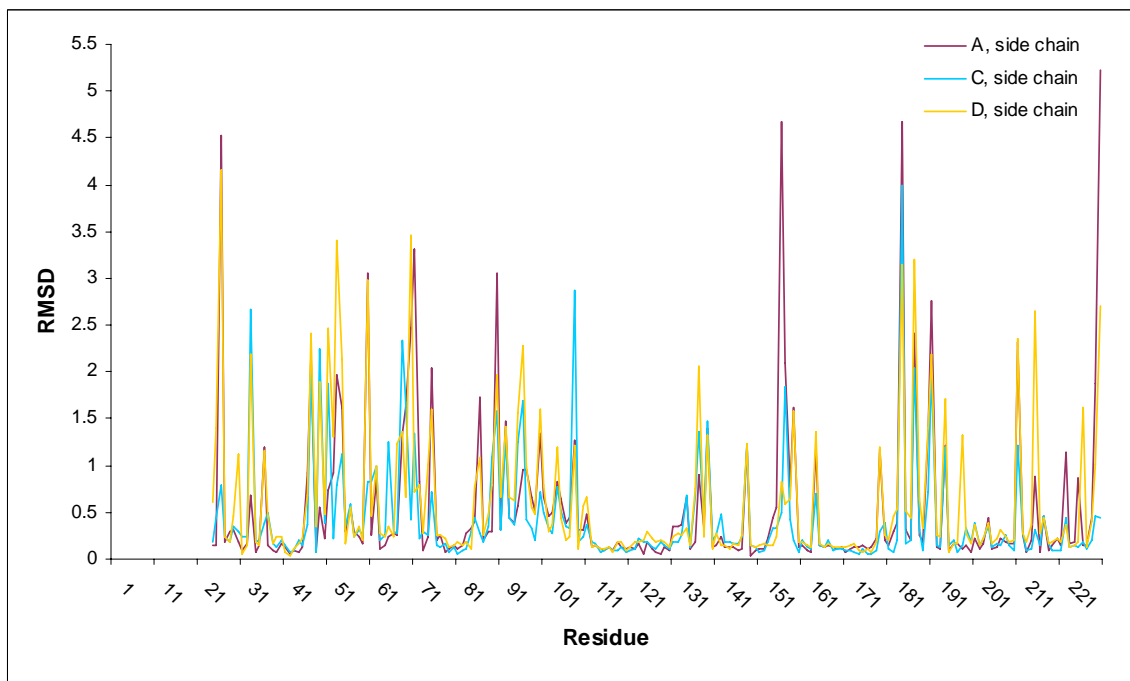
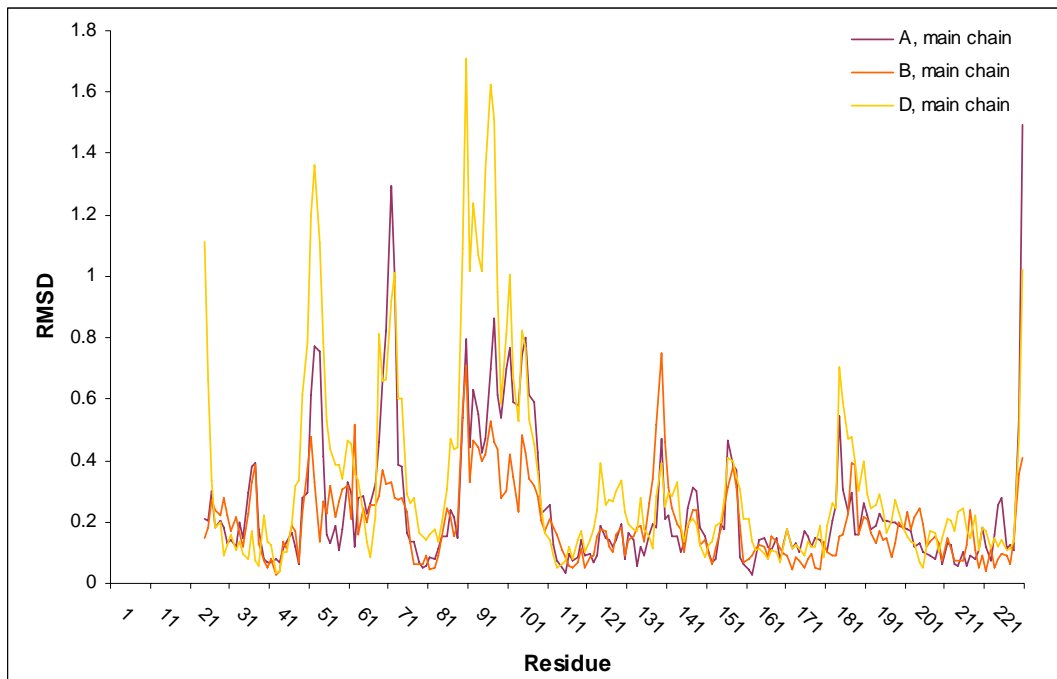


Figure AII-4. Rmsd values calculated from the superimposition of molecule A in purple, C in blue and D in yellow onto molecule B of the Hepes4mol structure. The rmsd values for the main chains are given in figure a) and the side chains rmsd values in figure b).

Molecule C

a)



b)

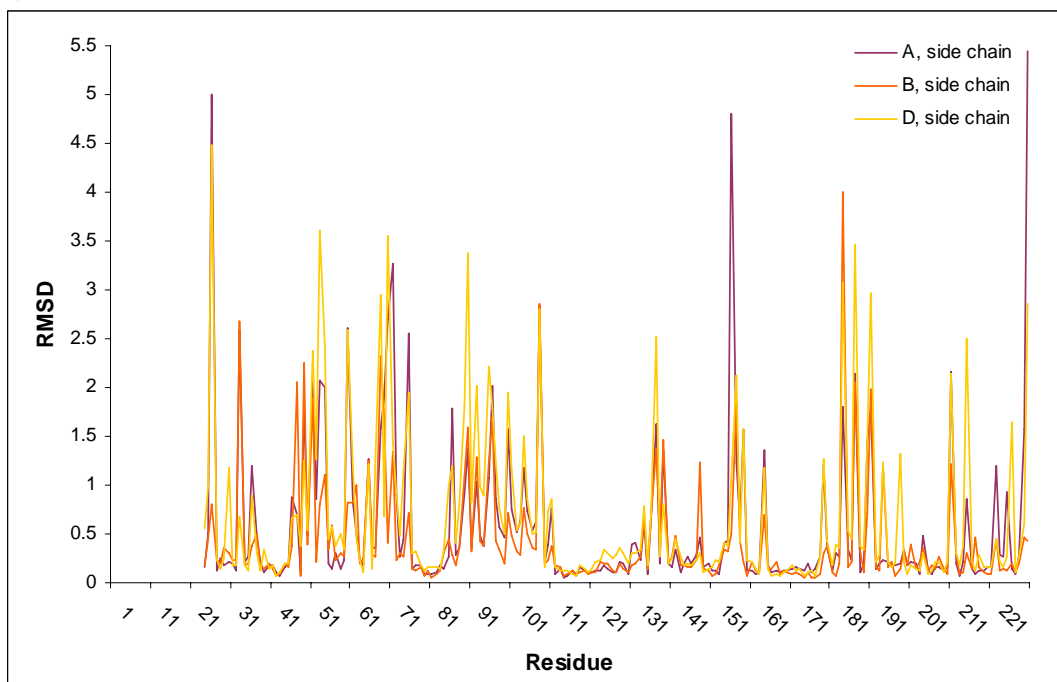
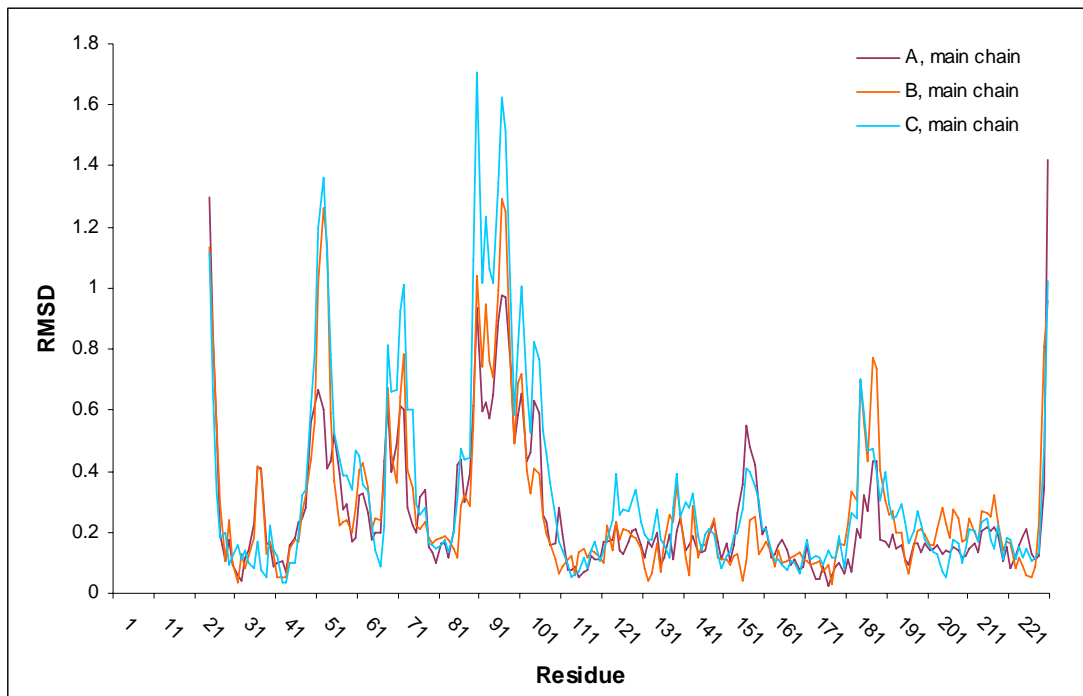


Figure AII-5. Rmsd values calculated from the superimposition of molecule A in purple, B in red and D in yellow onto molecule C of the Hepes4mol structure. The rmsd values for the main chains are given in figure a) and the side chains rmsd values in figure b).

Molecule D

a)



b)

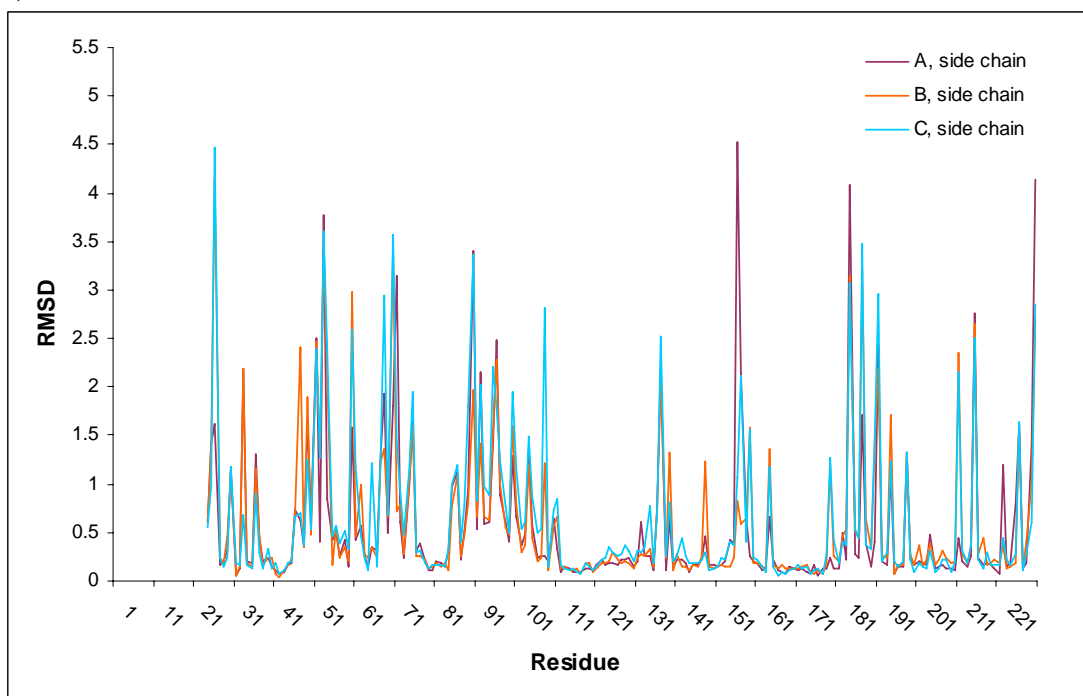


Figure AII-6. Rmsd values calculated from the superimposition of molecule A in purple, B in red and C in blue onto molecule D of the Hepes4mol structure. The rmsd values for the main chains are given in figure a) and the side chains rmsd values in figure b).

Superimposition of the main chain of the structure of Hepes2, Ches4 and Hepes4mol onto the deposited 2g7f structure of VcEndA.

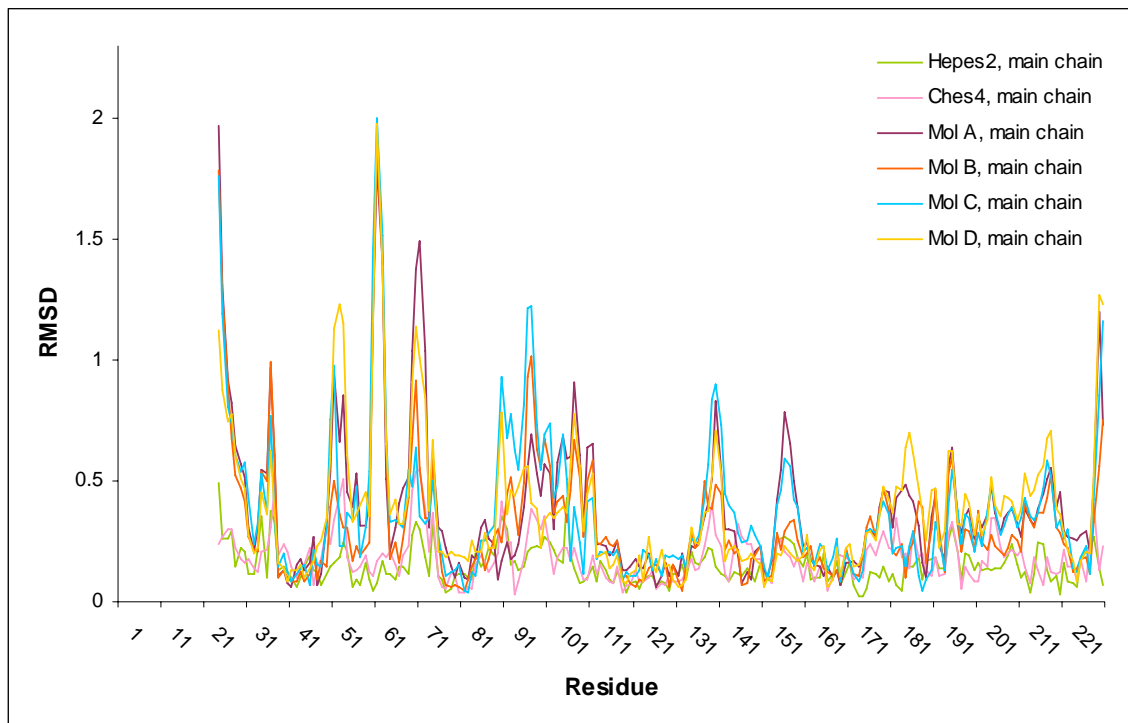


Figure AII-7. Rmsd values calculated from the superimposition of the main chain of the Hepes2 in green, Ches4 in pink and molecule A in purple, molecule B in red, molecule C in blue and molecule D in yellow of the Hepes4mol structures onto the deposited structure 2g7f of VcEndA in the Brookhaven Protein Data Bank.

Table AII-1. Root mean square deviation (RMSD) values for superimposition of molecule A-D in the Hepes4mol dataset. RMSD values for main chain atoms are above the diagonal, and RMSD values for all atoms are below.

	Mol A	Mol B	Mol C	Mol D
Mol A		0.217	0.243	0.258
Mol B	0.984		0.199	0.282
Mol C	0.963	0.681		0.332
Mol D	0.928	0.927	1.006	

Appendix III: Picture of unexplained ring shaped density in molecule C of Hepes4mol.

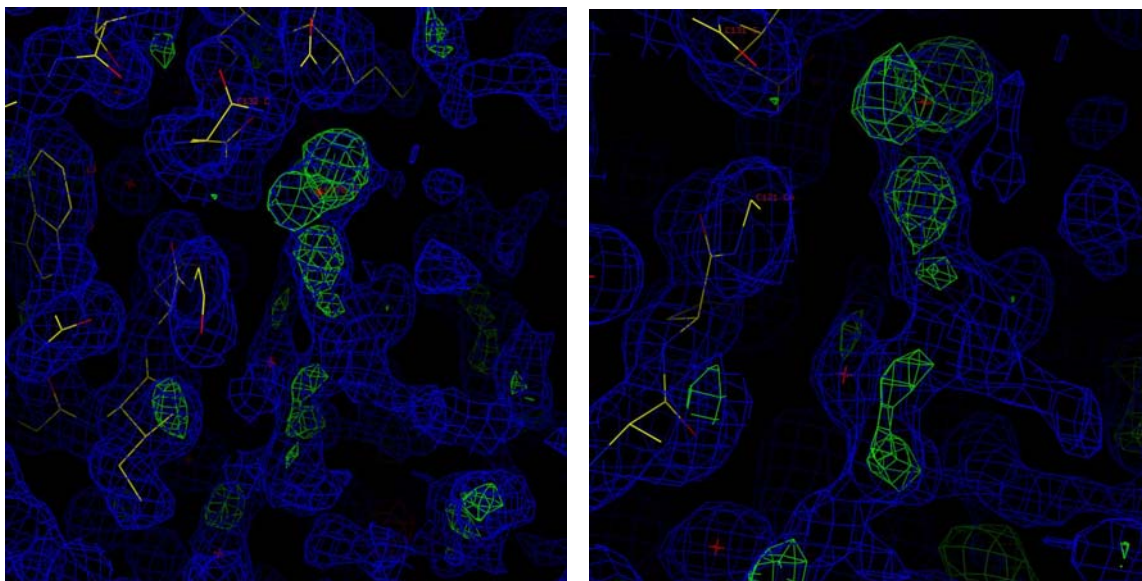


Figure AIII-1. Screenshot taken from the graphical program *O* shows a ring-shaped formation near residue 132 in molecule C in the map of Hepes4mol. The contrasts for the $2mF_o-DF_c$ density maps were reduced to 0.5 and 2.5 compared with normal values of 1.0 and 3.0.

Appendix IV: Pair wise comparison of distances of equivalent atom coordinates of residue Arg99 and Glu113 in the structures of Hepes2 and Ches4.

Table AIV-1. Pair wise comparison of equivalent atom coordinates in Å of the shifted residues Arg99 and Glu113 in the active site of Hepes2 and Ches4 regarding the deposited structure 2g7f.

	Arg99			Glu113		
	CZ	NH1	NH2	CD	OE1	OE2
Hepes2	2.65 Å	4.18 Å	3.06 Å	1.18 Å	1.67 Å	1.85 Å
Ches4	2.57 Å	4.24 Å	3.05 Å	1.03 Å	1.51 Å	1.71 Å

