PERSONA CIÈNCIA EMPRESA

Universitat Ramon Llull

TRABAJO DE FIN DE GRADO
(Grado en Química)

Título : Prediction of specificity in chitin deacetylases using a virtual docking protocol

Realizado por Juan Eiros

Dirigido por Dr. Xevi Biarnés

Barcelona, 17 de junio de 2014

Institut Químic de Sarrià

Final Degree Project

Prediction of specificity in chitin deacetylases using a virtual docking protocol

Author:
Juan Eiros

Director:
Dr. Xevi Biarnés

Laboratory of Biochemistry
Bioinformatics and Molecular Modelling Unit
Bioengineering Department

June 2014

"No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

Albert Einstein

Acknowledgements

I would like to acknowledge the help and guidance of Dr. Xevi Biarnés, for without him I would have not been able to immerse myself into bioinformatics. I have learnt from him a great deal of theoretical and practical issues about this branch of knowledge, and I would like to acknowledge his dedication towards my study. Also, I would like to appreciate the daily counsel of Javier Romero, and for his patience with my constant enquiries. Finally, I would like to thank the rest of the biochemistry department for having made my stay such a pleasure.

Contents

Abbreviations

CAZY	Carbohydrate-active enzymes
CDA	Chitin deacetylase
CE-4	Carbohydrate esteares family 4
COSs	Chitooligosaccharides
DA	Degree of acetylation
DP	Degree of polymerisation
MurNAc	N-acetyl muramic acid
$MurNH_2$	Muramic acid
GlcNAc	N-acetyl-D-glucosamine
$GlcNH_2$	D-glucosamine
RMSD	Root-mean-square deviation
PA	Pattern of acetylation
PDB	Protein Data Bank

Introduction

The present study is defined within two European Commission funded projects currently running at the Laboratory of Biochemistry at IQS. One of the main research topics of this laboratory is the biotechnological production of high-value chemical products by means of living organisms or its components. To this end, many different techniques are used in biotechnology, being two the most widespread[1]. On the one hand, bioprocess engineering uses microorganisms that can be grown using biorreactors, in order to obtain the products derived from their metabolic activity. On the other hand, enzymatic catalysis can be used, that is, the synthesis and engineering of proteins and peptides destined to catalyse chemical reactions. This seccond approach is the one used by the projects in which this work is defined. Biotechnological processes are an alternative and complement to conventional synthethic chemistry. It is beneficial in ecological terms, as biotechnology is a great asset in the field of Green Chemistry, a field that encourages the design and processes that minimise the use and generation of hazardous substances[1]. Additionally, the great specificity of enzymes is of great help when asymmetric synthesis is needed to produce a highly complex chemical molecule.

1.1 Enzymes and their applications

Enzymes are the biomolecules (generally, proteins) that determine the guidelines of the chemical transformations defining the metabolism of living beings. They act catalysing these reactions with great specificity, not only in the reaction they catalyse but also in the reacting substrates. As any other catalyst, enzymes do not alter the chemical equilibrium of the reaction, which depends only on the difference in free energy between the reactants and the products. Instead, they accelerate the consecution of the equilibrium state. The vast majority of metabolic reactions would be too slow in absence of enzymes.

Their catalytic capacity resides in the energetic stabilization of the transition state, thus diminishing the activation energy of the reaction. ($\Delta G^{\ddagger}$) (Figure 1.1).

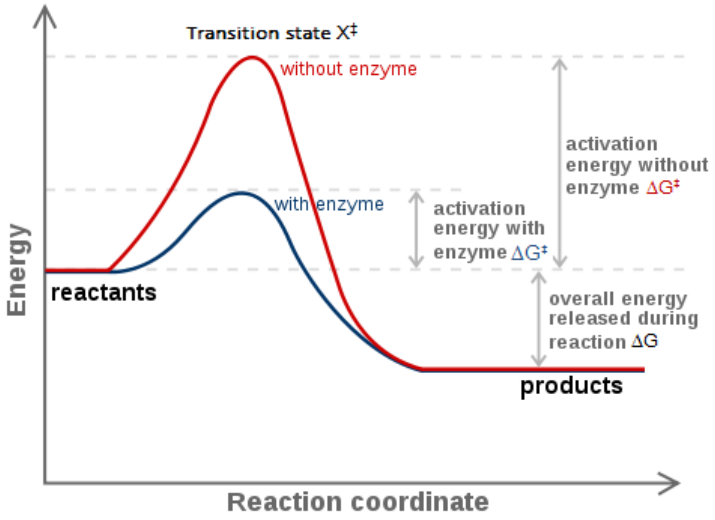


FIGURE 1.1: Enzymes accelerate reactions because they decrease $\Delta G^{\ddagger}$

The enzymes' catalytic capacity is explained from their ability to orientate the reactants (substrates and enzyme cofactors) in favourable orientations inside their catalytic pocket, forming an enzyme-substrate complex *ES*, that facilitates the formation of said transition state. The catalytic pocket is a three-dimensional space containing all the aminoacidic residues directly participating in the creation and breaking of covalent bonds.

Typical uses of enzymes as catalysts include lipases[2] and proteases[3]. Lipases hydrolyse naturally triglycerides, but also act upon other substrates, as simple n-alcohols and polyhydroxy compounds such as carbohydrates. Hydrolysis reactions of lipases can be carried out in a great range of carboxylic acids. Additionally, the synthethic inverse reaction to hydrolysis can be carried out in non-aqueous media. The most widespred use of lipases is the industrial use of lipases to produce fatty acid esters, with better purity and yields compared to the conventional chemical transesterification reaction. Proteases, in turn, constitute one of the most important group of industrial enzymes. They are widely used in the detergent, leather and dairy industries. They catalyse the hydrolisis reaction of the peptide bond that link aminoacids together in polypeptide chains.

1.2 Chitin and chitosan

Chitin is the second most abundant natural polymer on earth, second only to cellulose. Its structure consists of the linear union of N-acetyl-D-glucosamine residues by β-(1-4)-glycosidic bonds (Figure 1.2.a.). It is primarily found in the exoskeleton of arthropods and in cell walls of the majority of fungi. It is insoluble in water and organic solvents. From its partial deacetylation, chitosan (Figure 1.2.b.) and chitooligosaccharides (COSs, oligomers of chitosan) are obtained. Chitosan is not used to refer to a single chemical entity, but rather to a group of compounds that vary in their composition depending on the process by which chitin has been deacetylated. Chitosan is readily dissolved in water at pHs below 6.5. Thus, an operational definition of chitosan can be established: chitosan is chitin deacetylated enough to be soluble in water at acidic pHs[4]. It is one of the few natural polycationic biopolymers, as the amine groups in its D-glucosamine ($GlcNH_2$) residues are protonated inside the range of pHs where this polymer is soluble. In addition, the reactivity of these amine groups allows for the formation of gels with a great variety of cross-linking agents. A number of parameters define the physicochemical characteristics of chitosan: its degree of polymerisation (DP), defining the molecular mass and therefore establishing the difference between chitosan and COSs (DP 2-6); its degree of acetylation (DA), which dictates the charge distribution along the polymeric chain; and its pattern of acetylation (PA), which describes the distribution of GlcNAc and $GlcNH_2$ residues[5, 6].

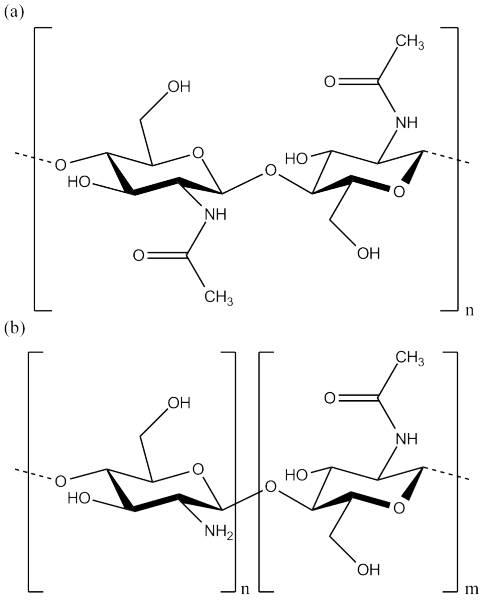


FIGURE 1.2: Chemical structure of chitin (a) and chitosan (b)

1.2.1 Applications and production of chitosan

Chitosan, and in particular COSs, are materials of great interest from the point of view of their application, thanks to a series of advantageous characteristics: (1) their biodegradability and biocompatibility with a great deal of organs, tissues and cells; (2) their physiological inertness and hydrophilicity; (3) the possibility to be processed in a variety of physical forms, such as fine powders, gels, membranes, beads, sponges, cottons and fibres[5]. These properties allow chitosan to be used extensively in agriculture, water treatment, cosmetic and food industry, as well as in biomedical and pharmaceutical industries (Table 1.1). As mentioned before, the DP, DA and PA of chitosan dictate its physicochemical characteristics and therefore its applications.

TABLE 1.1: Applications of chitosan and COSs. Adapted from Tsigos et al.[5]

Application	**Examples**
Water treatment	Recovery of metal ions and pesticides. Removal of phenols, proteins, radioisotopes, PCBs and dyes. Recovery of solid materials from food-processing wastes.
Agriculture	Seed coating. Fertilizer and fungicide.
Food additives	Clarification and deacidification of fruits and beverages. Preservative agent, antioxidant. Control of textures and flavours.
Biomedical and pharmaceutical materials.	Treatment of major burns. Artificial skin preparation. Antitumoral agent, blood anticoagulant. In drug - and gene - delivery systems.
Cosmetics	Skin and hair care products.
Chromatographic media	Immobilization of enzymes. Matrix for affinity chromatography.
Others	Synthetic fibres. Coating of paper.

Production of chitosan and COSs has traditionally been achieved at a large scale by chemical deacetylation of chitin, obtained from crab and shrimp shells as a byproduct of the seafood industry. Firstly, the shells are deproteinized and demineralized under

alkaline conditions and application of $CaCO_3$ to obtain pure chitin. Afterwards, chitin is subsequently deacetylated with the treatment of a strong base (NaOH or KOH at 40-50%) at high temperatures (80-160°C)[4, 5, 7]. Depending on the reaction conditions (base concentration, temperature and time of reaction) chitosan is obtained with a DA between 7% and 50%. The annual worldwide production of chitosan is estimated to be ca. 2000 tons. Nevertheless, the chitosan market is a very confidential one and there is a lack of reliable data. Today, bulk chitosans cost around 13 to 25 €/kg depending on the quality and quantity. Despite the vast amount of applications of chitosan which have been mentioned before, the majority of it is commercialised as a corporal fat blocker. It is sold as a dietary supplement to reduce weight, claiming that it 'traps' the fat in food, therefore reducing the fat absorption and increasing its excretion. Nevertheless, these claims lack scientific backing[8–10].

There are three critical disadvantages in the classical production of chitosan: (1) the energetic consumption of the process is considerable; (2) large amounts of concentrated alkaline solution are wasted, resulting in a daunting environmental pollution; (3) it leads to products with a broad range of DP, DA and PA. This last issue is a particular hindrance for the biomedical and pharmaceutical fields, since rigorous essays have to be performed in order to determine the biological activity of chitosan and COSs, which depend entirely on these factors. Furthermore, these fields ultimately affect human health, so it is of great concern that the reproducibility and repetitibility of these tests can be ensured.

Therefore, a demand for pharmacological grade chitosan has to be supplied. This kind of chitosan is sold up in small quantities for up to 30,000 €/kg, and must be obtained through alternative methods allowing the control over its physicochemical properties (therefore, its DP, DA and PA).

The present study is defined within two European projects (Chitobioengineering, funded within the 6th Framework Programme and Nano3Bio, funded within 7th Framework Programme) destined to develop biotechnological methods to produce sufficient quantities of defined chitosan polymers to suit the high price sector of the cosmetics, pharmaceutics and biomedical market. Namely, it aims at using chitin deacetylases (CDAs) as catalysts for the deacetylation reaction upon chitin oligomers (DP 2-6), which in contrast to chitin are soluble in aqueous solution and are more accessible for enzyme

reaction. CDAs are capable of carrying out the specific deacetylation of certain positions in a short chain of GlcNAc units, therefore obtaining COSs of defined DP, DA and PA. Moreover, the biotechnological production of COSs allows for the reaction to take place at mild conditions, without having to use high temperatures nor environmental hazardous reactants.

1.3 Chitin deacetylases

1.3.1 Deacetylation reaction and its mechanism

CDAs (EC 3.5.1.41) are a class of enzymes described in various fungi and insects. These proteins catalyse the removal of the acetamido group in the GlcNAc residues of chitin, chitosan and COSs (Figure 1.3). They belong to the 4th family of carbohydrate esterases (CE-4) as defined in the CAZY database[11]. CE-4 enzymes have a conserved region in their primary structure, which is referred to as the 'NodB homology domain' or 'polysaccharide deacetylase domain', where the deacetylation reaction takes place. Aside from CDAs, there are other CE4 enzymes with deacetylase activity on GlcNAc oligomers besides from their natural substrates, such as acetyl xylan estearase (EC 3.1.1.72), peptidoglycan GlcNAc deacetylase (EC 3.5.1.-) and COSs deacetylase (EC 3.5.1.-), amongst others.

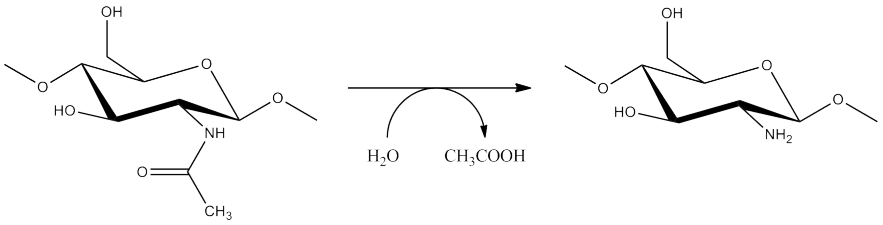


FIGURE 1.3: CDA catalyzed reaction. A unit of $GlcNH_2$ and a molecule of acetate are produced.

It has been proposed that CE4 estearases operate by metal-assisted general acid/base catalysis. This mechanism has been delineated thanks to the recent publication of the first relevant, and likely functional, crystal structure of a CDA in complex with substrate (*Vc*CDA-DP2)[12], although it had been previously hypothesized based on docking results[13]. The two-step mechanism (Figure 1.4) has been presupposed to be the same for all CE4 estearases based on the fact that they present a highly conserved

catalytic domain. The reaction begins with the nucleophilic attack of a water molecule to the carbonyl of the acetamido group. This water molecule is initially coordinating the metal, and is activated by the general base residue (Asp39, *Vc*CDA numbering). As a result of this attack, a tetrahedral oxyanion intermediate is formed, where the negative charge in the oxygen atom is stabilised by the metal coordination and the main chain nitrogen atom of Tyr169 alike. Finally, the C-N bond breaking is assisted by the protonation of the nitrogen atom by the general acid residue (His295) with the generation of a free amine in the deacetylated sugar pyranose ring and release of acetate.

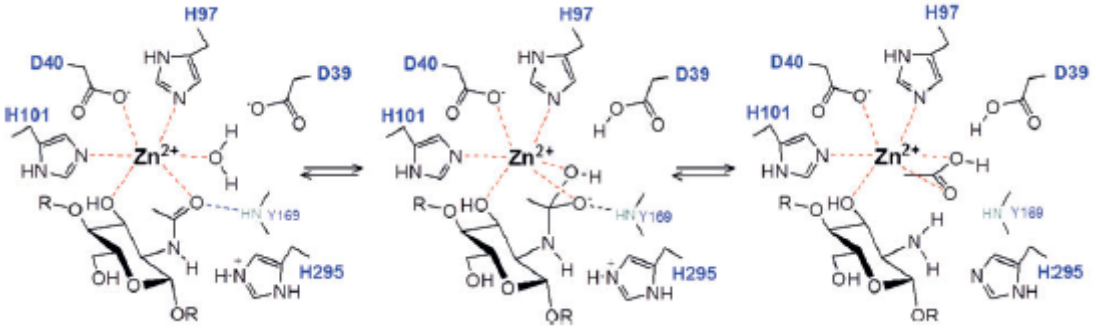


FIGURE 1.4: Acid/base catalytic mechanism of CE4 estearases. Aminoacidic residues are numbered based on *Vc*CDA's sequence. Extracted from Andrés et al. [12]

1.3.2 Specificity and modes of action

The deacetylation pattern and substrate specificity exhibited by CDAs, and related CE4 enzymes active on GlcNAc oligomers, is diverse. Whilst some are specific for a single position (e.g. *Rm*NodB[14] and *Vc*CDA[12]), others are able to attack the oligomeric chain multiple times (e.g. *Cl*CDA[13] and *Sl*AxeA[15]). At present, it is a major challenge to understand how this specificity is determined. Nevertheless, two means of action have been proposed: multiple attack mechanism (Figure 1.5.A) and multiple chain mechanism (Figure 1.5.B)[6]. In the multiple attack mechanism, binding of the enzyme onto the non-reducing end of the substrate is followed by a progressive deacetylation, whereupon fully deacetylated chitosan is produced. On the contrary, multiple chain mechanism-acting enzymes bind onto a specific position of the substrate and perform a single deacetylation. Afterwards, the product can be additionally processed by forming a new productive complex with the enzyme. For specific enzymes, only one deacetylation takes place. Others, such as *Cl*CDA, end up fully deacetylating the oligomer if enough time of reaction is allowed[13].

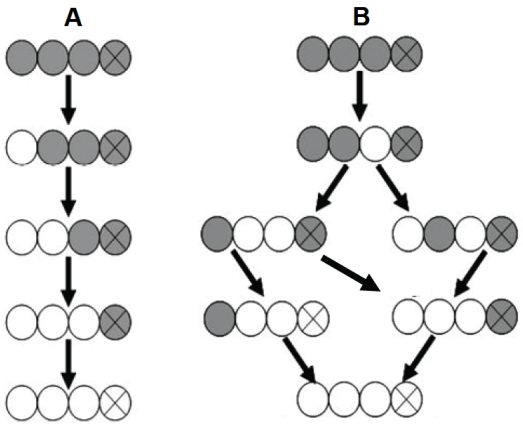


FIGURE 1.5: Multiple attack mechanism (A) and multiple chain mechanism (B). Extracted from Zhao et al.[6]

There are numerous CE4s enzymes of interest in order to produce COSs of high quality, showing different specificities (Table 1.2). All of these enzymes have been proven to be active on chitin oligomers, although some are not strict CDAs and therefore act naturally on other substrates, such as peptidoglycan.

TABLE 1.2: Summary of CE4 estearases of interest.

Organism	Protein name	PDB	Deacetylation product
Rhizobium meliloti	*Rm*NodB	-	○●●●
Bacillus subtilis	*Bs*PdaC	-	●●○●
Vibrio cholerae	*Vc*CDA	XXXX	●○●●
Colletotrichum lindemuthanium	*Cl*CDA	2IW0	●●○●
Streptococcus pneumoniae	*Sp*PgdA	2C1G	○■○■
Bacillus subtilis	*Bs*PdaA	1NY1	●□●□
Streptomyces lividans	*Sl*AxeA	2CC0	○○○○
Escherichia coli	*Ec*PgaB	3VUS	○○○○
Aspergilus nidulans	*An*CDA	2Y8U	?

● GlcNAc ○ $GlcNH_2$ ■ MurNAc □ $MurNH_2$

1.3.3 Structural features of CDAs

The polysaccharide deacetylase domain of CE4 enzymes adopts a $(\beta/\alpha)_8$ barrel topology. The central core is comprised of seven parallel β-strands that form a distorted β-barrel,

surrounded by α-helices. Inside the β-barrel, the highly conserved catalytic residues are situated. This sequence is LTXDDG, where X can be F (in most cases) or Y. A multiple sequence alignment (Figure 1.6) also indicates the presence of other conserved residues, corresponding to the two His residues coordinating the metal, and the catalytic His. Additionally, the presence of variable regions corresponds to loops surrounding the catalytic centre.

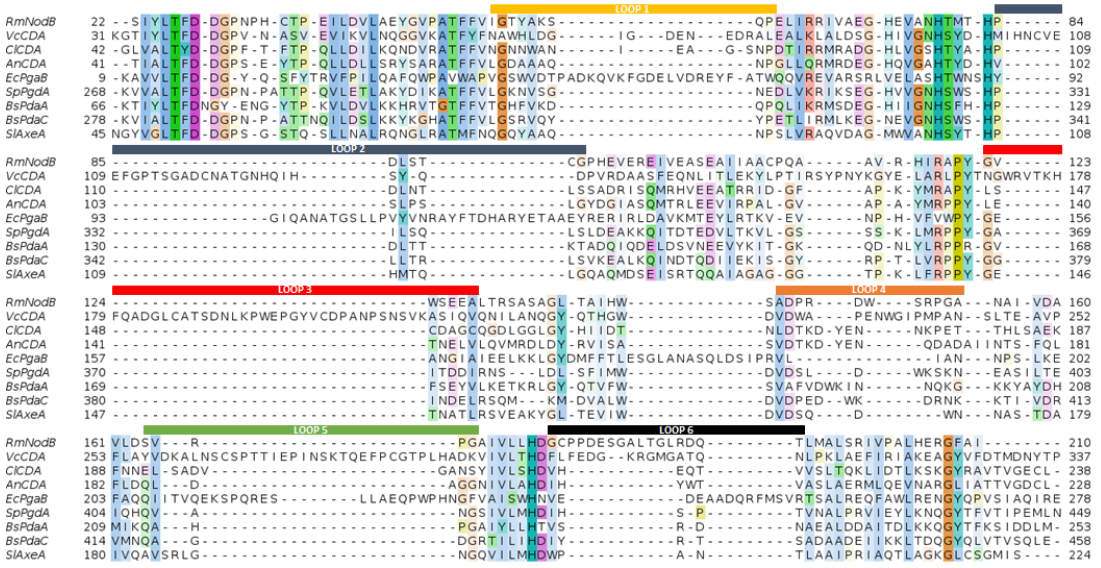


FIGURE 1.6: Multiple sequence alignment based on structure of the polysaccharide deacetylase domain of different CE4 enzymes.

Supperposition of CE4 crystallized enzymes structures of interest allows for the observation of the differences in their loops. These enzymes can possess other domains asides from the polysaccharide deacetylase domain, and such is the case for *Vc*CDA and *Sp*PgdA. While *Vc*CDA has two carbohydrate binding domains, *Sp*PgdA owns a unique N-terminal domain not identified in the Uniprot database[12, 16].

Combining the information extracted from the multiple sequence alignment and the superposition of the available structures of CE4 enzymes, differences in their loops can be examined in more detail. In Table 1.3, a comparison of the loops is presented. No structural data is available for *Rm*NodB nor *Bs*PdaC[17], as they have not yet been crystallised. For the currently crystallised enzymes, *Vc*CDA shows the largest loops 2, 3 and 5, differentiating it from the other CE4 enzymes.

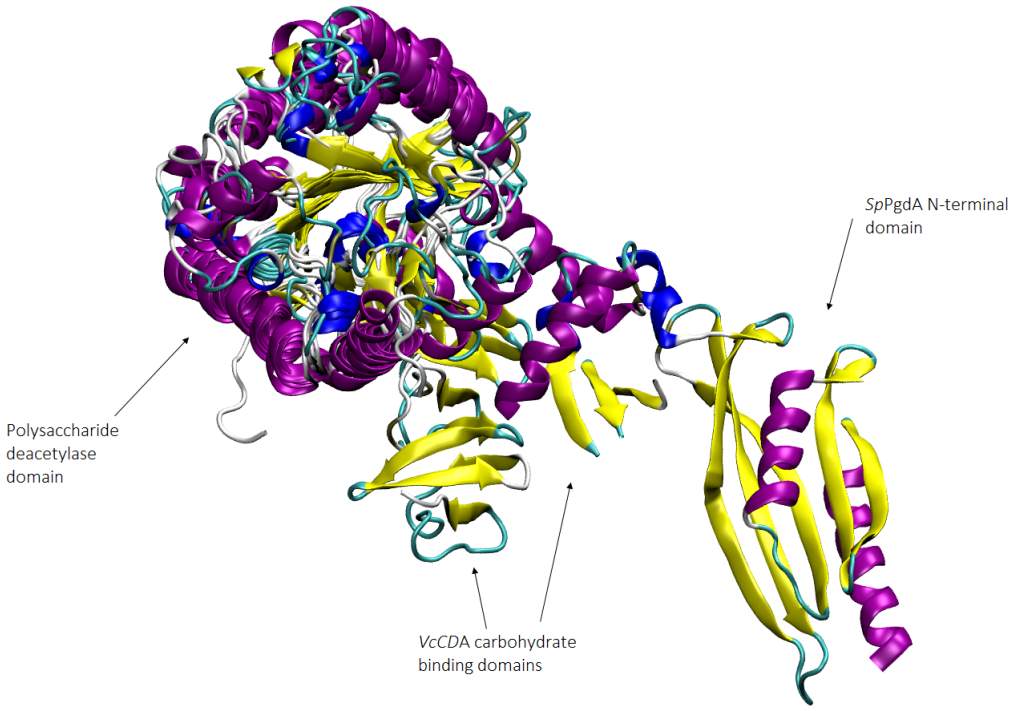
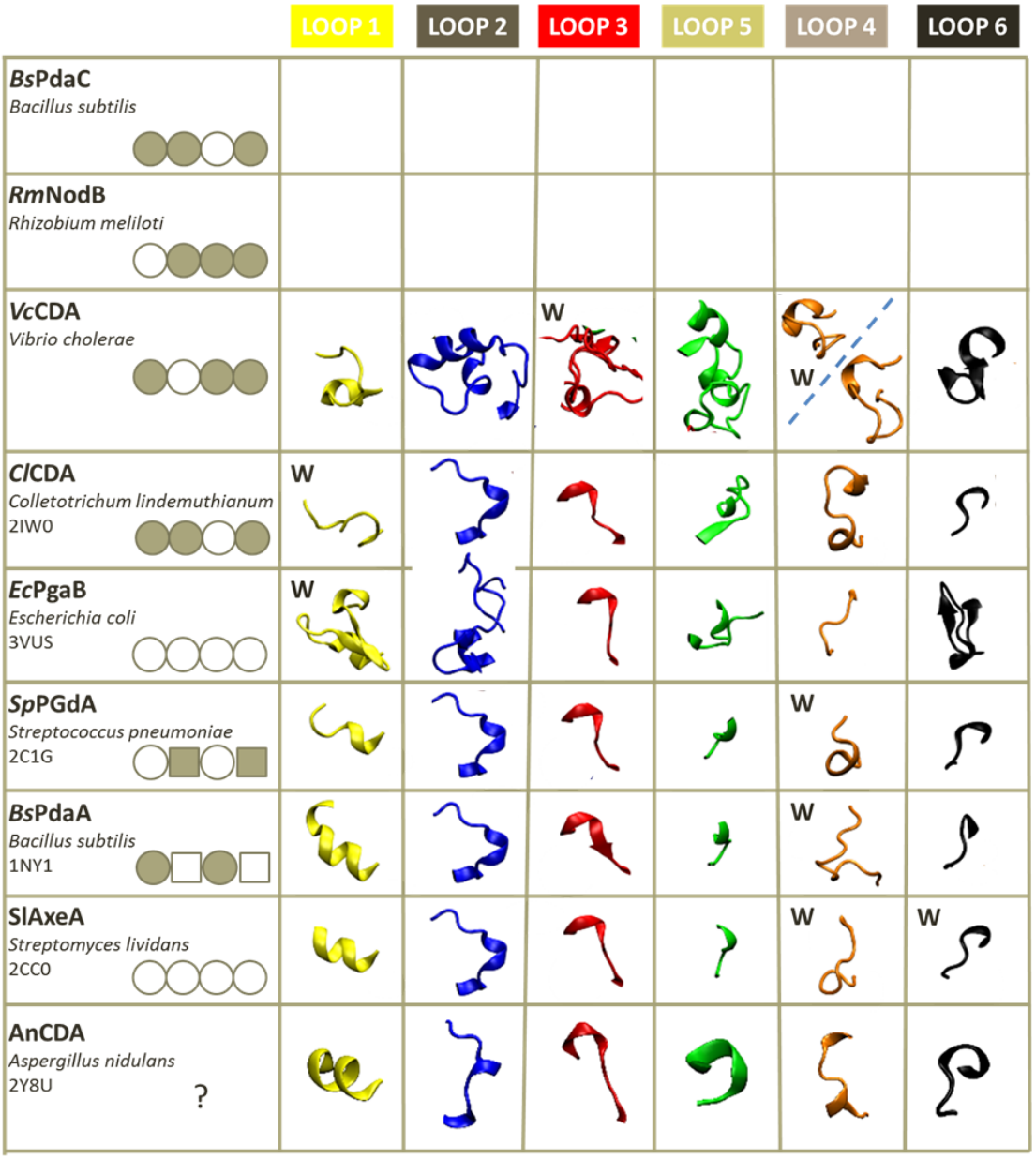


FIGURE 1.7: 3D structural superposition of CE4 enzymes

1.3.3.1 Subsite capping model

Given the sequential and structural diversities of loops in CE4 enzymes, and based on recent structural studies of *Vc*CDA[12], a new model was proposed in order to rationalise the specificity and deacetylation pattern of CE4 enzymes[12]. For a better understanding of the reaction, different subsites in the enzyme have to be defined based on the relative position of the substrate. Given an oligomer of GlcNAc with a DP of 4 (henceforth simply referred to as 'DP4'), up to 7 different subsites (-3 to +3) and 4 different modes of union can be described (A, B, C and D). In this model, subsite 0 corresponds to the GlcNAc unit that is being deacetylated (Figure 1.8). The new subsite-capping model proposes that the different loops decorating the active site of these enzymes, and their dynamics, modulate the binding of the substrates. Thus, 6 different loops have been described, where 3 of them (Loops 1, 2 and 6) regulate the availability of the negative subsites, and the rest (Loops 3, 4 and 5) act on the positive subsites. An alternative representation of the subsites is shown in Figure 1.9, where *An*CDA is represented as a surface in white, and the different loops surrounding the active site are coloured.

TABLE 1.3: Structural comparison of CE4 loops.



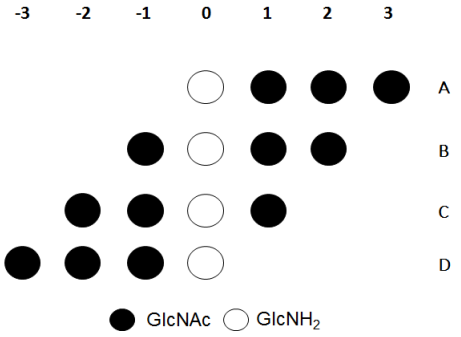


FIGURE 1.8: The four different modes of union for a 'DP4' molecule (A, B, C and D).

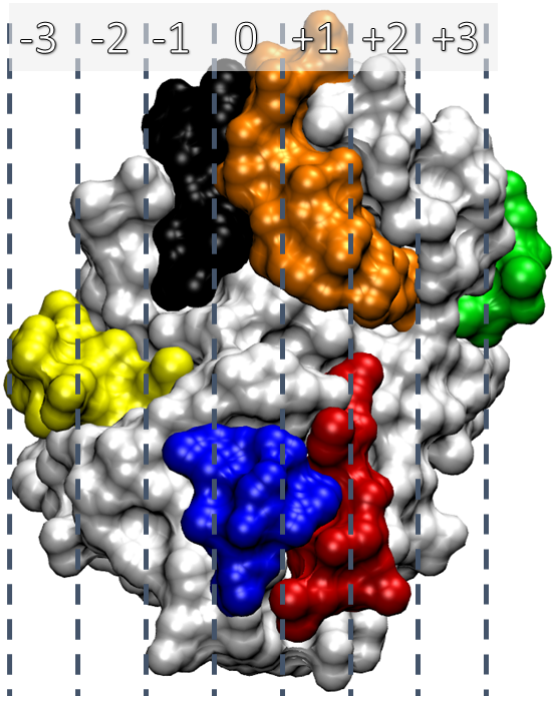


FIGURE 1.9: The seven possible subsites defined by the loops of a CE4 enzyme.

Objectives

Taking into account the need to produce COSs of high quality, and the recent postulation of the subsite capping model for CE4 enzymes, the objectives of this work are:

- To develop a simple computational docking procedure that allows the prediction of the pattern of acetylation of any given CE4 enzyme.
- To model the structure of the *Rm*NodB and *Bs*PdaC enzymes using public web services.
- To study the pattern of acetylation of different CE4 enzymes using the previously developed docking procedure.

Theoretical Background

3.1 Molecular recognition

Molecular docking is a computational tool developed to predict the bound conformation of a small molecule to a larger molecule, such as a protein, and the binding affinity of it. The basic process behind this technique is molecular recognition. This term is used to refer to the specific interaction between two or more molecules through non-covalent bonding. Molecular recognition events are continuously present in any chemical transformation process, and are crucial in biological systems. Examples highlighting this importance are receptor-ligand interactions between an enzyme and its substrates, antigen-antibody recognitions, cellular receptors, amongst others.

Protein-ligand interactions are the type of event which is of relevance in the present work. The common features of this interaction are detailed below:

- It involves non-covalent bonding, such as hydrogen bonds, ionic, hydrofobic, cation-pi interactions and metal complexation. All these interactions have different strenghts, and are usually all present in protein-ligand interaction events. Their 3D arrangement dictates the specificity of the interaction.
- Because it is highly specific, the guest and the host must exhibit molecular complementarity. This means that there is steric complementarity between the protein and the ligand. Electrostatical and solubility complementarity are also present. The surface properties of both the receptor and the ligand are complementary (e.g. lipophilic parts of the ligand tend to be in contact with lipophilic parts of the protein).

- The affinity is relatively high. Thus, the ligand is placed inside the enzyme in an energetically favourable conformation.

With this knowledge, structure based ligand design has gained considerable importance within the field of molecular modeling. The main aim is to be able to design ligands that target a specific protein, generally for drug design purposes. The computational docking approach is a novel way to tackle this objective, using the power of computation of modern computers and the knowledge of thermodynamics, which allow to evaluate the energetic contributions to binding in protein-ligand interactions.

3.2 The computational docking problem

The objective of computational docking is the development of a rapid method that allows the prediction of two questions: how does a certain ligand bind to a certain protein, and what is the stability of the formed complex. The methodology consists in the extensive evaluation of different protein-ligand conformations, until the most plausible one, in terms of predicted binding energy, is found.

The molecular docking problem can be faced through different approaches. The most simple one is to consider both protein and ligand to be rigid. This is currently considered to be a too rough approximation, as only 6 degrees of freedom are taken into account to define their relative orientation. The relative position between two rigid objects in a 3D space can be defined with a distance (R), two angles (F, G) and three dihedrals (W, Q, T). Nevertheless, flexibility in both ligand and receptor should be taken into account. However, this increases the complexity of the problem.

In the present work, only flexibility of the ligand has been taken into account. The free rotation of bonds inside the ligand can change the orientation of functional groups. Therefore, the complexity of the problem is augmented by L degrees of freedom, being L the number of dihedrals that define the number of rotatable bonds of the ligand. Flexibility of the protein has not been taken into consideration, as the different rotamers of the aminoacid sidechains increase the problem's complexity by another P dihedrals. It should be noted that if flexibility of the protein were to be considered, only some aminoacid sidechains would be made rotatable, as considering the flexibility of the protein as a whole renders the docking computation completely unapproachable. Overall,

a schematic representation of the protein-ligand configurations considered for this work is illustrated in Figure 3.1.

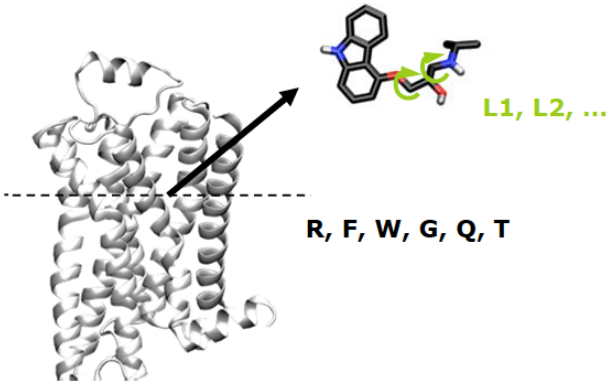


FIGURE 3.1: The protein-ligand configuration is represented by 6 + L degrees of freedom.

Taking into consideration the previously stated arguments, the three main ingredients of docking can be defined:

- First of all, an efficient search algorithm is necessary, that drives the search of the optimal binding pose. This is because it is virtually impossible to systematically analyse all possible orientations of the ligand to the receptor, and the internal rotamers (even considering the protein as a rigid body).
- Secondly, a scoring function will determine the goodness of the protein-ligand conformation which the search algorithm finds. This scoring function is a mathematical expression that weighs the energetic contributions to binding, and has to be easily computed.
- Finally, a docking calculation is never run once. A statistical analysis of several docking solutions has to be performed in order to obtain meaningful conclusions.

The search algorithm and the scoring function are the two features that differentiate docking programs from one another. In the present work, AutoDock has been used[18]. It has been developed at The Scripps Research Institute, and is freely available for all to use.

3.2.1 The search algorithm

There are two widely used search algorithms for docking purposes. The first one is the Metropolis Method, also known as Monte Carlo simulated annealing. The second one is the Genetic algorithm, used by AutoDock, and will be explained subsequently.

Genetic algorithm tries to simulate the process of evolution *in silico* (Figure 3.2). To that end, an initial population of chromosomes is generated randomly. Each chromosome has a number of genes equal to the number of degrees of freedom needed to describe the protein-ligand system (6+L). Thus, one chromosome (the genotype) describes unequivocally the conformation of the protein-ligand complex (the phenotype). Variability in the system is introduced by two means: inducing mutations randomly into the genes, and by 'sexual' reproduction between the individuals (crossover of the chromosomes). The whole simulation is composed by a series of generations. At each generation, a population of parents gives rise to a new offspring population by means of the two previously described operations. Then, a selection method is applied for each individual, and the parents have to compete with one another so that only the individuals with the best fitness give rise to offspring in the next generation. Solutions better suited to their environment reproduce, while worse ones die. This is similar to the process of evolution that happens in nature with biological systems, and in the case of this algorithm the environment to which the individuals have to adapt is the scoring function. Eventually, after a number of generations and energy evaluations, the last remaining offspring contains the best scored protein-ligand complex conformation.

3.2.2 The scoring function

Overall, scoring functions have to be easily computable, because the scoring of bindings drives forward the search algorithm. The rigorous theoretical calculation of binding free energies is difficult because it requires to take into account ensemble of configurations. Thus, scoring functions are not strict in physical sense because they simplify the calculation by decomposing the binding free energy into a sum of terms. There are three main classes of scoring functions (Table 3.1).

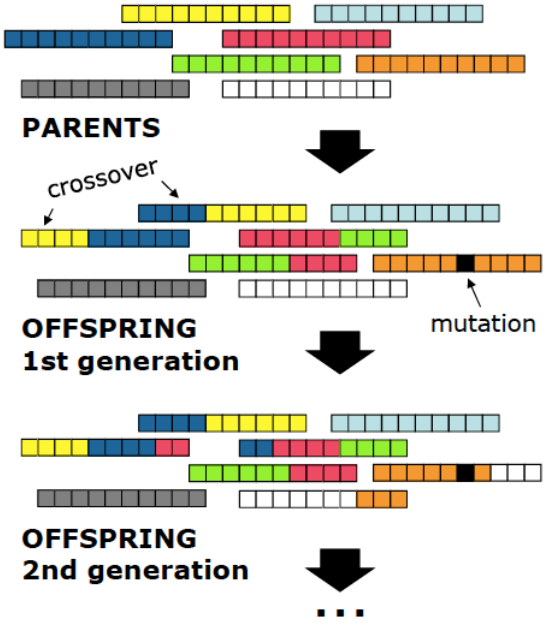


FIGURE 3.2: Simulation process of the Genetic algorithm

TABLE 3.1: Comparison of the three main types of scoring functions.

	Computational cost	**Physics Reliability**	**Robustness**
Force Field	High	High	High
Empirical	Low	High	Low
Heuristic	Low	Low	High

AutoDock uses a hybrid scoring function, based both on Force Field and Empirical methods. The force field (Equation 3.1) includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}).

$$\Delta G = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right) \quad (3.1)$$

Each of the pair-wise energetic terms includes empirical evaluations for dispersion/repulsion, hydrogen bonds, electrostatics and desolvation force-field parameters (Equation 3.2). The first term is a typical 6-12 Lennard-Jones potential for dispersion/repulsion interactions. The second term is used to describe hydrogen bond interactions based on a 10-12 potential. The E(t) function is used to describe the strict directionality of this interaction, based on the angle t from ideal H-bonding geometry. The third term is a typical Coulomb potential for electrostatics. Finally, the last term is a desolvation potential based on the volume of atoms (V) that surround a given atom and shelter

it from solvent, weighted by a solvation parameter (S) and an exponential term with distance-weighting factor σ of 3.5 Å.

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbond} \sum_{i,j} E\,(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{e(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-\frac{r_{ij}^2}{2\sigma^2})} \tag{3.2}$$

The weighting constants W are optimized by multilinear regression using a training set of protein-ligand complexes for which both the binding affinity and 3D structure is experimentally known. This is the predominant drawback of empirical scoring functions, as the binding affinity calculation is strongly dependant on the set of protein-ligand complexes that the software developer uses to calibrate these weighting constants.

One of the advantages of the AutoDock software is its use of precalculated grid maps. A map is calculated for each type of atom present in the ligand that will be docked onto the surface of the protein. This enhances the speed of the search algorithm, as the scoring function does not have to be computed from scratch every time a bound conformation is found. Instead, the numerical values are interpolated in the previously calculated grid. The grid is a three-dimensional lattice of regularly spaced points, which can be set by the user. Each point of the lattice stores the potential energy of a '*probe*' atom that is due to all the atoms in the protein.

3.2.3 The statistical analysis

Given the simplification character of scoring functions, and the space limitation in the search algorithm, a single solution of a docking procedure is never enough. The best strategy to follow is to repeat the same docking several times (even thousands) and analyze the results. The 'solution' for the docking process are the protein-ligand complexes with the best score and more repeated.

Normally, a cluster analysis is carried out in order to summarize big amounts of data. Generally, there are two ways to clusterize results. On the one hand, it can be convenient to cluster all the results in the root-mean-square deviation (RMSD) vs. the energy of

binding. This is the way that has been used in this work. On the other hand, one can clusterize by energy a number of results within a range of RMSD.

The utility of clusterizing the results is that different binding poses can be easily identified. Also, it is possible to check if the results are satisfactory by comparing them to an X-ray structure of the complex, if such information is available.

3.2.3.1 RMSD definition

For two different structures of the same molecule (a and b), the RMSD is calculated as shown in Equations 3.3 and 3.4. The summation in Equation 3.4 is over all N heavy atoms in structure a, the minimum is over all j atoms of the same type as i in structure b.

$$RMSD_{ab} = \max\left(RMSD'_{ab}, RMSD'_{ba}\right) \tag{3.3}$$

where

$$RMSD'_{ab} = \sqrt{\frac{1}{N}\sum_{i}\min_{j} r_{ij}^2} \tag{3.4}$$

Results and Discussion

4.1 The docking procedure

A docking procedure has been set up in order to predict the main deacetylating pattern of any given CE4 enzyme. This procedure consists of three main steps: (1) preparation of the protein and ligand structures; (2) computation of the docking using AutoGrid and AutoDock and (3) a statistical analysis of the docking results. Detailed description of each step is reported in Methods 6.1. In brief, (1) the three-dimensional structure of the protein can be obtained from the PDB database if it has been crystallised[19], or obtained by homology-modelling with the HHpred web-server[20]. Proteins are superposed in order to use a single grid box which explores the same volume for all the proteins. (2) The docking parameters vary depending on the ligand that is being docked. Namely, 400 dockings are performed for DP2, whereas for DP4 1600 dockings are performed because of the higher molecular complexity. (3) Reference structures for each mode of binding are first defined, and two different cutoffs are set, one based on the RMSD with respect to the reference structure and the other based on the interaction energy defined in AutoDock. This allows for a double criteria to clusterize the solutions: one that differentiates productive conformations with the lowest energies (double cutoff of RMSD & Energy) from those that have higher energies (single cutoff for RMSD).

The docking and modelling results have allowed to classify the studied enzymes based on their processivities. The results will be presented following this classification. First, non-processive enzymes will be discussed, and the importance of conformational changes in the protein will be adressed. Secondly, processive enzymes will be presented, and the importance of the availability of negative subsites will be discussed. Finally, and based on the previous results allowing for the validation of the docking protocol, a prediction of

the *An*CDA's deacetylation pattern will be given. Ultimately, drawbacks and suggestions for the improvement of the developed docking procedure will be adressed.

4.2 Non-processive enzymes

4.2.1 VcCDA

A recent structural study of *Vc*CDA has provided different structures of this enzyme in relevant states of its catalytic cycle, including the ligand-free (*apo*) and complexed forms with GlcNAc (DP1) and $(GlcNAc)_2$ DP2[12].

In order to simulate the binding of ligands to *Vc*CDA, the developed docking procedure has been performed with DP2' and DP3 on the *apo* structure of *Vc*CDA.

Regarding the DP2 docking, a cluster of results with low energies (Figure 4.1) that matches the geometry of the crystallised structure of DP2 in complex with *Vc*CDA (Figure 4.2) has been found. As can bee observed in Figure 4.1, the DP2 molecule explores positive subsites of the pocket, but the scoring function is better (red colour, low energy of binding) when the ligand is occupying subsites -1 and 0.

This result is particularly important because it validates the docking methodology used for all the docking experiments in this work. Additionally, from the experimental structure of DP2 in complex with the protein, the productive binding conformation can be inferred. When analysing the docking results, productive conformations are described when the carbonyl of the acetamido group is facing towards the metal atom, as well as the hydroxyl group from the C-3 of the sugar ring. As previously stated, a cluster of results with low energy matches this productive geometry in mode of union B.

The statistical analysis of the results, taking as reference for the RMSD calculation the frame of minimum energy in a productive binding mode, has been conducted. As can be seen in the RMSD vs. Energy plot (Figure 4.3), a cluster of results matches the energy level, and similitude in RMSD, as the frame of reference. A summary of the results of the docking procedure is shown in Table 4.1. To count the bound structures, the previously mentioned criteria is used. The details for the DP2 criteria can be found in the Methodology section. It should be stated that for a DP2 molecule, only two

modes of union (A and B) can be described, as opposed to a DP4 molecule (Figure 1.8). The data summarised in Table 4.1 allows for the observation that B mode of binding is preferred over mode of binding A, with a proportion of 9:1. Thus, negative subsites -2 and -3 are completely blocked, whereas subsite -1 is open for the ligand to enter. No significant changes are observed when comparing the two criteria of result selection, as the proportion of results mantains the same profile when the less restrictive cutoff is applied (B mode of union is the majoritary mode).

Taking into consideration the high proportion of molecules found in mode of union B, and weighing the fact that subsites -2 and -3 appear to be blocked, the docking protocol indicates that *Vc*CDA deacetylates DP2 specifically on the reducing end. This is in agreement with the experimental reported data[12]. Additionally, the mentioned experimental report has proven that this enzyme undergoes a significant conformational change when it binds to chitin oligomers. More specifically, Loop 4 of *Vc*CDA closes the active site when the deacetylation reaction takes place. Nevertheless, when docking DP2 onto the *Vc*CDA*apo* structure, no influence of the protein's conformational changes can be observed. This is because the DP2 molecule is small and can fit into the deep catalytic cleft of this enzyme, so the rigidity of the protein is not a hindrance for the docking procedure in this case.

TABLE 4.1: *Vc*CDA. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

double cutoff	n	%
A	2	8,7
B	21	91,3
nº dockings OK	23	
nº dockings done	400	
% success	5,75	
single cutoff	n	%
A	4	14,3
B	24	85,7
nº dockings OK	28	
nº dockings done	400	
% success	7,00	

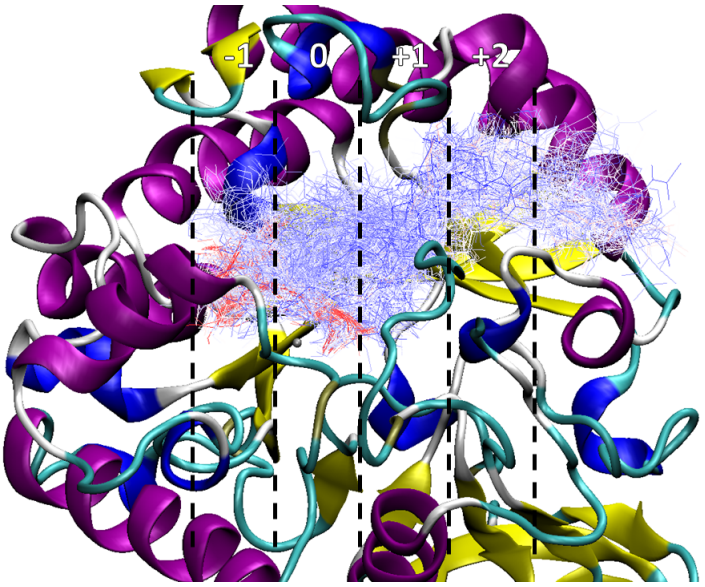


FIGURE 4.1: *Vc*CDA-DP2. Colour representation by energy of the 400 docking results. Frames coloured in red have a better (lower) energy of binding than blue ones.

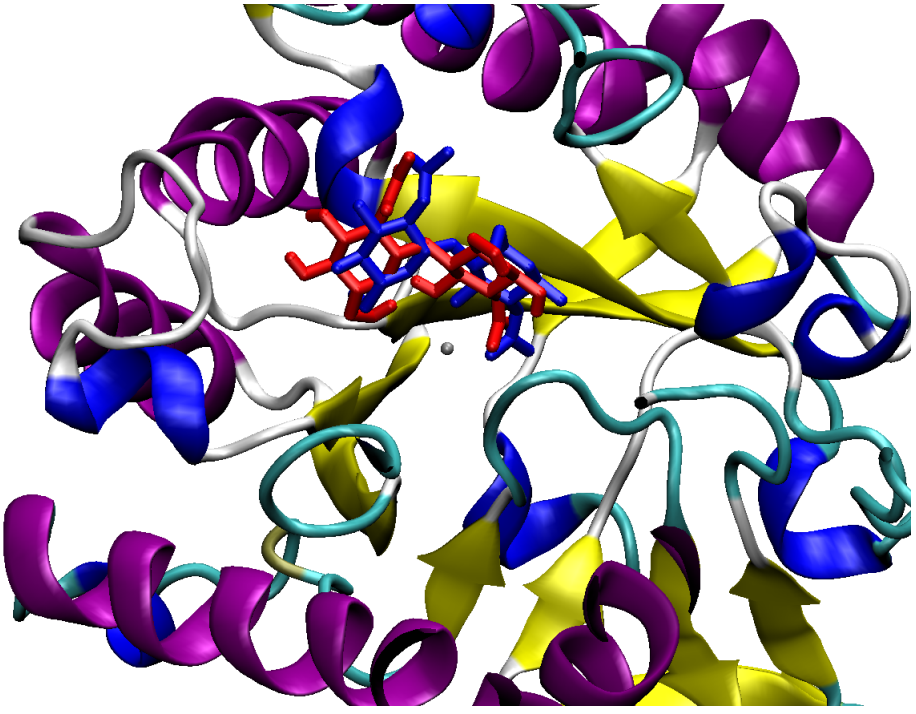


FIGURE 4.2: *Vc*CDA-DP2. Comparison of the crystallised DP2 structure (red) and the frame of lowest energy (blue).

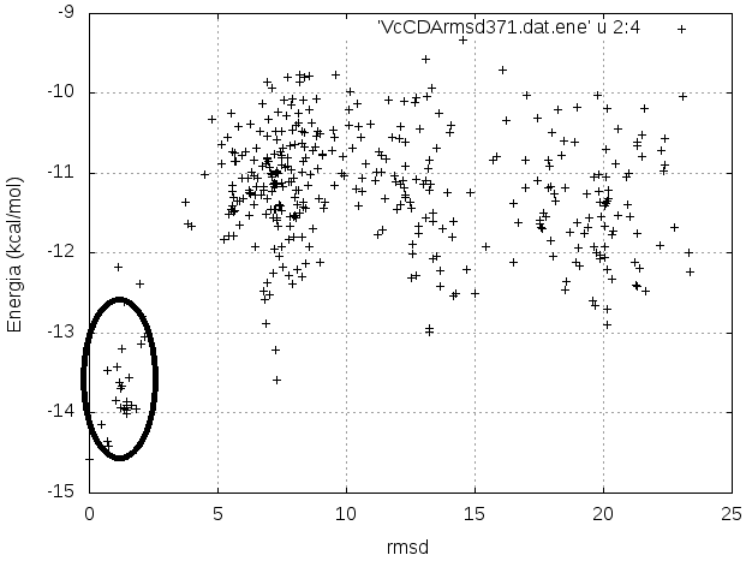


FIGURE 4.3: *Vc*CDA-DP2. Clusterization of 400 docking results of DP2 on *Vc*CDA*apo* structure. The RMSD is calculated for the whole ligand.

With regard to the dockings with DP3, no productive modes of union have been found for the 400 docking results, as the ligand seems not to fit properly into the catalytic pocket of the protein. This can be observed in Figure 4.4, where the results with lowest energy are represented. None of these ligands have a productive bound geometry, so the statistical analysis of the results has not been carried out due to a lack of reference molecules in productive conformations. The effects of conformational changes of the protein are more notable than with DP2 results. As already mentioned, the conformational changes of the protein when it binds to the ligand is a phenomenon that cannot be computed through docking. This conformational changes are accountable for the docking results of DP3 onto the *apo* structure of *Vc*CDA. Therefore, induced-fit conformational changes in *Vc*CDA appear to be decisive in the binding of ligands longer than DP3, and the developed docked protocol is not able to predict the deacetylation pattern for such ligands with this enzyme. Currently, this feature is subject of study in the Bioengineering department of IQS.

FIGURE 4.4: *Vc*CDA-DP3. Clusterization of DP3 ligands with an energy of binding below -15 kcal/mol. No productive binding modes can be found.

4.2.2 ClCDA

For this enzyme, 1600 dockings have been performed with DP4, because it is a more complex ligand than DP2. Clustering of the docking solutions based on RMSD calculations has been done as indicated in the corresponding Methodology section. A summary of the results is presented in Table 4.2. The data indicates that A and B are the two main modes of binding, both when counting the results with a double cutoff and with a single cutoff. Additionally, mode C of union has never been found, and D mode of union is of a much lower percentage than A and B modes of union for both criteria of selection. The fact that mode of union D rises its population when counting the structures only with the restriction of geometry, indicates that this mode of binding is of a higher energy than A and B. Indeed, these observations indicate that the docking protocol predicts that negative subsites -2 and -3 are not accessible. Nevertheless, *Cl*CDA has been reported to specifically deacetylate DP4 with a C mode of union (Figure 1.8)[13], although it ends up fully deacetylating the substrate at long reaction times. Strikingly enough,

the docking experiments have never found such position for the ligand, as previously stated.

TABLE 4.2: *Cl*CDA. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

double cutoff	n	%
A	12	67
B	5	28
C	**0**	**0**
D	1	6
nº dockings OK	18	
nº dockings done	1600	
% success	1,13	
single cutoff	n	%
A	42	71
B	13	22
C	**0**	**0**
D	4	7
nº dockings OK	59	
nº dockings done	1600	
% success	3,69	

*Cl*CDA's structure has a unique feature that might be accountable for this result. Loop 1 of this enzyme has a Trp residue that clearly blocks negative subsite -2. This can be observed in Figure 4.5, where the only frame found in position D is shown. This ligand is distorted, meaning that this Trp is indeed causing a strong steric impedance. Thus, ligands cannot bind in position C, because the non-reducing end of the molecule cannot occupy the negative subsite -2. Also, this is in agreement with the fact that the few solutions found in mode of union D are of a significant higher energy than A and B modes of union.

Based on the experimental evidence in the *Vc*CDA[12] study, it is plausible to think that the binding of a DP4 molecule onto *Cl*CDA provokes a conformational change, where Loop 1 moves, and the Trp residue allows access to the -2 subsite. Therefore, this conformational change could stabilize the enzyme-substrate complex in position C, inasmuch as it is more energetically favourable, and thus explaining the reason why the C mode of union is preferred experimentally. This hypothesis cannot be proved by docking alone, because the protein's structure is treated as a rigid body.

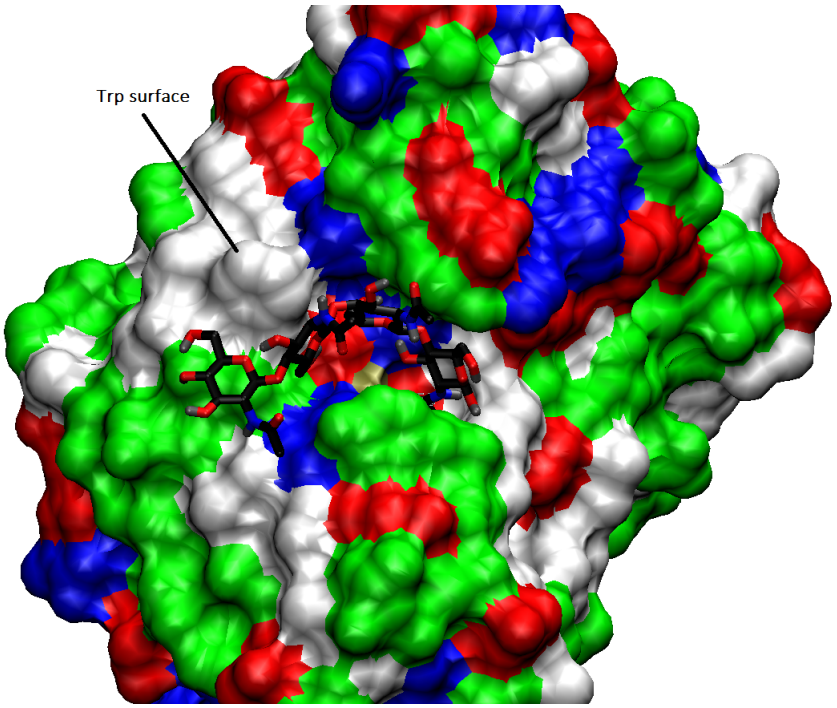


FIGURE 4.5: *Cl*CDA-DP4. DP4 ligand bound in a D mode of union. The ligand is distorted because of the steric impedance of the Trp residue occupying subsite -2.

4.2.3 RmNodB

*Rm*NodB has never been purified and crystallised, thus no 3D structure is available for this enzyme. A suitable model to perform docking experiments has been obtained, using the web server HHpred[20]. It is hosted by the Max-Planck Insitute for Developmental Biology, and is free to use. This website uses homology modeling, that is, the model is obtained identifying homologous sequences in the PDB with known structure, and using them as templates to build the 3D structure of the protein being modeled. The rapid and reliable identification of remote homologues of the queried sequence is the main strength of HHpred, and what makes it a relevant tool for bioinformatic investigation. The results for *Rm*NodB are illustrated in Table 4.3. The E-value is a parameter used by HHpred to assert the homologous relationship for a pair of proteins (the query and the template). It is defined as the average expected number of non-homologous proteins with a score higher than the one obtained for the database match. Therefore, An E-value much lower than 1 indicates statistical significance.

*Rm*NodB has the longest Loop 6 of all CE4 enzymes, so it is hard to obtain a good model of it by homology modeling using only one template. This is because, as expected, the

TABLE 4.3: *Rm*NodB. Up to 100 templates are detected by HHpred. They are scored automatically.

Rank	**PDB**	**E-value**
1	2CC0	9,10E-53
2	4M1B	1,50E-52
3	2J13	4,20E-51
4	1NY1	1,70E-49
5	2C1I	1,60E-49
6	2Y8U	2,00E-49
7	2C71	1,20E-47
8	2IW0	1,60E-46
9	2W3Z	2,80E-45
10	2VYO	2,20E-45
...		
32	2I5I	19
...		
80	4HD5	7,6

highest ranked templates are members of the CE4 family as their whole sequence has high similarity. Nevertheless, there are gaps in their sequences in the corresponding zone of Loop 6, because many CE4 enzymes have shorter Loops 6. Gaps in the sequence alignment are not desirable, because this implies that the modeling software does not have a template to base the three-dimensional structure of the sequence on, so a non-accurate model is obtained. Despite this, other models can be suitable to model the zone of Loop 6, because even if they do not have a good score for the whole sequence, they might be suitable to model the zone of Loop 6. This is the case for sequence number 32, a short primary sequence that matches that of *Rm*NodB's Loop 6 and has a defined tertiary structure. A comparison of the sequence alignment of template number 1 and template number 32 is shown in Figure 4.6 to illustrate this concept.

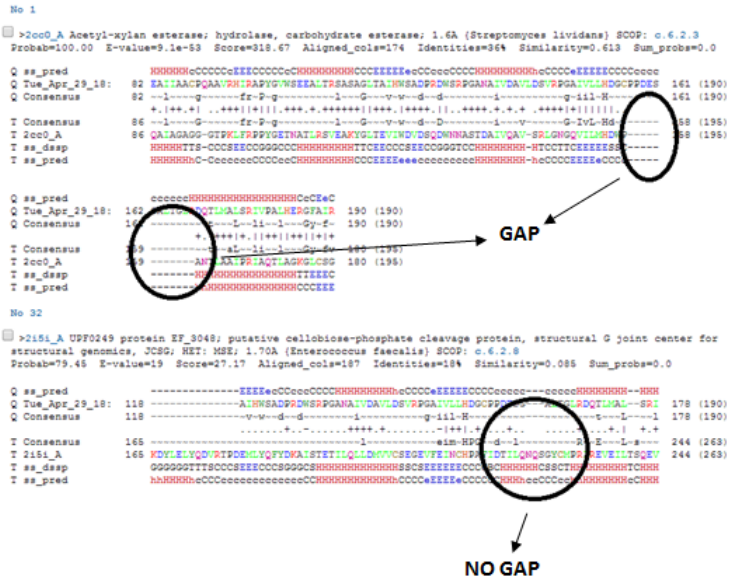


FIGURE 4.6: *Rm*NodB. Fragments of the sequence alignments of templates number 1 and number 32. While number 1 has a better overall score, template number 32 does not have a gap in the sequence corresponding to Loop 6.

Several models of *Rm*NodB structure are shown in Figure 4.7 which where generated by the single templates depicted in 4.3, as well as by a combination of templates (Hybrid models). The latter have proven to be the most adequate. Especially, model 1-32 (that is, merging template number 1 and template number 32) allows Loop 6 to be closed towards the body of the enzyme. Obtaining a good model of Loop 6 is crucial because it is not logical to have a loop completely separated from the scaffold of the protein, as this is something not observed for any crystallised structure of CE4 enzymes. Taking into consideration the structure of the models, 1600 dockings with DP4 were performed onto the 1-32 model of *Rm*NodB. As a summary of the results, the number of ligands with a double cutoff (RMSD and Energy) is presented in Table 4.4. The data indicates that subsites -2 and -3 are not available for the substrate, and only a very few amount of high energy structures manage to fit in (modes of union C and D in the single cutoff section of the Table). Therefore, A and B mode of union are the majority, being the A mode of binding clearly predominant, with a double and single cutoff (clusterization of A binding mode is shown in Figure 4.8). This can also be observed when visually analysing the docking results, as shown in Figure 4.9, where a cluster of ligands in position A is illustrated, and this corresponds to the specificity that has been reported in the literature[14].

On the whole, *Rm*NodB model 1-32 has proven to be a satisfactory model to perform the developed docking protocol with. The protocol is in agreement to the experimental findings on this enzyme. That is, it deacetylates specifically chitin oligomers at the non-reducing end, as negative subsites -2 and -3 are completely blocked, and negative -1 is only available at a very low ratio (13%).

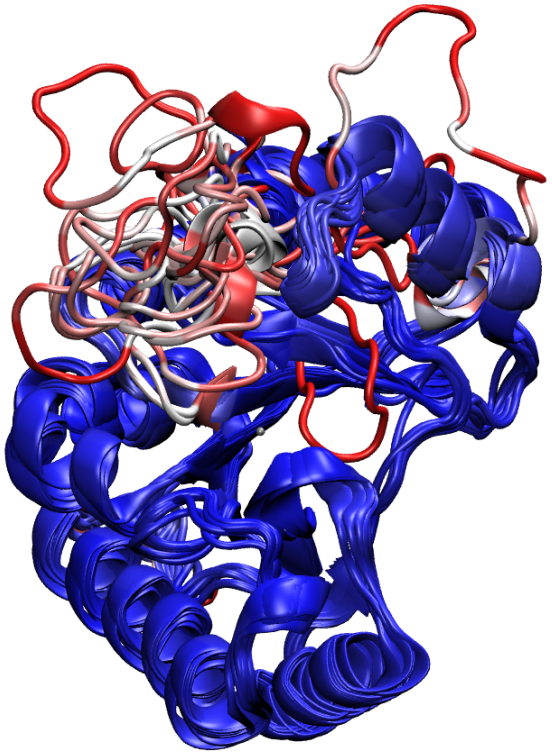


FIGURE 4.7: *Rm*NodB. Superposition of 15 different models of *Rm*NodB. Blue colour corresponds to conserved region, red colour corresponds to variable structure (Loop 6).

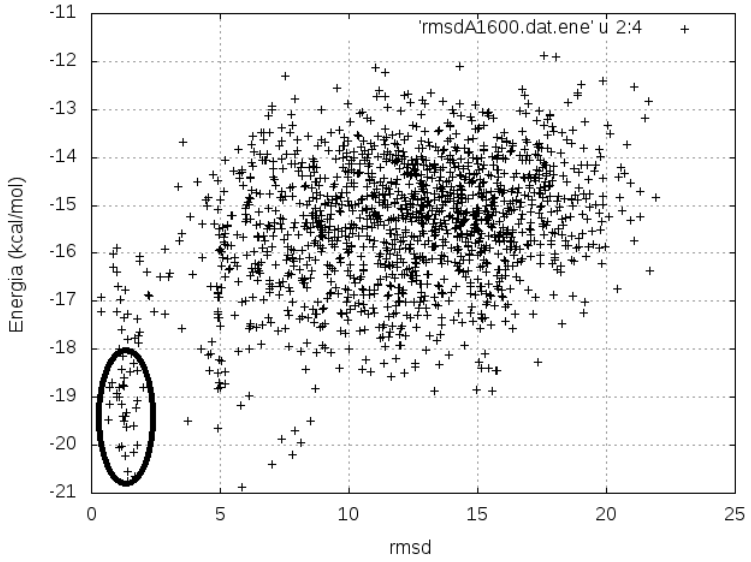


FIGURE 4.8: *Rm*NodB-DP4. Clusterization of 1600 docking results on the 1-32 *Rm*NodB model. The RMSD is calculated only for ring A of the ligand.

TABLE 4.4: *Rm*NodB-DP4. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

double cutoff	n	%
A	35	88
B	5	13
C	0	0
D	0	0
nº dockings OK	40	
nº dockings done	1600	
% success	2,50	
single cutoff	n	%
A	57	83
B	9	13
C	2	3
D	1	1
nº dockings OK	69	
nº dockings done	1600	
% success	4,31	

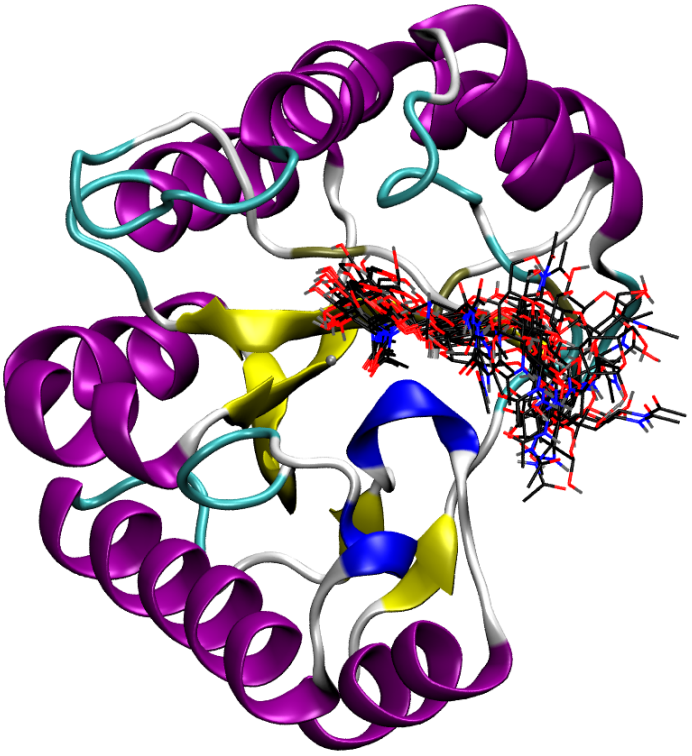


FIGURE 4.9: *Rm*NodB-DP4. Superposition of ligand structures bound to the enzyme in productive mode of union A.

4.3 Processive enzymes

4.3.1 SlAxeA

A summary of the docking results for this enzyme is presented in Table 4.5. The results show that all modes of union can be found, even with the double cutoff restriction. When the single cutoff restriction is applied, the number of structures for each mode of binding rises. This data differs from the behaviour previously shown by specific enzymes, for whom modes of binding C and D were almost nonexistent. This fact indicates that all seven subsites of this enzyme are open, from -3 to +3. Nevertheless, not all the modes of binding have the same energy. For instance, C mode of union rises the most in proportion when the single cutoff restriction is applied, indicating that this mode of binding has a higher energy compared to the rest. Moreover, mode of union A is always predominant, indicating that positive subsites have more favourable interactions with the ligand than negative subsites, despite the fact that they are all available. *Sl*AxeA is known to be a processive enzyme[15], and the results with single and specially double cutoff agree with

this fact. By comparing the number of structures in C and D mode of union in Table 4.5, it is evident that these modes of binding have a higher energy than the rest, reinforcing the idea that there are more favourable protein-ligand interactions in the zone of the positive subsites, thus lowering the overall energy of A and B modes of union. This might be accountable for the well-stablished Multipple attack mechanism depicted in Figure 1.5, where processive enzymes start to deacetylate on the non-reducing end of the ligand.

On the whole, dockings with *Sl*AxeA have highlighted the processive character of this enzyme. That is, every subsite from -3 to +3 is available for the substrate to enter. This feature can be clearly observed in Figure 4.10, where all the structures with a lower energy than -18 kcal/mol are represented. Compared to non-processive enzymes, *Sl*AxeA's loops are not impeding in any way the binding of the substrate to any particular subsite, and the enzyme can attack the ligand through a multiple attack mechanism.

TABLE 4.5: *Sl*AxeA-DP4. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

double cutoff	n	%
A	10	59
B	5	29
C	1	6
D	1	6
nº dockings OK	17	
nº dockings done	1600	
% success	1,06	
single cutoff	n	%
A	24	44
B	5	9
C	10	19
D	15	28
nº dockings OK	54	
nº dockings done	1600	
% success	3,38	

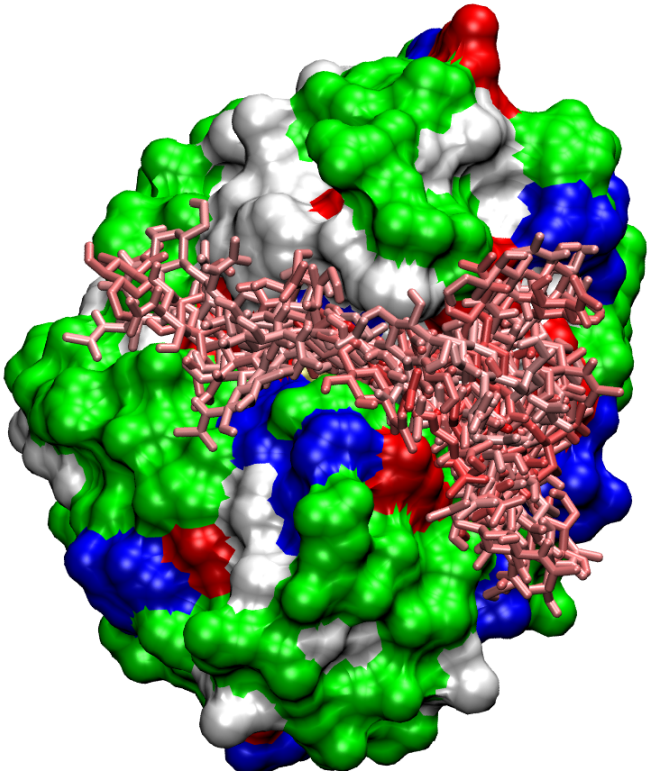


FIGURE 4.10: *Sl*AxeA-DP4. Visual representation of docking solutions with an energy of binding below -18 kcal/mol.

4.3.2 BsPdaC

As with *Rm*NodB, no structural data is available for *Bs*PdaC. Therefore, a 3D model of this protein is necessary in order to perform dockings on it. *Bs*PdaC is a multi-domain protein: it owns the N-terminal unknown domain presented in Section 1.1.3, and a C-terminal polysaccharide deacetylase domain, which all CE4 enzymes own. It is highly likely that this protein is orthologous to *Sp*PgdA, because they both share the same two domains and the same natural function, that is, both deacetylate GlcNAc units of peptidoglycan. Accordingly, the N-terminal domain of *Bs*PdaC was not included in the HHpred query because it does not intervene in the deacetylation process, so there is no need to obtain a model for it. Querying only the C-terminal domain, which is responsible for the reaction, allowed for the identification of more accurate templates for its structure.

An overview of the HHpred web server results is presented in Table 4.6. As opposed to *Rm*NodB, *Bs*PdaC has no 'special' loops. That is, all of its loops are standard with regard to their length, so the alignments do not have substantial gaps. Several models

were obtained, as with *Rm*NodB. Many were discarded, based on punctual errors in their structure. Mainly, most models were not satisfactory because of the orientations of key aminoacid residues in the structure (LTFDDG conserved sequence, the two His residues coordinating the metal, etc.). The four best models are presented in Figure 4.11 which correspond to an automatic selection of the best 20 templates by the server, and a manual selection of templates no. 2, no. 5 and no. 8. This four models have slight differences in the orientation of aromatic residues, which usually play a key role in steric interactions. Nevertheless, they all have correct orientations of the LTFDD motif, and the 3 key His residues (the catalytic one and the two coordinating the metal).

TABLE 4.6: *Bs*PdaC. Top 10 templates found for the modelisation of *Bs*PdaC polysaccharide deacetylase domain.

Rank	**PDB**	**E-value**
1	4M1B	6,50E-50
2	2C1I	4,60E-50
3	2CC0	1,00E-49
4	2J13	7,70E-49
5	2Y8U	4,20E-48
6	1NY1	2,20E-47
7	2C71	6,90E-47
8	2IW0	9,50E-45
9	2VY0	1,00E-43
10	2W3Z	2,30E-41

The docking protocol has been run 1600 times for each of the four models, to determine if there are any differences amongst them. A summary of the results is shown in Table 4.7. On the one hand, model no. 2 and the Automatic model show the same features. Both models indicate that modes of union C and D are blocked when applying the more strict criteria, so subsites -2 and -3 appear to be blocked. On the contrary, this two modes experiment a substantial rise in their proportion when considering the geometry of binding only, regardless of the interaction energy. On the other hand, models no. 5 and no. 8 behave similarly. For these two models, mode of union C has been found when applying the double cutoff, albeit at a low proportion. Also, all modes of union have been found when applying the single RMSD cutoff, as with the rest of models. In

general, these results seem to indicate that all 7 subsites are available for the substrate to enter, suggesting *Bs*PDaC to be a processive enzyme. Nevertheless, the activity and specificity on chitin oligomers of this protein has been reported, showing the same deacetylation pattern as *Cl*CDA[17]. This is not in accordance to the results obtained with the docking procedure which do not show any preference for the mode of union C. No conformational changes should be expected for this enzyme, as its loops are not different from the rest of the CE4 family. In fact, the modeled loops resemble a lot to those of *Sl*AxeA, differentiating it from *Cl*CDA in the sense that its loops are of the standard length. In an attempt to merge the docking results with the experimental evidence, it is possible that *Bs*PdaC is a processive enzyme that acts with a multiple chain mechanism (Figure 1.5). If that were the case, it might explain why it appears to be specific for mode of union C. This would only happens at low times of reaction, but if enought time was given, the enzyme would end up fully deacetylating the substrate, as all the subsites have been predicted to be accessible. Nevertheless, the docking results cannot explain why mode C is preferred, as the majoritary mode is A for all the models (and as with all the other processive enzymes).

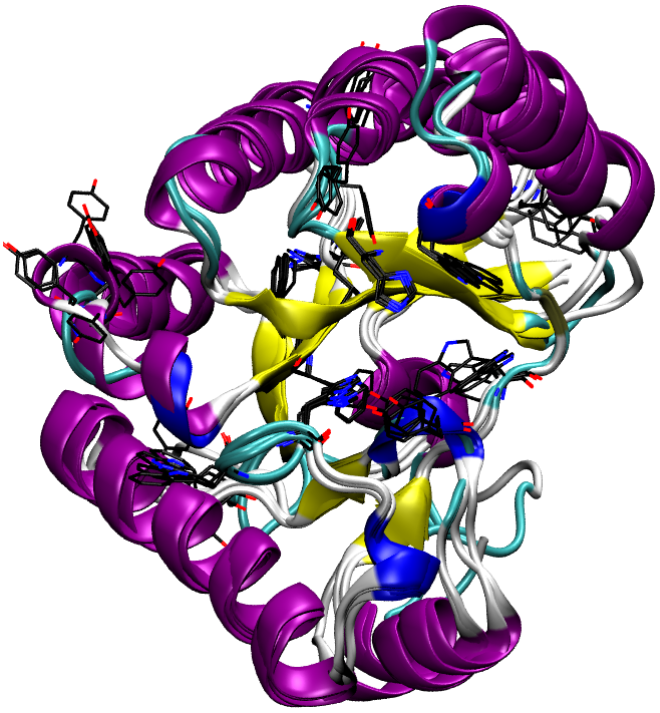


FIGURE 4.11: *Bs*PdaC. Superposition of models no. 2, no. 5, no. 8 and Automatic. Differences in the orientations of Trp395 and Tyr319 (aromatic residues in the core and Loop 1, respectively) can be observed.

TABLE 4.7: *Bs*PdaC. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

	no. 2		no. 5		no. 8		Automatic	
double cutoff	n	%	n	%	n	%	n	%
A	9	50	4	36	9	75	4	67
B	9	50	5	45	2	17	2	33
C	0	0	2	18	1	8	0	0
D	0	0	0	0	0	0	0	0
nº dockings OK	18		11		12		6	
nº dockings done	1600		1600		1600		1600	
% success	1,13		0,69		0,75		0,34	
single cutoff	n	%	n	%	n	%	n	%
A	42	40	27	44	45	69	30	68
B	41	39	12	20	9	14	3	7
C	11	10	11	18	7	11	3	7
D	12	11	11	18	4	6	8	18
nº dockings OK	106		61		65		44	
nº dockings done	1600		1600		1600		1600	
% success	6,60		3,81		4,06		2,75	

4.3.3 AnCDA

Based on the previous results with the other studied enzymes, the docking protocol has been tested on *An*CDA to predict its deacetylation pattern. This enzyme has been reported to be active on chitin oligomers, as it is a strict CDA, although no pattern of acetylation has been described[21]. Its structure has been determined experimentally, and its loops are very similar to *Sl*AxeA's.

As with the other enzymes, 1600 dockings were performed on this protein with DP4. An overview of the docking results is shown in Table 4.8. As with other studied enzymes, C mode of union is not found at low energies, but it is when the criteria of energy is not applied. If the single cutoff distribution of percentages is compared to that of *Sl*AxeA's, it can be infered that they act similarly. Moreover, a visual representation of the ligands

with an energy of binding below -18 kcal/mol, clearly shows that positive and negative subsites of this enzyme are open alike for the subsite to enter (Figure 4.12).

Therefore, after having examined the docking results, it can be predicted that *An*CDA will deacetylate in the same way as *Sl*AxeA, that is, it will be a processive enzyme and will act with a multiple attack mechanism.

TABLE 4.8: *An*CDA. Results for ligands with a double cutoff (RMSD & Energy) and single cutoff (RMSD).

double cutoff	n	%
A	11	69
B	4	25
C	0	0
D	1	6
nº dockings OK	16	
nº dockings done	1600	
% success	1,00	
single cutoff	n	%
A	21	51
B	5	12
C	6	15
D	9	22
nº dockings OK	41	
nº dockings done	1600	
% success	2,56	

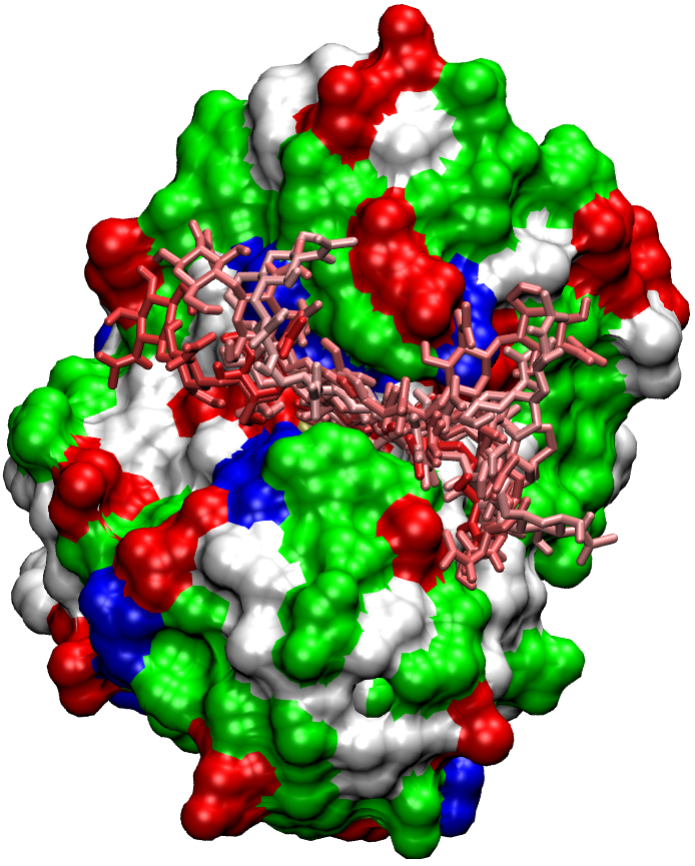


FIGURE 4.12: *An*CDA-DP4. The ligands on a docked conformation can enter both positive and negative subsites.

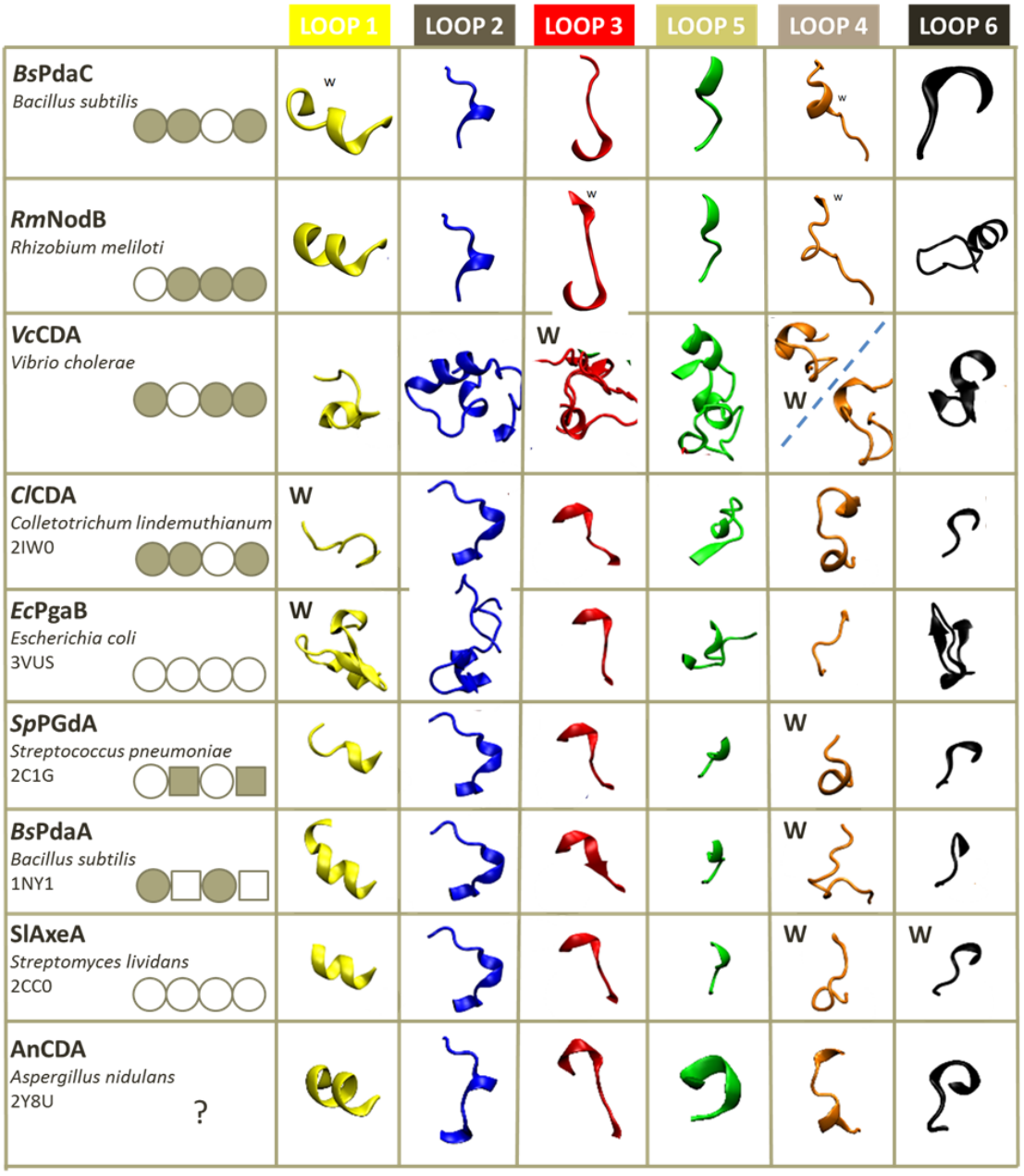
4.4 Summary

The developed docking protocol has allowed to predict the deacetylating pattern of different CE4 enzymes, based on the distribution of results for each binding mode of the ligand. The proportions of the docked results for each mode imply the availability of subsites of the enzyme. For specific enzymes, negative subsites are blocked, and one of the binding modes is always predominant. *Cl*CDA is a special case of specific enzyme, as it is theorised that a conformational change happens during the binding of the ligand molecule, so the results with the rigid *apo* structure are not in agreement with the experimental specificity. This is supported by the experimental findings of *Vc*CDA's structural changes and verification of such with DP2 and DP3 ligands. With regards to the *Rm*NodB, a semi-automatic modelling of the structure has proven sufficient to yield results which are in line with the experimental specificity of this enzyme (activity on the non-reducing end of the ligand).

As for processive enzymes, they have all of their subsites available for the ligand to enter. Nevertheless, positive subsites show a higher stabilization of the enzyme-substrate complex, which might be accountable for the experimentally-determined multiple attack mechanism of such proteins. With regards to *An*CDA, it is predicted that it will be a processive enzyme, due to its high resemblance with *Sl*AxeA. With regards to the *Bs*PdaC models, they all give results that do not concur with what is reported in the literature. In this case, and taking into account the experimental data, it is possible that this is a processive enzyme that acts with a multiple chain mechanism, as conformational changes are discarded for a lack of differentiating characteristics of its loops' structures.

Finally, after models have been obtained for *Rm*NodB and *Bs*PdaC, the CE4 loops table has been updated, as shown in Figure 4.9. The loops for *Rm*NodB and *Bs*PdaC are taken from models no. 1-32 and no. 2, respectively.

TABLE 4.9: Updated structural comparison of CE4 loops.



4.5 Drawbacks and improvements

First of all, it should be noted that the percentage of success of the developed docking protocol is very low. The percentage of success is defined herein as the percentage of all the structures that fit the selection criteria (double or single cutoff) with respect to the number of dockings that have been performed.

Obviously, this number is greater when the recount of structures is done with a single cutoff, as compared to when its performed with a double cutoff. Nevertheless, the percentage of success is below 10% in all of the performed docking experiments, and is usually around 1%. This can be explained due to the fact that the DP4 is a complex ligand, with many (24) torsional degrees of freedom. This is best observed when comparing the percentages of success of dockings with DP2 and dockings of DP4, being the first slightly greater (DP2 has only 12 torsional degrees of freedoms). Therefore, it could be argued that the developed docking protocol is at the limit of the docking technique. That is, dockings with DP5 or even longer oligomers would prove to be very inefficient, as a much larger number of dockings should be run in order to obtain meaningful results, rendering the purpose of dockings useless. This percentage could be enhanced by modifying the docking protocol. Mainly, the grid box could be set to be of a smaller volume, in order to reduce the space where the docking algorithm can find solutions. Nevertheless, this approach has not been considered as it is desireable that the ligand can fit in the greatest volume possible, in an attempt to restrict the docking process in the slightest. Furthermore, it could be argued that in order to rise the percentage of success, the number of energy evaluations could have been increased, in order to refine the search algorithm. Notwithstanding, this parameter was controlled, and it was proven that the algorithm converged at the set number of energy evaluations, as detailed in the Methodology section 6.1.3.

Finally, another flaw of the developed docking protocol is that, for processive enzymes, the multiple chain mechanism of action seems to never be predicted. It has been theorized for *Bs*PdaC because experimental data is available that contradicts the docking results, but this might not always be the case when studying the deacetylation pattern of a CE4 enzyme. The explanation for this is that, with the available results for processive enzymes, mode of binding A seems to always be predominant and thus it is logical to think that multiple attack mechanism is their mode of action. To test this flaw, the process of multiple deacetylation has been simulated. To this end, a different software (AutoDock Vina[22]) has been used. This software uses a different (heuristic) scoring function that makes it much faster than AutoDock, and is more user-friendly because it performs the statistical analysis of the results automatically. The results are presented in Table 4.10. The ligand is deacetylated on the position that vina finds for the DP4 molecule, and is subsequently docked again, to study the position where Vina docks it.

For *Sl*AxeA, the first two steps of a multiple attack mechanism are obtained. For the model no. 2 of *Bs*PdaC, a multiple chain mechanism is obtained.

Table 4.10: Two consecutive deacetylation reactions are presented.

Protein	**DP4 position**	**Monodeacetylated position**
*Sl*AxeA	A (-7,52 kcal/mol)	B (-6,9 kcal/mol)
*Bs*PdaC	C (-6,8 kcal/mol)	B (-7,0 kcal/mol)

Conclusions

After the realisation of this work, it can be concluded that:

- The developed docking procedure allows the prediction of a CE4's deacetylation pattern, based on whether subsites are available for the substrate or not. Specific enzymes exhibit a clear pattern of docked structures, where one of the binding modes is clearly predominant, whereas processive enzymes show a more even distribution of results.
- The use of the developed docking protocol upon the *Rm*NodB model allows to predict that subsites -2 and -3 are closed, and that this enzyme specifically deacetylates the DP4 ligand on the non-reducing end. The predicted specificity is in agreement with the experimental results, therefore, using a fast and free web-server like HHpred to model this enzyme is suitable to obtain significant information through the developed docking protocol.
- The use of the developed docking protocol upon the *Bs*PdaC models does not concur with the experimental results for this enzyme. In an attempt to combine the docking and experimental results, it is proposed that this enzyme is processive with a multiple chain mechanism of action. Therefore, at low times of reactions, this enzyme specifically deacetylates the DP4 on the C mode of union; but at higher times, it will fully deacetylate the ligand because all of its subsites are open.
- *An*CDA shows the same behaviour and structural characteristics as *Sl*AxeA. Therefore, it can be possible that this enzyme deacetylates in the same processive way, by means of a multiple attack mechanism, as all subsites from -3 to +3 are available for the substrate to enter, with initial preference towards binding mode A.

Methodology

6.1 Docking protocol

The developed docking protocol is based on consecutive AutoGrid and AutoDock calculations[18]. The procedure consists of three main steps: (1) preparation of the protein and ligand structures; (2) computation of the docking using AutoGrid and AutoDock and (3) a statistical analysis of the docking results.

6.1.1 Preparation of protein and ligand structures

The three-dimensional structure of the protein is obtained from the PDB database[19] (PDB accession codes from Table 1.2). Subsequently, the `.pdb` file has to be modified in order for the software to work with it. First, all the crystallisation water molecules from the file have to be removed, as well as possible ions, co-solvents or complexed ligand molecules. This can be done by manually erasing the coordinates corresponding to the molecules to be erased from the `.pdb` file, or by following the steps detailed in the AutodockTools software tutorial. This software is a grafic user interface for AutoDock developed by the same laboratory that develops AutoDock[23].

After the water and other molecules have been removed, the `.pdb` file is opened with AutoDockTools, in order to add polar hydrogens, assign the type of atoms according to AutoDock4.2 (current version) and to compute the Gaisteger partial charges of every atom on the protein. This process is detailed in the AutoDockTools software tutorial, and this generates another file, with the `.pdbqt` extension. This is the file extension that can be used by AutoGrid and AutoDock.

This preparation process is performed similarly for the ligand `.pdb` file. The ligand coordinates are as well downloaded from the PDB database. The ligand itself is obtained by erasing the coordinates of the protein that is complexed with the ligand. The PDB codes were the ligand structures have been obtained from are: 1ZU0 for DP2, 2R0H for DP3, 1LZC for DP4.

Last but not least, if no structural data of the protein is available, it is obtained using the HHpred modelation web-server[20]. The parameters selected for the search of the homologue templates of the query sequence are:

Input format	FASTA
HMM database	pdb70_14Jun14
MSA generation method	HHbits
Max. MSA Generation iterations	3
Score secondary structure	yes
Alignment mode	local

The selection of the best templates is a more manual process, detailed for *Rm*NodB and *Bs*PdaC in the correspondent subsection of the Results and Discussion. This is because the selection of the models is done by checking that there are no substantial gaps in the primary sequence alignment of the query and the template, and that there are no errors in the 3D structure (bad orientation of key residues, for instance).

6.1.2 AutoGrid parameters

AutoGrid is the tool that AutoDock uses to calculate the grid map of the receptor previous to the docking itself. The parameters used by AutoGrid are stored in a `.gpf` file (grid parameter file). The generic `.gpf` that has been used is presented below.

```
npts 120 75 54                          # num.grid points in xyz
gridfld receptor.maps.fld               # grid_data_file
spacing 0.375                           # spacing(A)
receptor_types A C HD N NA OA SA Zn     # receptor atom types
ligand_types C HD OA N                  # ligand atom types
```

```
receptor receptor.pdbqt                # macromolecule
gridcenter 1.5 0.4 12.0                # xyz-coordinates or auto
smooth 0.5                             # store minimum energy w/in rad(A)
map receptor.C.map                     # atom-specific affinity map
map receptor.HD.map                    # atom-specific affinity map
map receptor.OA.map                    # atom-specific affinity map
map receptor.N.map                     # atom-specific affinity map
elecmap receptor.e.map                 # electrostatic potential map
dsolvmap receptor.d.map                # desolvation potential map
dielectric -0.1465                     # <0, AD4 distance-dep.diel;>0, constant
```

A visual representation of the gridbox is illustrated in Figure 6.1.

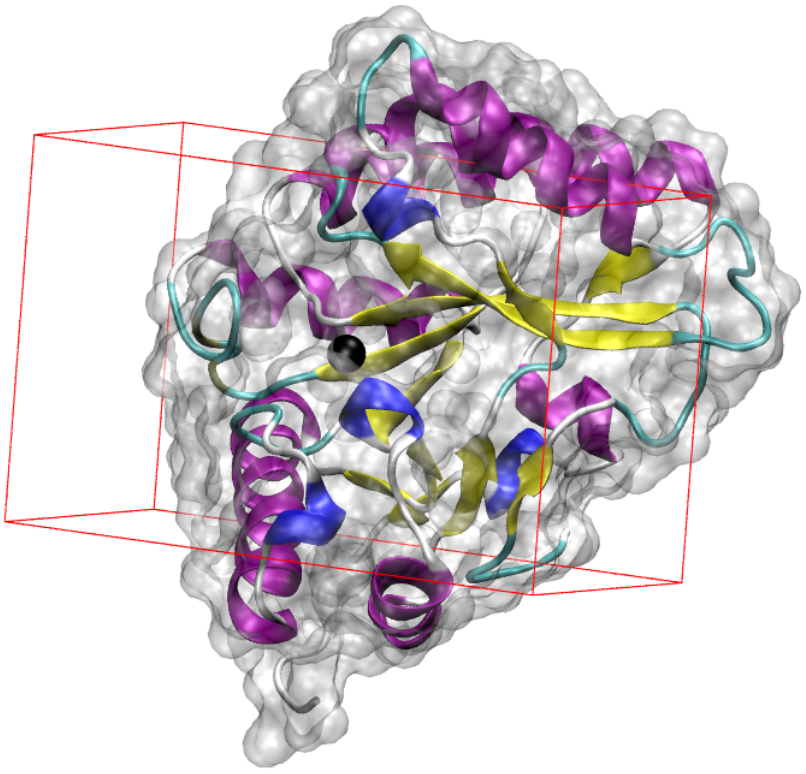


FIGURE 6.1: Visual representation of the box that defines the 3D space where the grid will be calculated, and where docking solutions will be found.

6.1.3 AutoDock parameters

The docking parameters used by AutoDock are stored in a `.dpf` document (docking parameter file). This parameters are listed below:

```
autodock_parameter_version 4.2  # used by autodock to validate parameter set
outlev 1                        # diagnostic output level
intelec                         # calculate internal electrostatics
seed pid time                   # seeds for random generator
ligand_types C HD OA N          # atoms types in ligand
fld receptor.maps.fld           # grid_data_file
map receptor.C.map              # atom-specific affinity map
map receptor.HD.map             # atom-specific affinity map
map receptor.OA.map             # atom-specific affinity map
map receptor.N.map              # atom-specific affinity map
elecmap receptor.e.map          # electrostatics map
desolvmap receptor.d.map        # desolvation map
move ligand.pdbqt               # small molecule
about 4.333 26.093 29.588       # small molecule center
tran0 random                    # initial coordinates/A or random
axisangle0 random               # initial orientation
dihe0 random                    # initial dihedrals (relative) or random
tstep 2.0                       # translation step/A
qstep 50.0                      # quaternion step/deg
dstep 50.0                      # torsion step/deg
torsdof 6                       # torsional degrees of freedom
rmstol 2.0                      # cluster_tolerance/A
extnrg 1000.0                   # external grid energy
e0max 0.0 10000                 # max initial energy; max number of retries
ga_pop_size 300                 # number of individuals in population
ga_num_evals 7000000            # maximum number of energy evaluations
ga_num_generations 50000        # maximum number of generations
ga_elitism 1                    # number of top individuals to survive to next
                                generation
ga_mutation_rate 0.02           # rate of gene mutation
ga_crossover_rate 0.8           # rate of crossover
ga_window_size 10               #
ga_cauchy_alpha 0.0             # Alpha parameter of Cauchy distribution
ga_cauchy_beta 1.0              # Beta parameter Cauchy distribution
```

```
set_ga                          # set the above parameters for GA or LGA
sw_max_its 300                  # iterations of Solis & Wets local search
sw_max_succ 4                   # consecutive successes before changing rho
sw_max_fail 4                   # consecutive failures before changing rho
sw_rho 1.0                      # size of local search space to sample
sw_lb_rho 0.01                  # lower bound on rho
ls_search_freq 0.06             # probability of performing local search on
                                individual
set_psw1                        # set the above pseudo-Solis & Wets parameters
unbound_model bound             # state of unbound ligand
ga_run 100                      # do this many hybrid GA-LS runs
analysis                        # perform a ranked cluster analysis
```

One of the parameters that was optimized was the number of energy evaluations. This parameter is controlled becacuse it is important to have a value that does not imply a high computational cost and therefore a long calculation time, but at the same time its important that the algorithm converges in terms of energy. A representation of the number of energy evaluations vs. energy of binding, is shown in Figure 6.2, to illustrate the fact that 7 milion evaluations are enough (and in fact, excessive) for the search algorithm to converge.

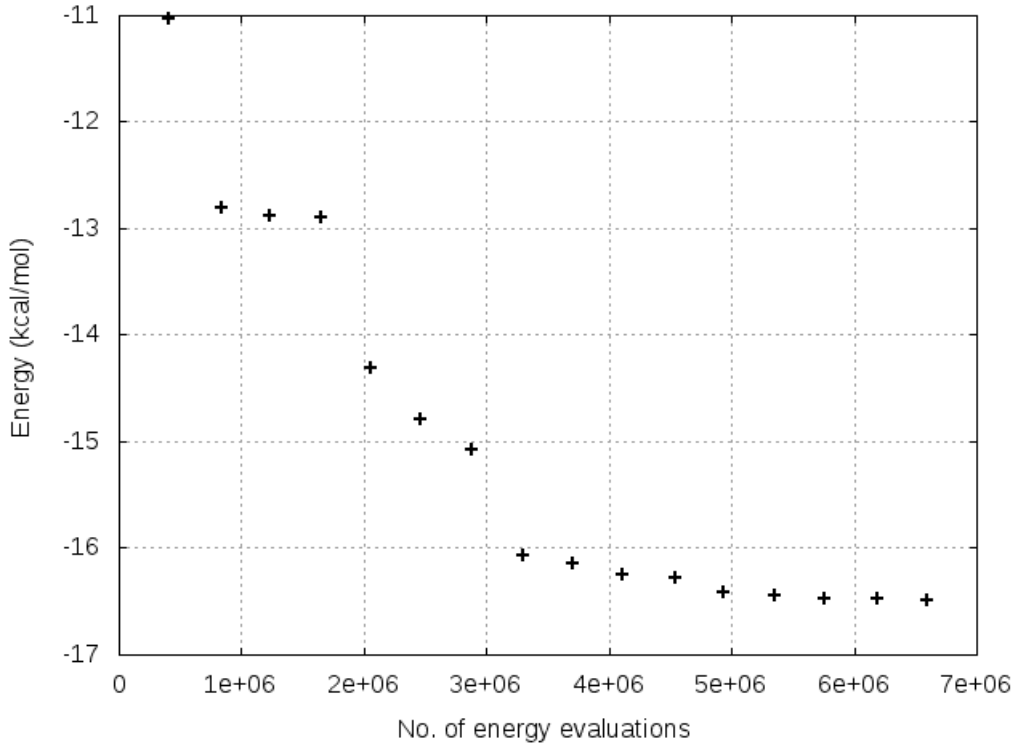


FIGURE 6.2: Evolution of the binding energy with the number of energy evaluations for a docking procedure.

6.1.4 Statistical Analysis

The clusterization of DP4 docking results has been done as follows. First of all, taking into consideration the four productive binding modes that have been described (A, B, C and D) (Figure 6.3), reference structures for each mode have been searched through all of the docking results. That is, for each mode of union, a reference molecule has been selected based only on the geometry of the sugar ring that is in subsite 0. This geometry has to be similar to the observed position in the experimental crystallisation of DP2 inside the catalytic pocket of *Vc*CDA (carbonyl of the acetamido group oriented towards the metal atom, as well as the C3-hydroxyl coordinating the metal).

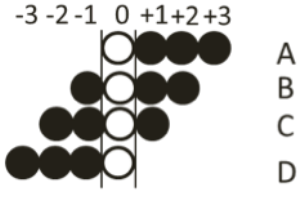


FIGURE 6.3: The four possible binding modes for a molecule of DP4.

Once these reference structures have been found (Figure 6.4), the RMSD is calculated 4 times for each docking solution. That way, the RMSD is calculated for each of the four pyranose rings of the DP4 molecule, therefore reducing the 'noise' in the RMSD calculation if it were to be done for the whole ligand. This is because the variation in position of the 3 pyranose rings that are not in the 0 subsite provoke great shifts in the RMSD value. Finally, structures are counted based on a double criteria. On the one hand, structures with a productive geometry and low energy are counted with a double cutoff (RMSD <2.0 and Energy <-18 kcal/mol). On the other hand, less energetically favourable productive conformations are counted by applying a single RMSD cutoff (RMSD <2.0). With this protocol, productive binding modes of higher energy are considered too.

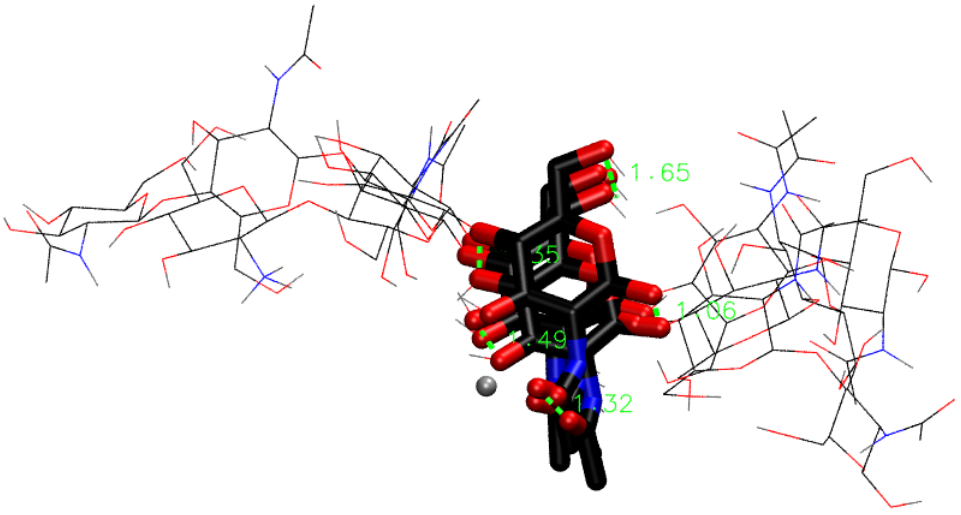


FIGURE 6.4: The four reference structures for each of the four binding modes. Distances in green are expressed in Å. Therefore, an RMSD below 2.0 is considered to be the same mode.

For DP2, the reference molecules for the RMSD calculation are selected from the docking results itself. Additionally, the RMSD is computed for the whole ligand one time, and not for each ring, as the complexity of the molecule is low enough to avoid the 'noise' in the RMSD calculation.

Bibliography

[1] Shijie Liu. *Bioprocess engineering. Kinetics, Biosystems, Sustainability, and Reactor Design*, chapter 1. Introduction. Elsevier, 2013.

[2] Geoffrey Hills. Industrial use of lipases to produce fatty acid esters. *European journal of lipid science and technology*, 105(10):601–607, 2003.

[3] Adil Anwar and Mohammed Saleemuddin. Alkaline proteases: a review. *Bioresource Technology*, 64(3):175–183, 1998.

[4] Maria Hayes. *Marine Bioactive Compounds*, chapter 4. Chitin, Chitosan and their Derivatives from Marine Rest Raw Materials: Potential Food and Pharmaceutical Applications. Springer, 2012.

[5] Iason Tsigos, Aggeliki Martinou, Dimitris Kafetzopoulos, and Vassilis Bouriotis. Chitin deacetylases: new, versatile tools in biotechnology. *Trends in Biotechnology*, 18(7):305–312, 2000.

[6] Yong Zhao, Wan-Taek Ju, Gyung-Hyun Jo, Woo-Jin Jung, and Ro-Dong Park. Perspectives of chitin deacetylase research. *BIOTECHNOLOGY OF BIOPOLYMERS*, page 131, 2011.

[7] Edwin L Johnson and Quintin P Peniston. Process for the manufacture of chitosan, March 25 1980. US Patent 4,195,175.

[8] MD Gades and JS Stern. Chitosan supplementation does not affect fat absorption in healthy males fed a high-fat diet, a pilot study. *International Journal of Obesity & Related Metabolic Disorders*, 26(1), 2002.

[9] Matthew D Gades and Judith S Stern. Chitosan supplementation and fat absorption in men and women. *Journal of the American Dietetic Association*, 105(1):72–77, 2005.

[10] C Ni Mhurchu, SD Poppitt, AT McGill, FE Leahy, DA Bennett, RB Lin, D Ormrod, Leigh Ward, C Strik, and A Rodgers. The effect of the dietary supplement, chitosan, on body weight: a randomised controlled trial in 250 overweight and obese adults. *International journal of obesity*, 28(9):1149–1156, 2004.

[11] AFMB CNRS Université d'Aix Marseille. Carbohydrate esterase family 4, June 2014. URL `http://www.cazy.org/CE4.html`.

[12] Eduardo Andrés, David Albesa-Jové, Xevi Biarnés, Bruno M Moerschbacher, Marcelo E Guerin, and Antoni Planas. Structural basis of chitin oligosaccharide deacetylation. *Angewandte Chemie International Edition*, 2014.

[13] David E Blair, Omid Hekmat, Alexander W Schüttelkopf, Binesh Shrestha, Ken Tokuyasu, Stephen G Withers, and Daan MF van Aalten. Structure and mechanism of chitin deacetylase from the fungal pathogen colletotrichum lindemuthianum. *Biochemistry*, 45(31):9416–9426, 2006.

[14] Michael John, H Röhrig, Jurgen Schmidt, Ursula Wieneke, and Jeff Schell. Rhizobium nodb protein involved in nodulation signal synthesis is a chitooligosaccharide deacetylase. *Proceedings of the National Academy of Sciences*, 90(2):625–629, 1993.

[15] Edward J Taylor, Tracey M Gloster, Johan P Turkenburg, Florence Vincent, A Marek Brzozowski, Claude Dupont, François Shareck, Maria SJ Centeno, José AM Prates, Vladimír Puchart, et al. Structure and activity of two metal ion-dependent acetylxylan esterases involved in plant cell wall degradation reveals a close similarity to peptidoglycan deacetylases. *Journal of biological chemistry*, 281(16):10968–10975, 2006.

[16] David E Blair, Alexander W Schüttelkopf, James I MacRae, and Daan MF van Aalten. Structure and metal-dependent mechanism of peptidoglycan deacetylase, a streptococcal virulence factor. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15429–15434, 2005.

[17] Kaori Kobayashi, I Putu Sudiarta, Takeko Kodama, Tatsuya Fukushima, Katsutoshi Ara, Katsuya Ozaki, and Junichi Sekiguchi. Identification and characterization of a novel polysaccharide deacetylase c (pdac) from bacillus subtilis. *Journal of Biological Chemistry*, 287(13):9765–9776, 2012.

[18] Garrett M Morris, David S Goodsell, Robert S Halliday, Ruth Huey, William E Hart, Richard K Belew, and Arthur J Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14):1639–1662, 1998.

[19] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[20] Tübingen Dept. of Protein Evolution at the Max Planck Institute for Developmental Biology. Hhpred - homology detection and structure prediction by hmm-hmm comparison, June 2014. URL `toolkit.tuebingen.mpg.de/hhpred`.

[21] Yun Wang, Jin-Zhu Song, Qian Yang, Zhi-Hua Liu, Xiao-Mei Huang, and Yan Chen. Cloning of a heat-stable chitin deacetylase gene from aspergillus nidulans and its functional expression in escherichia coli. *Applied biochemistry and biotechnology*, 162(3):843–854, 2010.

[22] Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010. ISSN 1096-987X. doi: 10.1002/jcc.21334. URL `http://dx.doi.org/10.1002/jcc.21334`.

[23] The Scripps Research Institute, June 2014. URL `http://autodock.scripps.edu/resources/adt`.