UNIVERSITÀ DEGLI STUDI DI SASSARI

Dipartimento di Chimica, Via Vienna, 2, Sassari

Dottorato di Ricerca in Biochimica, Biologia e Biotecnologie Molecolari - XXI Ciclo

Protein structure and function relationships: application of computational approaches to biological and biomedical problems

Ph.D. Dissertation

Presented by : Paolo Mereghetti

Supervisor : Maria Luisa Ganadu

Ai miei genitori

Contents

Acknowledgements

I gratefully acknowledge the financial support of the University of Sassari.

I thank Prof. Maria Luisa Ganadu for scientific help during my doctoral project and for excellent supervision.

Also I thank Prof. Piercarlo Fantucci and Prof. Luca De Gioia for the valuable collaboration and for fruitful discussion.

I thank Dr. Rebecca Wade and all the Molecular and Cellular Modelling group for the time spent at the European Media Laboratory Research of Heidelberg.

I also acknowledge the Dr. Elena Papaleo and Marco Pasi for helpful discussion and significant collaboration.

I thank the Pensione San Pietro for the accommodation during my stay in Sassari

Finally I would like to thank my little star Laura for her contribution in my thesis, but, most important, for that one in my life.

Riassunto

Le proteine rappresentano uno dei costituenti fondamentali degli esseri viventi, all'interno dei quali svolgono i pi svariati ruoli. La caratterizzazione della loro funzione non è solo importante per la comprensione del funzionamento della molecola in se, ma anche del sistema biologico del quale fa parte.
Per poter conoscere in dettaglio i meccanismi di funzionamento di una proteina, indipendentemente dal ruolo da essa svolto (enzima, recettore, strutturale, ecc.), è necessario conoscere la relazione esistente tra la sua struttura tridimensionale e la attività. Condizione necessaria per lo studio della relazione struttura-attività è la disponibilità della struttura tridimensionale. Ad oggi solo per l'1% delle sequenze proteiche depositate in banca dati è nota la struttura tridimensionale. I metodi sperimentali per la risoluzione della struttura delle proteine sono temporalmente e economicamente onerosi, il che ne limita il loro utilizzo a poche proteine di particolare rilevanza. Il recente sviluppo dei metodi computazionali per la predizione della struttura delle proteine, li ha resi, per molti casi, una valida alternativa e/o supporto ai metodi sperimentali. Uno dei limiti principali dei metodi computazionali, al quale si sta recentemente ponendo particolare attenzione, è l'incapacità di determinare con precisione l'accuratezza della struttura tridimensionale ottenuta. Per questo motivo abbiamo sviluppato un metodo di validazione di strutture proteiche (predette mediante metodi computazionali) in grado di misurarne quantitativamente l'accuratezza.
Il metodo, descritto nel capitolo 2 è basato su una combinazione di reti neurali artificiali. Le reti neurali artificiali sono dei modelli matematici sviluppati in analogia al funzionamento dei neuroni. Essendo in grado di descrivere complesse relazioni, sono state scelte per mappare la relazione tra struttura tridimensionale e accuratezza della stessa. Un vantaggio importante delle reti neurali artificiali è la capacità di apprendere da esempi. Addestrata quindi su un insieme di strutture ad accuratezza nota la rete neurale è in grado, data una struttura tridimensionale, di predirne l'accuratezza. Questo strumento è stato fondamentale in alcuni studi di correlazione struttura-attività, riportati in questo lavoro, in cui è stato necessario predire la struttura tridimensionale della proteina d'interesse.

In particolare, nel capitolo 3, è riportato lo studio dell'interazione di una nuova classe di dolcificanti naturali, gli stevioli, con il recettore umano del dolce. Gli stevioli sono estratti dalle foglie della Stevia rebaudiana, la quale contiene dieci varianti glicosilate dello stevioside, il costituente principale.

Questi dolcificanti sono recentemente diventati di grande rilevanza per via del loro alto potere dolcificante e basso potere calorico. Inoltre, essendo stati riscontrati effetti ipoglicemici, diuretici e cardiotonici, associati all'utilizzo di queste molecole, esse risultano interessanti non solo per l'industria alimentare ma anche per il settore farmaceutico.
Il recettore del dolce (t1r2-t1r3) è un recettore transmembrana eterodimerico accoppiato ad una proteina G. È composto da una regione citoplasmatica, un dominio transmembrana ed una regione extracellulare. L'interazione con i dolcificanti avviene in siti localizzati sia nella regione extracellulare che nel dominio transmembrana. Non essendo disponibile la struttura sperimentale del recettore, è stata predetta e validata utilizzando il metodo precedentemente descritto.
Attraverso uno studio di docking in-silico è stata studiata l'interazione tra gli stevioli e tutti i possibili siti di legame del recettore t1r2-t1r3. Questa analisi ha permesso di identificare dei siti di legame preferenziali per gli stevioli, in particolare, il sito localizzato nella regione transmembrana sembra essere il pi adatto legare questa classe di compost.

Le proteine non sono entità rigide, piuttosto, sono strutture caratterizzate da una particolare flessibilità. Le proprietà dinamiche associate alla struttura sono infatti fondamentali per l'attività svolta dalla proteina stessa. Sempre nell'ottica dello studio della relazione struttura-funzione è stata studiata la relazione tra proprietà dinamiche e la attività di alcuni enzimi psicrofili (capitolo 4). Gli enzimi psicrofili sono enzimi adattati a lavorare a bassa temperatura, presentano infatti un elevata efficienza catalitica alle basse temperature comparati con le controparti mesofile.
Appoggiandoci ai risultati e ai modelli proposti dalla letteratura è stato effettuato un accurato studio comparativo (enzima psicrofilo vs mesofilo) delle proprietà termodinamiche di due diversi rappresentanti appartenenti alla famiglia delle elastase e delle uracil-DNA-glycolsylase. Nello studio, effettuato mediante dinamica molecolare, è stato possibile trovare, in accordo con precedenti evidenze, che l'adattamento alle basse temperature è correlato con la differente flessibilità strutturale dell'enzima psicrofilo rispetto al mesofilo, la quale, influenza le modalità con cui l'enzima interagisce con il substrato.

La struttura tridimensionale può essere particolarmente utile quando si stanno studiando malattie che dipendono da mutazioni in una particolare proteina. È utile per analizzare l'impatto della mutazione sulla struttura e quindi sulla attività della proteina stessa.

Come esempio sono stati riportati tre casi (capitolo 5) in cui uno studio strutturale è stato utilizzato per supportare dati genetici e biochimici per l'analisi dell'impatto di mutazioni puntiformi sull'attività della proteina e il loro ruolo nella malattia associata. In particolare, sono state studiate tre rare malattie infantili associate a gravi disordini metabolici, che coinvolgono mutazioni puntiformi in proteine mitocondriali.

A causa del continuo aumento della potenza di calcolo e delle enormi quantità di informazioni biologiche depositate nelle banche dati, i metodi computazionali per lo studio della struttura delle proteine hanno raggiunto uno sviluppo tale da poter essere considerati un supporto fondamentale, ed, in alcuni casi, indispensabile, ai metodi sperimentali. È necessario tuttavia sottolineare l'importanza dell'utilizzo combinato di entrambe gli approcci. I metodi sperimentali richiedono l'elaborazione e l'analisi dei dati prodotti, mentre i metodi computazionali necessitano dei dati sperimentali per poter essere accurati.

Abstract

Proteins are fundamental constituent of the living systems in which they hold many different functions. The characterization of the protein functions can also help to explain not just the working of individual molecules, but of whole systems. Regardless of the kind of protein (enzyme, receptor, etc.), the detailed knowledge of the protein roles resides in a deep cognition of the relationship between the three-dimensional (3D) structure and its function. The study of the structure-activity relationship require, as a necessary condition, the availability of the 3D structure. At now, the 3D structure is known only for the 1% of the protein sequences stored into the databases. The experimental determination of the 3D structure is restricted to some relevant proteins, as they are temporarily and economically expensive. The recent development of the computational methods for the prediction of the protein structure, makes, in many cases, these methods a valuable choice and/or a support for the experimental ones. Recently a great effort is done to one of the major limits of the computational methods, the inability to get the accuracy of a predicted protein.
For this reason, we developed a tool for the validation of the predicted 3D protein structure, that is able to quantify their accuracy. The method, described in chapter 2, is based on the combination of multiple artificial neural networks. The artificial neural networks are mathematical models developed in analogy to the real neurons. Being able to describe complex relationship, they have been chosen to map the relation between the 3D structure and its accuracy. One of great advantages of the artificial neural networks, is the ability to learn from examples. Hence, trained on a set of structure with known accuracy, the neural network is able, given a 3D structure, to predict its accuracy. This tool has been of primary importance in some structural-activity relationship studies, described in this work, in which the prediction of the 3D structure of the protein under study, has been necessary.

In particular, in chapter 3, an interaction study between a new class of natural sweeteners (steviol glycosides) and the human sweet taste receptor, has been described. The steviol glycosides are contained in the leaves of the Stevia rebaudiana. From this plant ten different steviol glycosides can be extracted, among them, the stevioside is the main component.
The relevance of these sweeteners is recently increased due to their non-caloric property and their high sweetening power.

Moreover, hyploglycemic, diuretic and cardiotonic effects associated to these molecules makes them an important target not only for the food industry but also for the pharmaceutical one. The sweet taste receptor (t1r2-t1r3) is an heterodimeric transmembrane G-protein coupled receptor. It is composed by a cytoplasmatic, a transmembrane and an extracellular domain. The sweeteners can interact with sites localized in the extracellular or in the transmembrane region. Given that the three-dimensional structure of the receptor is not known, it has been predicted and evaluated using the method previously described. The interaction between the steviol glycosides and all the possible binding sites of the receptor, has been analyzed by means of an in-silico docking study, which allowed to identify the preferential binding site for the steviol glycosides. In particular, the transmembrane binding site seems to be the suitable for this class of compounds.

Protein structures are characterized by a peculiar flexibility which is fundamental for its activity. Following the line of the structure-activity analysis, the relationship between the dynamical properties and the function of some psychrophilic enzyme has been studied (chapter 4). The psycrophilic enzymes are adapted to work at low temperature, compared to the mesophilic enzymes, they show an high catatlytic efficiency at low temperature.
Supported by literature models and results, an accurate comparative study (psychrophile vs mesophile) of the thermodynamic properties of two different enzymes belonging to the elastases and the uracil-DNA-glycosylases families has been done. This study, carried out with molecular dynamics simulations, revealed, according to previous evidences, that the low temperature adaptation is related to the different flexibility of the psychrophilic compared to the mesophilic enzyme. This difference influences how the enzymes interact with their substrates.

The analysis of the protein structure can be useful when we are dealing with diseases which are dependent on mutations in a given protein. It help to identify how the mutation impairs the protein function and which is its impact on the disease.
As examples, here, we report three cases in which a structural study has been used to support biochemical and genetical data for the analysis of the impact of point mutations on the protein structure and function and its effect on the associated disease. In particular, we have studied three different serious rare diseases which involve grave metabolic disorder associated to point mutations in mithocondrial proteins.

The continuous increasing of the computing power and the huge quantity of biological information stored into the databases, leads to the computational methods to reach a development good enough to be considered a crucial support for the experimental methods. It is worth to underline the importance of the combined use both the approaches. The experimental methods require the processing and the analysis of the data, whereas the computational methods need the experimental data to be accurate.

Chapter 1

Introduction

The greatest challenge to any thinker is stating the problem in a way that will allow a solution.
Bertrand Russell (1872 - 1970)

1.1 The structure and function paradigm

Living organisms are complex systems made up of many different elements that interact each others in a highly-organized fashion. Hence, to understand how living systems work, a detailed knowledge of each element they are made up and each kind of interaction, is required [1]. Among the different components that constitute living systems, proteins represent one of the most important element. Proteins are complex macromolecules which hold many different functions, and the detailed knowledge of their roles reside in a deep cognition of the relationship between the three-dimensional (3D) structure of the protein and its function.
Indeed, the characterization of the protein functions requires the knowledge of its 3D structure, which can also help to explain not just the working of individual molecules, but of whole systems.
The first crude model of a protein 3D structure, the whale myoglobin [2], was obtained in 1957. Afterward, in the 1960 the first protein structure was solved at atomic detail. Since then, the number of structures solved each year has risen exponentially, giving us more than 51,900 structures as of July 2008 [3]. Despite the high number of known macromolecular structures, they represent only the

<1% of the deposited protein sequences [4].
There are now more than 600 completely sequenced genomes of cellular organisms, contributing to more than five million unique protein sequences in the publicly accessible database [5]. The experimental determination of the protein function would be hardly possible due to the huge amount of data.
Currently, approximately 20%, 10%, 7% and 1% of annotated proteins in the *Homo sapiens*, *Drosophila melanogaster*, *Mus musculus*, and *Caenorhabditis elegans* genomes, respectively, have been experimentally characterized [5].
However, as the volume of data has increased, computational methods for the protein structure and function prediction have increased too. Computational methods have become more and more accurate during the years and now represent a fundamental support for the experimental techniques. Experimental and computational techniques take a mutual advantage one from the other, i.e. experimental methods need the computational treatment of huge amount of data, and theoretical models interpreting the results and direct the development of new experiments. On the other hand, computational methods require experimental informations to be accurate and to correctly reproduce the experimental results. Hence, it turns out to be noticeable the importance of the computational techniques in the contribution that could be given to the experimental methods.

1.2 Protein structure prediction

Prediction is very difficult, especially about the future.
Niels Henrik David Bohr (1885 - 1962)

In the last years several different methods have been developed for the prediction of the 3D structure of proteins [6, 7, 8]. All these methods can be classified into two broad categories: the template-based modelling (TBM) and the free modelling (FM) [6]. Here, we concisely describe the rationale of each category, for more detailed explanations see reviews [6, 7, 8, 9, 10], and book [11].

1.2.1 Template-based modelling

One of key contributions derived from the knowledge of protein 3D structure is the understanding of protein evolution. As two protein sequences diverge over evolutionary time, their structures will tend to remain mostly the same; the overall fold of the protein will be conserved even where there are significant insertions and deletions in the sequence [3]. As a consequence, a three-dimensional model of a protein of interest (target) can be built from related protein(s) of known structure [template(s)] that share statistically significant sequence similarity [12].
The traditional comparative modelling procedure consists of many serial steps usually repeated iteratively until a satisfactory model is obtained: *i.* finding one or more suitable template protein(s) related to the target; *ii.* aligning target and template sequences; *iii.* identifying structurally conserved regions; *iv.* building structural frameworks by copying the aligned conserved regions or by satisfying the spatial restraints from templates; *v.* constructing the unaligned regions, usually loop regions; adding side-chain atoms.
The identification of a suitable template(s) can be performed by a sequence search tool such as **BLAST** [13] or **PSI-BLAST** [14], which allow to identify similar proteins using only sequence information. This approach, named comparative modelling, is unable to identify evolutionary distant homologue proteins (remote homologues), because they share low sequence similarity with the target protein. The introduction of more sophisticated methods [15] that derive their power from profile-profile comparison and the effective use of structural information has significantly increased the remote homologue detection capability [12]. This kind of approach is referred to as fold-recognition (or

BLAST
Basic Local Alignment Search Tool. Local alignment method that allows to search a target protein (or nucleic acid) against a sequence database

PSI-BLAST
Position Specific Iterated - BLAST. Method that allows to search a target protein (or nucleic acid) against a sequence database using sequence profile

threading), because they attempt to detect similarities between the target protein and the template(s) that are not accompanied by any significant sequence similarity (figure 1.1).

1.2.2 Free modelling

The existence of similar structures in the Protein Data Bank (PDB) [4] is a necessary condition for the successful template based modelling. An important question is how complete the current PDB structure library is . When structural analogs do not exist or could non be identified in the PDB library, the structure prediction has to be generated without template. This kind of predictions has been termed as *ab initio* or *de novo*, or in general free modelling [6]. One of the most successfully technique of free modelling is the fragment assembly used by Baker and coworkers in the development of the ROSETTA [16] software. In the latest version of ROSETTA [17, 10] structures were built in two stages: in the first step structures were assembled in a reduced knowledge-based model, then in the second stage, Monte Carlo simulations with an all-atom physics-based potential are performed to refine the details of the low-resolution models. Another successful free modelling approach, called TASSER [18], constructs 3D models based on a purely knowledge-based approach.
Significant efforts have been made on the purely physics-based approach for protein folding and structure prediction [6]. This kind of approach try to develop proper parametrized physics-based functions that are able to correctly describe the potential energy surface of a protein, hence the 3D structure is predicted by a complete simulation of the folding process. Although a purely physics-based *ab-initio* simulation has the advantage of revealing the pathway of protein folding, the best current free modelling results come from those methods which combine both knowledge-based and physics-based approaches [6].

1.2.3 Usefulness of comparative models

Comparative models may be used to identify residues involved in catalysis, binding or structural stability, to examine protein-protein or protein-ligand interactions, to correlate genotypic and phenotypic mutation data, and to guide experimental design. When either the similarity between target and template is low (less than 30%) or the models were built using free modelling techniques, they should only be used as structural frameworks to study overall feature of the protein and certainly not for deriving accurate measures of distances, en-

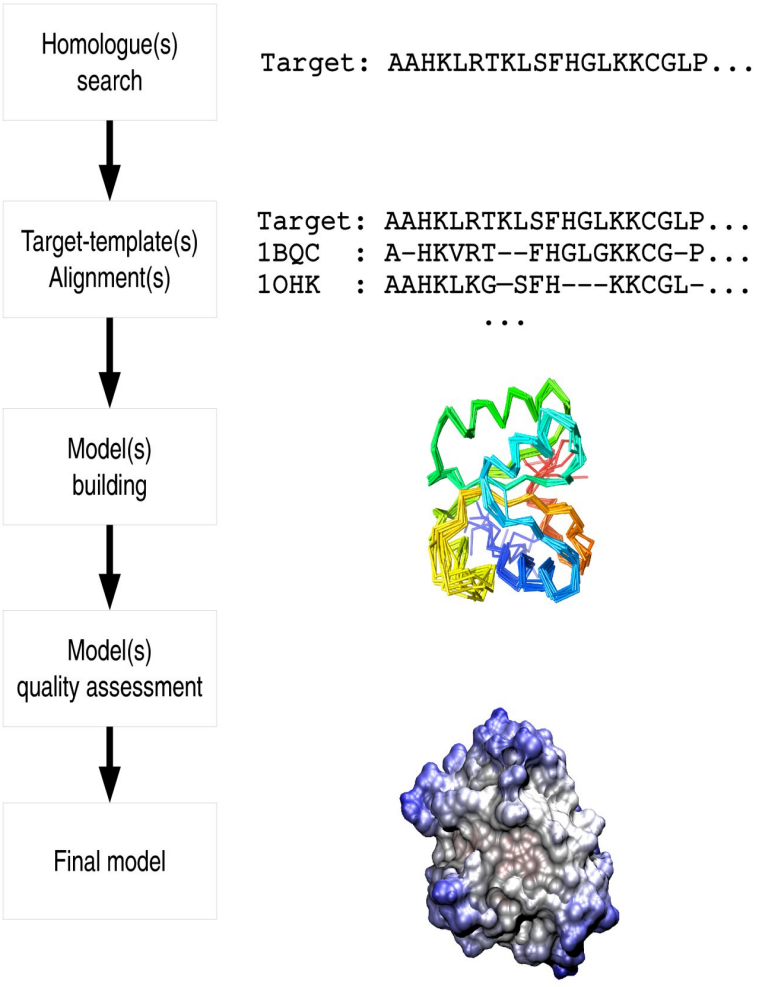


Figure 1.1: Homology modelling diagram.

ergies or for drug design tasks [19]. Moreover, it should be always taken into account that the accuracy of the model is not uniform throughout the structure and that functionally important regions are likely to be better conserved, at least for orthologous proteins, than the rest of the structure [19]. In general, the use of the predicted model strictly depends on its accuracy, that in turn is related to the method used for its prediction. Figure 1.2 shows, the main uses of three-dimensional structure of proteins depending on the method they have been obtained, are shown.

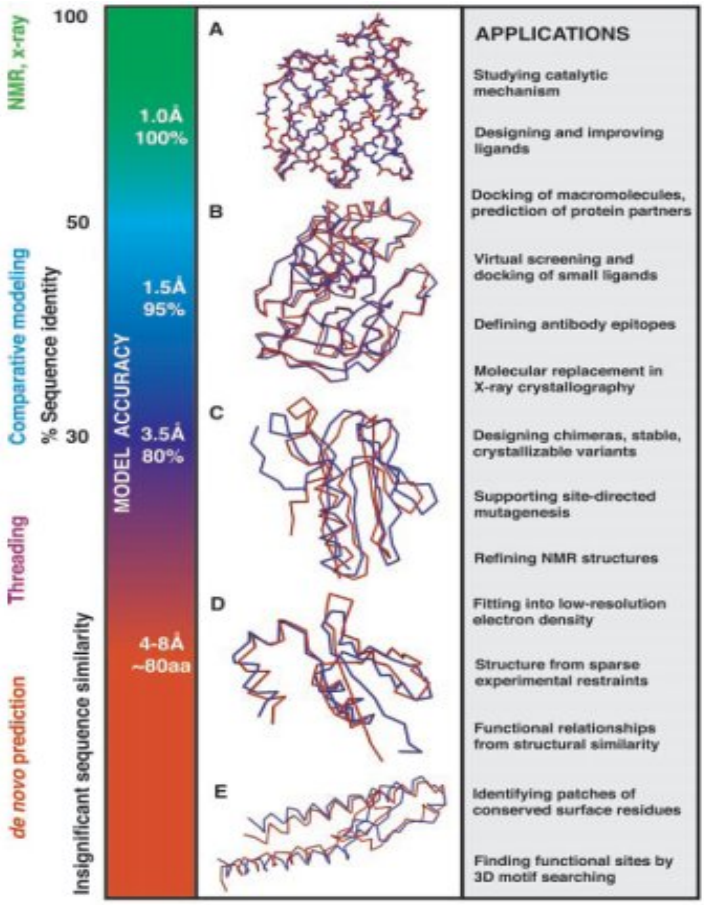


Figure 1.2: Accuracy and application of protein structure models. Shown are the different ranges of applicability of comparative protein structure modelling, threading, and de novo structure prediction; the corresponding accuracy of protein structure models.

The state-of-the-art of the current modelling methods, has been biannual monitored in the CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiment since the 1994 [20]. Experimental structural biologists who are about to solve a protein structure are asked to render available the sequence of the protein before the crystal structure is stored in the PDB. Than predictors produce models for this protein using computational methods and store it into the CASP database. At the end of the CASP session the predictions are assessed comparing them to crystal structure, and some conclusions about the state of the art of the different methods, are drawn. The experiment runs blindly, that is, the assessors do not know who the predictors are until the very end of the experiment. At the same time to the CASP another experiment named CAFASP [21] has started to assess the performance of the automatic server for protein structure prediction. This allow to evaluate how much human expert knowledge is important to obtain better models. As mentioned before, the accuracy of a protein structure model determines its usefulness in investigating a biological problem. Hence it is of primary importance an effective method able to assess the quality of a model. The relevance of this task is proven by the introduction in the last CASP (CASP7 [22]), of a category to objectively test and compare methods for model quality assessment.

1.3 Protein quality assessment

A little inaccuracy sometimes saves tons of explanation.
Hector Hugh Munro (Saki) (1870 - 1916)

When we talk about protein quality assessment methods, first of all we need to properly define the term "quality" of a protein model, in order to develop methods able to compute or predict it. We can refer to quality as a measure that quantifies the accuracy of a 3D model, i.e. the similarity of the model with the experimentally determined native structure (the X-ray structure is usually taken as native structure). The more the model is similar to the experimental structure the more it will be accurate. At first sight, it might seem that the evaluation of the correctness of a model is a straightforward task once the experimental structure is available, but matters are not so easy. To get a measure of similarity the first task is to optimally superimpose the model on the experimental structure. Protein structures, however, have an intrinsic flexibility which implies the existence of more than one optimal superposition. The simplest way is to perform a least square fitting on all the atoms of the two proteins and compute the rmsd. It is defined as the square root of the squared differences between the coordinates of corresponding atoms, and therefore, it will weight more regions that are not well superimposed with respect to the rest. Some regions may be more flexible than others, and the rmsd may penalize them to give an underestimation of the quality of the models. We need to ascertain that our quality measure takes this into account and does not penalize a model if it does not correctly reproduce regions of the experimental structure that have significant experimental uncertainty [23, 19]. Many quality measures have been developed to accomplish this task: GDT-score [24], LG-score [23], TM-score [25] MaxSub [26], a comprensive description of the model quality assessment measure but the TM-score is given in ref [23].

Decoys
Wrong protein structure obtained with a computational method. The decoys that are more similar to the native structure are defined native-like structure.

The definition of the measure of quality is only half of our goal; we need to develop a method able to *predict* the model quality without the existence of the native structure. Many methods of protein models validation have been developed during the last years [27, 28, 29, 30, 31, 32, 33, 34]. Most have focused on finding the native structure or native-like structures in a large set of **decoys** , while a few have focused directly on the quality assessment problem. These methods predict the overall global quality of a protein structure model by analyzing various structural features, such as non-bonded interactions, solvent

exposure, secondary structure, hydrophobic interactions, and stereochemical features.
The first technique has been to use methods such as PROCHECK [27] and WHATCHECK [35] in order to evaluate stereochemistry quality of protein model. These methods were developed in order to check the extent to which a model deviates from real X-ray structures based on a number of observed measures. However, such evaluations are often insufficient to differentiate between stereochemically correct models.
A variety of energy-based programs have been developed more specifically for the discrimination of native-like models from decoy structures. These programs were based either on empirically derived physical energy functions or statistical potentials derived from the analysis of known structures [28, 34, 36]. For some time, statistical potential methods such as PROSAII [37] and VERIFY3D [38] had been popular use for rating model quality. More recently, methods such as PROQ [32] , FRST [39], MODCHECK [40], and AIDE [41] have proved to be more effective at enhancing model selection. These methods combine many structural parameters related to the model accuracy by means of parametric models (usually linear models) or learning-based models such as neural networks or support vector machines. Here, we have developed a novel software (AIDE : Artificial Intelligence Decoys Evaluator) for protein quality assessment that is able to predict different quality indicators given a predicted 3D protein model [41]. The software is based on multiple neural networks as they are able to quantitatively describe complex relationships, such as that between the 3D protein structure and its accuracy. A detailed description of the AIDE software is given in chapter 2; to a better understand of how it works a brief description of the neural networks is given here.

1.3.1 Artificial neural networks: a brief introduction

The artificial neural networks are mathematical models that were developed in analogy to a network of biological neurons. In the brain, the highly interconnected network of neurons communicates by sending electric pulses through the neural wiring consisting of axons, synapses and dendrites. In 1943, McCulloch and Pitts modeled a neuron as a switch that receives input from other neurons and, depending on the total weighted input, it is either activated or remains inactive. Each input coming from another cell is multiplied by a weight which corresponds to the strength of the signal between the two neurons. These weights can be both positive (excitatory) and negative (inhibitory) [42].

It was shown that simple networks of such neuron models, named perceptrons, have properties similar to the brain: they can perform pattern recognition and they could learn from examples [43, 44].
In the 1969 Minsky and Papert [45], showed that simple perceptrons could solve only the very limited class of linearly separable problems. Nonetheless, the development of the error back-propagation method [46] allows to train more complex networks, being able to solve problems not linearly separable which are a more interesting and more diffused class of problems.
A neuron model is described in figure 1.3. The output of a neuron is the weighted sum of the input vector eventually filtered by a function f. More formally, given an input vector $\vec{x} = \{x_i\}, i = 1, ..., n$ the output of the neuron j is given by

$$o_j = f\left(\sum_i^n w_{i,j}x_i + b_j\right) \tag{1.1}$$

Where $w_{i,j}$ are weights form the input element i to the neuron j, b_j is a threshold value associated to each neuron, and f is the function of the neuron.

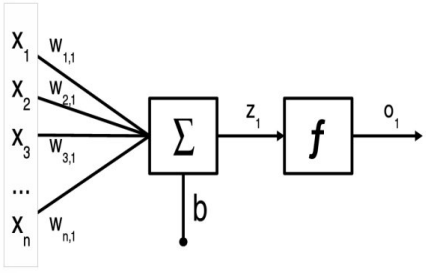


Figure 1.3: Single neuron model. x elements of the input vector, b neuron bias, z weighted sum, f neuron function, o neuron output.

Many neurons can be combined to create more complex networks. Depending on the type of neurons and how the neurons are connected to each others, different kind of neural networks can be created. Here, we only focus on the most common type of neural network that is the feed-forward neural network, in which each neuron of a layer is connected to all the neurons of the next layer and the information flows from the input to the output without loops. For an overall overview of the neural networks types see ref. [47, 48]. An example of feed-forward neural network is showed in figure 1.4.

In a multi-layers feed-forward neural network we always have an input vector,

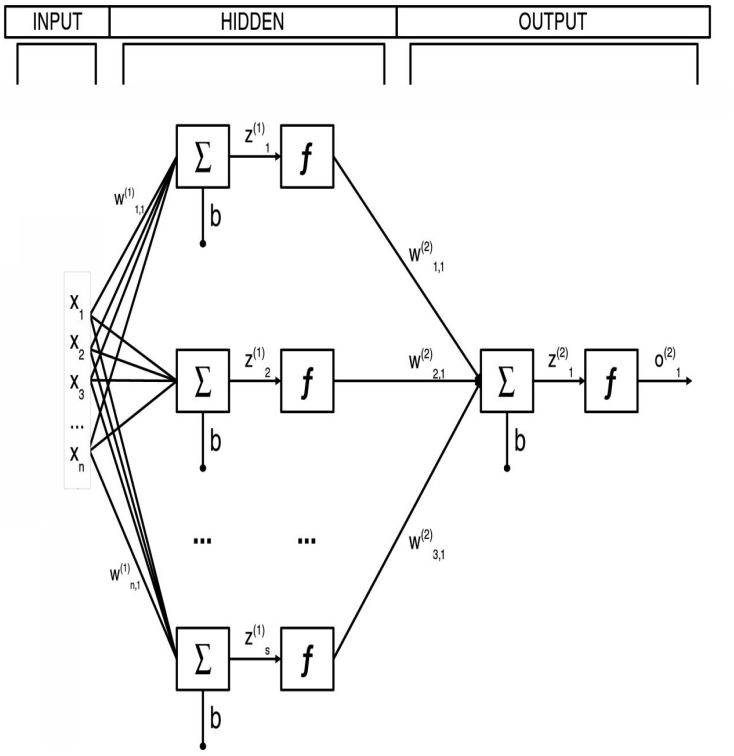


Figure 1.4: Simple example of feed-forward neural network. x elements of the input vector, b neuron bias, w network weights, z weighted sum, f neuron function, o neuron output.

Hidden
The hidden layer is composed of neuron whose output is not visible by the user.

one or more **hidden** layers, and one output layer composed of one or more neurons. For example the output computed by the network showed in figure 1.4 is the following:

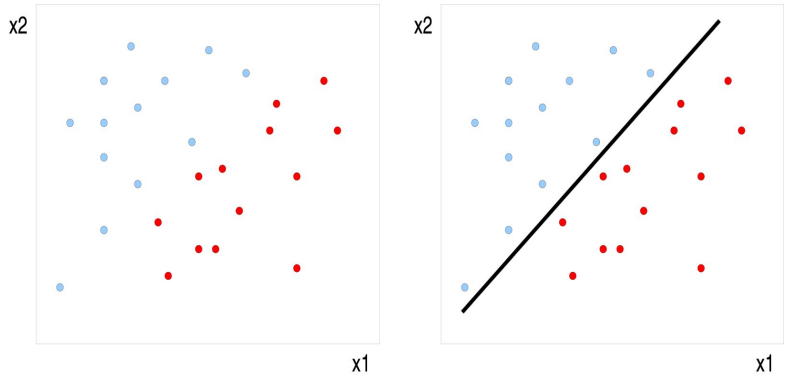
$$o_j = f^{(2)}\left(\sum_i^s w_{j,k}^{(2)} f^{(1)}\left(\sum_i^n w_{i,j}^{(1)} x_i + b_j\right) + b_k\right) \tag{1.2}$$

Where $w_{i,j}^{(1)}$ are the weights of the hidden layer, $w_{j,k}^{(2)}$ are the weights of the output layer, b_j and b_k are the threshold values associated to the hidden neurons and to the output neuron, respectively. $f^{(1)}$ is the function of the neurons of the hidden layers (assumed to be the same for each neuron) and $f^{(2)}$ is the function of the neuron of the output layer.

How they work: the two dimensional classification problem

To understand how a neural network works we can consider a simple example: we want to classify some objects into two classes. Each object is described as a two dimensional vector. Data can be represented as points in a two dimensional plane as shown in figure 1.5(a).

We use only one neuron with a two dimensional input vector and a single



(a) Find a function that best divides the the data into the two classes (red or blue)

(b) The black line represents the fitted linear function (the linear neuron) that best divides the data into the two classes (red or blue)

Figure 1.5: Two dimensional classification example.

output value. Using the equation 1.1, and considering a linear function, we

can compute the output of this neuron as the sum of the product between the weights and the input values, $output = \left(\sum_{i=1}^{2} w_i x_i + threshold\right)$. We need some training data which are described by a two dimensional vector and a value which represents the class the data belongs to. Training starts by setting all the weights in the neuron to small random numbers. Now, for each input example the network gives an output, which starts randomly. We measure the squared difference between this output and the desired output (the known class). The sum of all these numbers over all training examples is called the total error of the network. If this number was zero, the network would be perfect, and the smaller the error, the better the network. By choosing the weights that minimize the total error, one can obtain the neural network that best solves the problem at hand. In this example we identify the best line that divides the points into two classes, as showed in figure 1.5(b).
The power of neural networks appear when we are dealing with complex non-linear classification or regression problems, indeed, multi-layers feed-forward neural networks are universal function approximator, as proven by the Cybenko theorem [49]. The failure to map a function arises from poor choices for parameters or an insufficient number of hidden neurons. In a feed-forward neural network with hidden units the algorithm used for adjusting the weights and thresholds is called, as previously mentioned, back-propagation [46]. Many other algorithms other than the back-propagation have been developed to train neural networks, here we get a little bit inside of the back-propagation algorithm because it is the first developed and one of the most used.

Back-propagation In back-propagation we have to find the weights w that minimize the total error function

$$E = \frac{1}{2} \sum_{k \in output} (o_k - t_k)^2 \tag{1.3}$$

This task is done iteratively updating the weights using the steepest descent method [a]. The weights are updated starting from the output layer going back to the hidden layers. There are some learning parameters that need tuning when using back-propagation, and there are other problems to consider. (See appendix 7.1.1 for a detailed explanation for the algorithm.) For instance, gradient descent does not guaranteed to find the global minimum of the error

[a] **S**teepest descent:
Given a function $E(w)$, the parameters w_{min} that minimize E can be found iteratively by searching on the steepest descent direction of the function: $w_{new} = w_{old} - \eta \frac{\partial E(w)}{\partial w}$, where η define the stepsize. The process stops when a termination criteria is reached

function, so the result of the training depends on the initial values of the weights. Moreover another problem is the over-fitting. Over-fitting occurs when the network has too many parameters to be learned from the number of examples available, that is, when a few points are fitted with a function with too many free parameters [47]. To estimate the generalization performance of the neural network, one needs to test it on independent data, which have not been used to train the network. This is usually done by cross-validation, where the data set is split into, for example, ten sets of equal size. The network is then trained on nine sets and tested on the tenth, and this is repeated ten times, so all the sets are used for testing. This gives an estimate of the generalization ability of the network; that is, its ability to classify inputs that it were not trained on [47, 48, 42].

1.4 Analysis of 3D structure

We may, I believe, anticipate that the chemist of the future who is interested in the structure of proteins, nucleic acids, polysaccharides, and other complex substances with high molecular weight will come to rely upon a new structural chemistry, involving precise geometrical relationships among the atoms in the molecules and the rigorous application of the new structural principles, and that great progress will be made, through this technique, in the attack, by chemical methods, on the problems of biology and medicine.
Linus Carl Pauling (1901 - 1994)

As previously discussed, protein structures have a central role in many different fields. In the pharmaceutical industry, for example, the knowledge of a protein structure is fundamental for the development of new drugs. Indeed, to obtain specific drugs one must deeply know how they interact with their target (often a protein), and this task can be exclusively done knowing the 3D structure of the target.
A related field in which the knowledge of the 3D structure of a protein has assumed a central role is the food industry. The development of new food additives (flavours, sweeteners, stabilizer, emulsifier, etc.) and the study of the existing ones may requires to know the way they interact with a specific receptor.
Moreover the knowledge of the three-dimensional structure of a protein can be useful for studying structural, dynamical and thermodynamical properties of the protein itself or for analyzing the effect of mutations on the protein function. Some of the possible analysis that require the knowledge of the 3D structure of a protein are listed in table 1.1.

Some of the analysis listed in table 1.1 have been used into this work, hence we decide to briefly describe them in the following sections.

1.4.1 Drug design and docking

The development of new drugs is undoubtedly one of the most challenging tasks of today's science. The very high complexity of the problem requires the collaboration of pharmaceutical industry, biotech companies, regulatory authorities, and academic researchers. In the last few years, the field of drug development

has become more productive than ever before. This rapid development is basically due to the parallel increase of the experimental high-throughput techniques and the computational methods. Genomic and proteomic studies allow to obtain information related to the target; in particular, they are important for the identification of new putative targets, or to analyze the transmission pathways on which a given target is involved with. Moreover, the emergence of combinatorial chemistry enables the production of very large libraries of compounds, which could be tested using the continuously developed high-throughput tests [50, 51]. A variety of computational approaches can be very helpful at different stages of the drug-design process: in the first step, it is important to reduce the number of possible ligands, while at the end, during lead-optimization stages, the emphasis is on decreasing experimental costs and reducing times [52]. Many enhancement have been made in the computational approaches including protein flexibility, refinement of the final complexes, and estimation for the binding free energies. Molecular dynamics (MD) simulations have played a dominant role to improve docking procedures;

Docking

Docking is designed to find the correct conformation of a ligand and its receptor. This task is not trivial at all because several factors influence the binding process. In particular the mobility of both the ligand and the receptor, the effect of the protein environment on the charge distribution over the ligand, and their

Analysis	**SP**[1]	**DK**[2]	**MS**[3]	**NMA**[4]
Drugs design	X	X	X	X
Protein-protein interactions	X	X	X	X
Development of food additives	X	X	X	X
Dynamical properties	X		X	X
Thermodynamical properties	X		X	
Mutation analysis	X	X	X	
Mutation design	X	X	X	
Protein design	X	X	X	

Table 1.1: Examples of studies based on the three-dimensional structure of a protein. The main computational techniques used in each study are also shown. 1. SP: structure prediction, 2. DK: docking, 3. MS: molecular simulations (e.g. molecular dynamics, montecarlo, and brownian dynamics simulations), 4. Normal mode analysis.

interactions with the surrounding water molecules, highly complicate the quantitative description of the process. The idea behind this technique is to generate a comprehensive set of conformations of the receptor complex, and then to rank them according to their stability. The most popular docking programs include Delos[REF], DOCK [53], AutoDock [54],FlexX [55], and GLIDE [56], among others. From a theoretical point of view docking aims at correct prediction of the structure of the complex $[E + I] = [EI]$ under equilibrium conditions: $[EI]aq \rightleftharpoons [E]aq + [I]aq$ The free energy of binding (G) is related to binding affinity by equations 1.4 and 1.5:

$$G = -RTln(K_A) \tag{1.4}$$

$$K_A = K_i^{-1} = \frac{[E + I]}{[E][I]} \tag{1.5}$$

Posing
The process of determining whether a given conformation and orientation of a ligand fits the active site. This is usually a fuzzy procedure that returns many alternative results

Ranking
Docked complexes are evaluated using a scoring function and ordered starting from the best scored compound

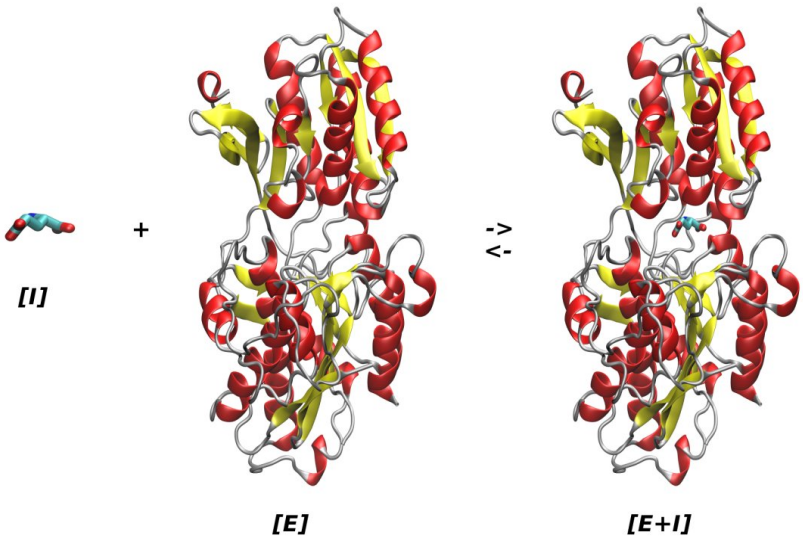


Figure 1.6: The figure illustrates the binding of a glutamate to the metabotropic glutamate receptor subtype 1 (mGluR1). The figure refers to the crystal structure of the glutamate receptor complexed with the glutamate. PDB ID: 1ewk, chain A.

Prediction of the correct structure **posing** of the [E+I] complex does not require information about K_A. However, prediction of biological activity **ranking** requires this information [57].

The commonly used "keylock" paradigm was put forward by Emil Fischer to explain the manner in which the enzyme hexokinase exerted its specificity. The "key-lock" theory said that the enzyme was a rigid negative of the substrate and that the substrate had to fit into to react. During the last years the increasing collection of structural data that describe protein structures and protein complexes shown that proteins are not rigid objects, but rather, they could be better represented as a collection of structures which are in equilibrium between each others. The "key-lock" theory has then be overtaken by a more "flexible" paradigm which represents a more accurate description of most biological complexes. The ligand protein interactions resemble more a "hand and glove" association, where both parts are flexible and adjust to complement each other (induced fit). Both can modify their shape and their complementarity to increase favourable contacts and reduce unflattering interactions, maximizing the total binding free energy [52, 57].

Considering the computational docking problem, we need a computational technique that is able to identify the ligand conformation and position (and eventually the target conformation) that maximize the free energy of binding, (on figure 1.7 the flowchart of a general docking procedure is shown).

Many methods have been developed to perform this task. It is possible to group the computational docking methods into three categories depending on the degrees of freedom that they take into account.

Rigid docking If both the ligand and the target are considered as rigid entity. Usually it is performed keeping fixed the target position and moving the ligand inside the binding pocket. The most frequently used search strategies are random search methods such as Monte Carlo[b], Tabu search [58][c], or genetic algorithms[d].

[b] **Monte Carlo algorithm**: *i.* Generate an initial configuration of a ligand in an active site consisting of a random conformation, translation and rotation. *ii.* Score the initial configuration. *iii.* Generate a new configuration and score it. *iv.* Use a criterion to determine whether the new configuration is retained. *v.* Repeat previous steps until the desired number of configurations is obtained.

[c] **Tabu search algorithm**: *i.* Make n small random changes to the current conformation. *ii.* Rank each change according to the value of the chosen fitness function. *iii.* Determine which changes are 'tabu' (that is, previously rejected conformations). *iv.* If the best modification has a lower value than any other accepted so far, accept it, if it is in the 'tabu'; otherwise, accept the best 'non-tabu' change. *v.* Add the accepted change to the 'tabu' list and record its score. *vi.* Go to the first

[d] **Genetic algorithms** are a class of computational problem-solving approaches that adapt the principles of biological competition and population dynamics. Model parameters are

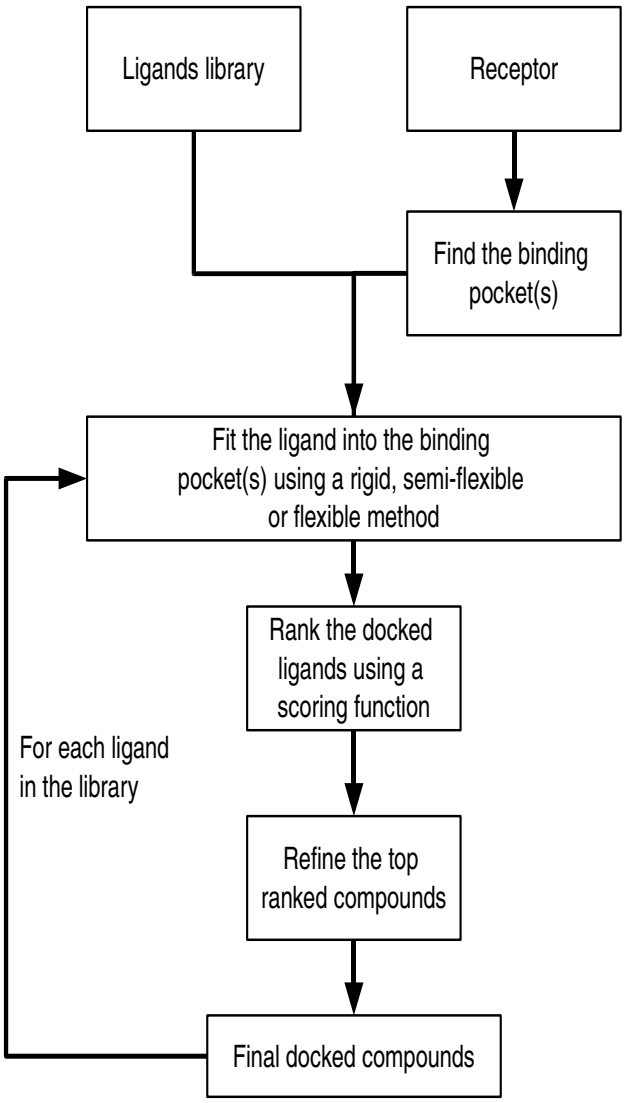


Figure 1.7: Molecular docking flowchart.

Semi-flexible docking In the semi-flexible docking the ligand is allowed to modify its conformation but the target keeps a rigid structure. The searching algorithm could be stochastic, systematic search methods, or deterministic simulation search methods (molecular dynamics). The systematic search tries to explore all the possible positions inside the binding pocket and all the conformations of the ligand. A true systematic search is unfeasible because the combinatorial explosion of the possible conformations [57].

Flexible docking If both the ligand and the target are allowed to modify their conformations. The introduction of the target flexibility in computational docking methods is a difficult task because it leads to an impressive increase of the computational costs. Many different approaches have been developed to introduce the flexibility without overloading the computational costs [57, 52, 59].

Combination with molecular dynamics A common method is to combine the fast and inexpensive protocols with accurate and costly molecular dynamics simulations. At first fast methods were used to rank ligands according to an energy function and excluding the ligands that are almost surely unable to bind, thus molecular dynamics (MD) simulations (which allow to incorporate protein flexibility) is performed on the top ranked compounds to obtain more accurate estimation of the free energy of binding [59, 52].

Ensemble methods Another more cost effective way to introduce the flexibility is to start from different conformations of the target protein, and then performing fast docking procedures which consider the target structure as a rigid object.

The way to produce the different starting target conformations varies according to the method used. One common way is to perfor a **normal mode analysis** (see chapter 3 for details) on the protein structure and taking some relevant conformations. Another way is to perform molecular dynamics simulation and to extract from the trajectory some significant frames [59, 52].

encoded in a 'chromosome' and stochastically varied. Chromosomes yield possible solutions to a given problem and are evaluated by a fitness function. The chromosomes that correspond to the best intermediate solutions are subjected to crossover and mutation operations analogous to gene recombination and mutation to produce the next generation. For docking applications, the genetic algorithm solution is an ensemble of possible ligand conformations

Scoring function The evaluation and ranking of predicted ligand conformations is a crucial aspect of structure-based virtual screening. The development of reliable scoring functions is a fundamental task because it must be able to discriminate between incorrect pose to the correct one. Scoring functions can be divided into three classes: Force field based, empirical scoring functions, and knowledge-based [57].

Force fields based Force field based scoring functions uses physics-based energy functions to compute the binding energies between target and ligands. Force field are mathematical models properly parametrized to describe the potential energy of a molecule. The functional form of the force fields follows from quantum mechanical calculations [60]. For a short description of the force field see 1.4.2.

Empirical scoring functions These scoring functions are composed by a sum of several parametrized scoring functions fitted to reproduce experimental data, such as binding energies and/or conformations. The design of empirical scoring functions is based on the idea that binding energies can be approximated by a sum of individual uncorrelated terms. The coefficients of the various terms are obtained from regression analysis using experimentally determined binding energies and structural information [57].

Knowledge-based scoring functions This type of approach exploits the ever increasing knowledge base of experimentally determined target-ligand structures relying explicitly or implicitly on Boltzmann's principle: frequently observed states correspond to low energy states of the system [60].

Refinement of the docked complexes

The most accurate and convenient approach to address the docking problem seems to be a two-step protocol. Fast and less accurate algorithms are first used to scan large databases of molecules and to reduce their size to a reasonable number of hits. Thus a more accurate and time-consuming method which can refines the conformation of the complexes is applied [61, 62]. Molecular dynamics simulations is a valid alternative for structural refinement of the final docked complexes. As previously mentioned MD is able to incorporate flexibility of both ligand and receptor, and enhancing complementarity between them, and thus accounting for induced fit. Moreover, the evolution of the complexes over the simulation indicates its stability: incorrectly docked structures are unstable

and lead to the disruption of the complex, while realistic complexes will show stable behaviour. In addition, MD allows the incorporation of explicit solvent molecules and their interactions in the simulations of the docked systems is very important for understanding the role of water and its effect on the stability of the ligand-protein complexes [52].

1.4.2 Molecular dynamics simulations and protein functions

The importance of the molecular dynamics simulations can be realized looking at the number of times that has been cited in the previous sections. It represents an extremely useful technique that could be used for several kind of studies.
Molecular dynamics simulations is a computational method used in the theoretical study of biological molecules, because it permits to calculate the time dependent behaviour of a molecular system. MD simulations have provided detailed information on the fluctuations and conformational changes of proteins and nucleic acids (the time of biological process in proteins is reported in Table 1.2). These methods are now routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes, and to relate the results with the function of the protein. It is also possible to study the effect of explicit solvent molecules on protein structure and stability to obtain time-averaged properties of the biomolecular system, such as density, conductivity, and dipolar moment, as well as interactions energies and entropies. MD is useful not only for rationalizing experimentally derived information at the molecular level, but it is very helpful in the determination of the structure by X-ray or NMR. In this work the molecular dynamics simulations were used as method for analyzing biological fluctuations of psychrophilic proteins around the native state (see chapter 4), as a tool for refining predicted protein models obtained by homology modelling(see chapter 3) and as a docking method for identifying and analyzing the binding of small molecules to protein target (see chapter 3).

Molecular dynamics simulations: a brief introduction

The Molecular dynamics method was first introduced by Alder and Wainwright in the late 1950's [63, 64] to study the interactions of hard spheres: many important insights concerning the behaviour of simple liquids emerged from their studies. The next major advance was achieved in 1964, when Rahman carried out the first simulation using a realistic potential for liquid argon. The first Molecular dynamics simulation of a realistic system was done by Rahman and Stillinger in their simulation of liquid water in 1974 [65] and the first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) [66].

Molecular dynamics simulations are in many respects very similar to real experiment; when we perform a real experiment, we proceed as it follows. We

Local Motions (0.01 to 5 Å, 10-15 to 10^{-1} s)	Atomic fluctuations Sidechain Motions Loop Motions
Rigid Body Motions (1 to 10 Å, 10^{-9} to 1s)	Helix Motions Domain Motions (hinge bending) Subunit motions
Large-Scale Motions ($>$ 5 Å, 10^{-7} to 10^{4} s)	Helix coil transitions Dissociation/Association Folding and Unfolding

Table 1.2: Time of motions in proteins. The amplitude of motions is expressed as a RMSD value.

prepare a sample, we connect it with the measuring instrument, and we measure the property of interest during a certain time interval. In MD we follow exactly the same approach. First, we prepare the sample: we select a model system consisting of N particles and we solve Newton's equations of motion for this system until the properties of the system no longer change with time (system equilibration). After equilibration, we perform the actual measurement.

Statistical Mechanics Molecular dynamics simulations generate information at microscopic level, including atomic positions and velocities. However, this kind of information cannot be compared with the experimental data, because no real experiment provides us with such detailed information. A typical experiment measures an average property, such as pressure, energy, heat capacities, etc., averaged over a large number of conformations and, usually, over the time of the measurement. The conversion of this microscopic information to macroscopic observables requires to treat the systems from a statistical point of view. In **classical** molecular mechanics each microscopic state is uniquely defined by its coordinates and momenta, and a Molecular dynamics simulation generates a sequence of points in **phase space** as a function of time; these points belong to the same ensemble, and they correspond to the different conformations of the system. An ensemble is a collection of all possible configurations which have different microscopic states but an identical macroscopic or thermodynamic state [67].

Classical
The word classical means that the electrones are not explicitely considered and the nuclear motion obeys the laws of Newton's classical mechanics

Phase space
6N-dimensional space defined by coordinates $\mathbf{r}^N$ and momenta $\mathbf{p}^N$

There exist different ensembles with different characteristics:

- Microcanonical ensemble (NVE): the thermodynamic state characterized

by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. This corresponds to an isolated system.

- Canonical Ensemble (NVT): this is a collection of all systems whose thermodynamic state is characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T.
- Isobaric-Isothermal Ensemble (NPT): this ensemble is characterized by a fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T.
- Grand canonical Ensemble (mVT): the thermodynamic state for this ensemble is characterized by a fixed chemical potential, m, a fixed volume, V, and a fixed temperature, T.

In statistical mechanics, averages corresponding to experimental observables defined in terms of ensemble averages, which is the average taken over a large number of replicas of the system considered simultaneously. Given a particular ensemble, an average of a quantity $f(\mathbf{r}^N, \mathbf{p}^N)$ over all possible states is called *ensemble average* and it is denoted as $\langle f(\mathbf{r}^N, \mathbf{p}^N)\rangle$. Considering, for example, the NVE ensemble in classical mechanics the ensemble average can be computed averaging over all conditions compatible with the imposed values of N, V, and E.

$$\langle f(\mathbf{r}^N, \mathbf{p}^N)\rangle = \frac{\int_E f(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N}{\Omega(N, V, E)} \tag{1.6}$$

where $\Omega(N, V, E) = \int_E d\mathbf{r}^N d\mathbf{p}^N$. In Molecular dynamics simulations we can study the average behaviour computing the natural time evolution of the system and averaging the quantity of interest $f(\mathbf{r}^N, \mathbf{p}^N)$ over a sufficiently long time. This average is called *time average* and is denoted as $\overline{f(\mathbf{r}^N, \mathbf{p}^N)}$ to distinguish it from an ensemble average, and can be computed as

$$\overline{f(\mathbf{r}^N, \mathbf{p}^N)} = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\mathbf{r}^N, \mathbf{p}^N; t)\, dt \tag{1.7}$$

Combining the equations 1.6 and 1.7 we can obtain the following fundamental equation of statistical mechanics (1.8) that is usually referred to as "ergodic hypothesis" [68] and it states that the time average equals the ensemble average:

$$\overline{f(\mathbf{r}^N, \mathbf{p}^N)} = \langle f(\mathbf{r}^N, \mathbf{p}^N)\rangle_{ensemble} \tag{1.8}$$

The basic idea is that if the system is free to evolve in time indefinitely, that system will eventually pass through all possible states. Because the simulations

are of fixed duration, one must be certain to sample a sufficient amount of phase space. A Molecular dynamics simulation must be sufficiently long so that enough representative conformations have been sampled.
The Molecular dynamics simulation method is based on Newton's second law or the equation of motion, $F = ma$, where F is the force exerted on the particle, m is its mass and a is its acceleration. From a knowledge of the force on each atom, it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time. From this trajectory, the average values of properties can be determined. The method is deterministic; once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past. Consider a system consisting of N particles moving under the influence of the internal forces acting between them. The spatial positions of the particles as functions of time will be denoted by $\mathbf{r}^N(t) = (\mathbf{r}_1(t), ..., \mathbf{r}_N(t))$, and their velocities, $\mathbf{v}^N(t) = (\mathbf{v}_1(t), ..., \mathbf{v}_N(t))$. If the forces, $\mathbf{F}^N = (\mathbf{F}_1, ..., \mathbf{F}_N)$, on the N particles are specified, then the classical motion of the system is determined by Newton's second law

$$m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i \tag{1.9}$$

where $\mathbf{F}_i$ is the force exerted on particle i, m_i is the mass and $\ddot{\mathbf{r}}_i$ is the acceleration (second derivative of the position) of particle i respectively.

The total energy of an isolated system is defined by its Hamiltonian $\mathcal{H}(\mathbf{p}^N, \mathbf{r}^N)$ which is a function of the coordinates $\mathbf{r}^N$ and momenta $\mathbf{p}^N$ of the constituting particles: $\mathcal{H}(\mathbf{p}^N, \mathbf{r}^N) = \mathcal{K}(\mathbf{p}^N) + \mathcal{U}(\mathbf{r}^N)$ where $\mathcal{K}(\mathbf{p}^N)$ is the kinetic energy and $\mathcal{U}(\mathbf{r}^N)$ is the potential energy of the systems. The force acting on each particle can also be expressed from the potential energy as:

$$\mathbf{F}_i = -\frac{\partial \mathcal{U}}{\partial \mathbf{r}_i} \tag{1.10}$$

The equation 1.9 constitutes a set of $3N$ coupled second-order differential equations. A unique solution to equation 1.9 is obtained by choosing a set of initial conditions, $\mathbf{r}^N(0)$,$\mathbf{v}^N(0)$. Newton's equations completely determine the full set of positions and velocities as functions of time and thus specify the classical state of the system at time t. Except in special cases (two-bodies systems and special cases of three-bodies systems), an analytical solution to the equations of motion, equation 1.9, is not possible. An MD calculation, therefore, employs an iterative numerical procedure, called a numerical integrator, to obtain an

approximate solution.
The atomic initial position $\mathbf{r}^N(0)$ are usually taken from the experimentally obtained structure such as the X-ray crystal structure of the protein or the solution structure determined by NMR spectroscopy. The initial velocities, $\mathbf{v}_i(0)$, are often chosen randomly from a Maxwell-Boltzmann distribution [e]

And the initial accelerations are determined combining the equations 1.10 and 1.9.

$$\mathbf{a}_i = -\frac{1}{m_i}\frac{\partial \mathcal{U}}{\partial \mathbf{r}_i} \tag{1.12}$$

Integration Algorithms Numerous numerical algorithms have been developed for integrating the equations of motion: position Verlet algorithm [69], Leap-frog algorithm [68], velocity Verlet algorithm [70], Beeman's algorithm [68].

A common way to derive the integration algorithms is by approximating the position and the velocity by a Taylor series expansion:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta t + \frac{1}{2}\mathbf{a}_i(t)\delta t^2 + \frac{\delta t^3}{3!}\dddot{\mathbf{r}}_i + \mathcal{O}(\delta t^4) \tag{1.13}$$

Where $\mathbf{r}_i(t)$ is the position, $\mathbf{v}_i(t)$ is the velocity (the first derivative with respect to time), $\mathbf{a}_i(t)$ is the acceleration (the second derivative with respect to time), etc.

$$\mathbf{v}_i(t+\delta t) = \mathbf{v}_i(t) + \mathbf{a}_i(t)\delta t + \frac{1}{2}\ddot{\mathbf{v}}_i(t)\delta t^2 + \frac{\delta t^3}{3!}\dddot{\mathbf{v}}_i + \mathcal{O}(\delta t^4) \tag{1.14}$$

Where $\mathbf{v}_i(t)$ is the velocity, $\mathbf{a}_i(t)$ is the acceleration (the first derivative of the velocity with respect to time), etc.

One of the first algorithm used for the molecular dynamic simulations is the position Verlet algorithm 1.15 [69],

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \tag{1.15}$$

which uses positions and accelerations at time t and the positions from time $t-\delta t$ to calculate new positions at time $t+\delta t$. The Verlet algorithm uses no explicit velocities. The advantages of the Verlet algorithm are that it is straightforward, and the storage requirement is modest. The disadvantage is that the algorithm

[e] The probability that an atom i has a velocity $v_{i,x}$ in the x direction at a temperature T is given by:

$$p(v_{i,x}) = \left(\frac{m_i}{2\pi k_B T}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2}\frac{m_i v_{i,x}^2}{k_B T}\right] \tag{1.11}$$

is of moderate precision. A more detailed description of the Verlet algorithms is in the appendix 7.2.1.
The integration algorithms must satisfy two fundamental properties that are imposed by the Newton's equations of motion.
One is the time-reversibility, that implies that we can reverse the trajectory into the phase space. And the other, the most important, is the conservation of the total energy of the system. For most of the molecular dynamics application the Verlet-like algorithms are perfectly adequate [68].

Force Fields In order to provide a picture of the microscopic behaviour of a system from the laws of classical mechanics, MD requires, as an input, a description of the interparticle interactions. The quality of the results of an MD simulation depends on the accuracy of this description. One common approach is to model the interactions with *empirical force fields*.

Empirical
The term empirical refers to the fact that the parameters of the individual functions are fitted to experimental data and/or ab initio quantum chemical calculations.

An **empirical** force field is built up from a set of equations (called the potential functions) properly parametrized to represent the potential energy of the system. The force field describes the potential energy ($\mathcal{U}$) of the system as a function of the atomic position of all atoms in the system. The value of energy is calculated as a sum of internal, or bonded, terms which describe the bonds, angles and bond rotations in a molecule, and a sum of non-bonded terms. which account for interactions between non-bonded atoms or atoms separated by 3 or more covalent bonds. The most common terms included in almost all force fields for biological molecules are the following.

- bonds: $E_b = \frac{1}{2}k_b(r_b - r_b^0)$
  this term represents the interaction between atom pairs where atoms are separated by one covalent bond. This is an approximation of the energy of a bond as a function of the displacement from the ideal bond length r_b^0. The force constant, k_b, determines the strength of the bond. Both r_b^0 and k_b values are specific for each pair of bound atoms, and they are often evaluated from experimental data.

- angles: $E_\theta = \frac{1}{2}k_\theta(r_\theta - r_\theta^0)$
  this term is associated with alteration of bond angles, θ, from an ideal value r_θ^0, which is represented by an harmonic potential as in the previous. Values of k_θ and r_θ^0 depend on chemical type of atoms constituting the angle.

- dihedrals: $E_\gamma = \frac{1}{2}k_\gamma[1 - \cos(n\gamma - \gamma^0)]$
  the torsion angle potential function models the presence of steric barriers

between atoms separated by 3 covalent bonds. The motion associated with this term is a rotation, described by a dihedral angle and coefficient of symmetry (n) around the middle bond.

- van de Waals: $E_{ij}^{VdW} = -\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}}$
  The van der Waals interaction between two atoms results from a balance between repulsive and attractive forces: the former arises at short distances where the electron-electron interaction is strong whereas the latter is longer range than the repulsion but as the distance become short, the replusive interaction becomes dominant. This gives rise to a minimum in the energy. Positioning of the atoms at the optimal distances stabilizes the system. The van der Waals interaction is most often modelled using the Lennard-Jones 6-12 potential which expresses the interaction energy using the atom-type dependent constants A and C.

- Coulomb: $E_{ij}^{coulomb} = \frac{C\ q_i\ q_j}{\epsilon\ r_{ij}}$
  The electrostatic interaction between a pair of atoms is represented by Coulomb potential; ϵ is the effective dielectric function for the medium and r_{ij} is the distance between two atoms having charges q_i and q_j.

The empirical force fields have several approximation and limitations, which may result in inaccuracies in the calculated potential energy.

Solvent Solvent, usually water, has a fundamental influence on the structure, dynamics and thermodynamics of biological molecules, both locally and globally. One of the most important effects of the solvent is the screening of electrostatic interactions. It is important to include solvent effects in a MD simulation. This can be done at several levels:

The simplest implementation uses a distance-dependent dielectric, which has been used for a long time as a convenient and cheap, but also fairly crude, approximation of solvent effects. There are more sophisticated recent efforts to modulate the dielectric screening as a function of the solvent-excluded volume, as in the EEF (effective energy function) 1 model [71], with improvements that adjust the screening function according to the distance of a charge site from the surface. Another simple and fully empirical implicit solvent model is based on the atomic solvent accessible surface area, and aims to reflect observed hydrophobicity or hydrophilicity depending on the residue and/or the atom types of solvent exposed atoms [72]. More sophisticated implicit solvent models make use of the Poisson theory [73] for calculating the electrostatic component of the free energy of solvation for a set of charges embedded in a low-dielectric cavity

surrounded by a solvent environment represented as a dielectric continuum [72]. The computational cost for solving the Poisson equation limits its use for many applications such as the Molecular dynamics. A widely used compromize between accuracy and computational cost makes use of the Generalized Born formalism [74], which is an approximation of the linearized Poisson equation. In some cases the implicit solvent models are not enough accurate to represent our systems, in these cases we can introduce explicitly the molecules of the solvent around our molecule. This leads to an increase of the computational costs but results in a more accurate treatment of the system.

Bibliography

[1] Ludwig Von Bertalanffy. *General System Theory: Foundations, Development, Applications*. George Braziller, 1968.

[2] JC Kendrew, RE Dickerson, BE Strandberg, RG Hart, DR Davies, DC Phillips, and VC Shore. Structure of myoglobin: a threedimensional fourier synthesis at 2a resolution. *Nature*, 185:422–427, 1960.

[3] Roman A. Laskowski and Janet M. Thornton. Understanding the molecular machinery of genetics through 3d structures. *Nature review genetics*, 9:141–151, 20078.

[4] 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 Acid Research*, 28:235–242, 2000.

[5] David Lee, Oliver Redfern, and Christine Orengo. Predicting protein function fromsequence and structure. *Nature review molecular cell biology*, 8:995–1005, 2007.

[6] Yang Zhang. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, 18:342–348, 2008.

[7] Jones D T. Protein structure prediction in the postgenomic era. *Curr. Opin. Struct. Biol.*, 10:371–379, 2000.

[8] Baker D and Sali A. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.

[9] Oliver C Redfern, Benoit Dessailly, and Christine A Orengo. Exploring the structure and function paradigm. *Curr. Opin.Struct. Biol.*, 18:394–402, 2008.

[10] Das R and Baker D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[11] David M Webster. *Protein Structure Prediction: Methods and Protocols*. Humana Press, 2000.

[12] Krzysztof Ginalski. Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, 16:172–177, 2006.

[13] Altschul Stephen F, Gish Warren, Miller Webb, Myers Eugene W, and Lipmanl David J. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[14] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, and Lipman D J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[15] Krzysztof Ginalski, Nick V Grishin, Adam Godzik, and Leszek Rychlewski. Practical lessons from protein structure prediction. *Nucleic Acid Research*, 33(6):1874–1891, 2005.

[16] KT Simons, R Bonneau, I Ruczinski, and D Baker. Ab initio protein structure prediction of casp iii targets using rosetta proteins. *Proteins*, Suppl 3:171–176, 1999.

[17] Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim D E, Meiler J, Misura K M, and Baker D.

[18] Wu S T, Skolnick J, and Zhang Y. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol*, 5:17, 2007.

[19] Domenico Cozzetto, Alejandro Giorgetti, Domenico Raimondo, and Anna Tramontano. The evaluation of protein structure prediction results. *Mol. Biotechnol*, 39:1–8, 2008.

[20] Moult J, Pedersen J, Judson R, and Fidelis K. A largescale experiment to assess protein structure prediction methods. *Proteins*, 23:ii–v, 1995.

[21] Fischer D, Elofsson A, and Rychlewski L. Olympic games of protein structure prediction; fully automated programs are being evaluated vis-a-vis human teams in the protein structure prediction experiment. *Protein engineering*, 13:667–670, 2000.

[22] Casp7, http://predictioncenter.org/casp7.

[23] Cristobal S, Zemla A, Fischer D, Rychlewski L, and Elofsson A. A study of quality measures for protein threading models. *BMC Bioinformatics*, 2:5, 2001.

[24] Zemla A, Venclovas C, Moult J, and Fidelis K. Processing and evaluation of predictions in casp4. *Proteins*, 45:13–21, 2002.

[25] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, 2004.

[26] N Siew, A Elofsson, L Rychlewski, and D Fischer. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 2000.

[27] R A Laskowski, M W MacArthur, D S Moss, and J M Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[28] Gatchell D, Dennis S, and Vajda S. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins*, 41:518–534, 2000.

[29] Vendruscolo M, Najmanovich R, and Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins*, 38:134–148, 2000.

[30] T Lazaridis and M Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, 288:477–487, 1999.

[31] MJ Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[32] Bjorn Wallner and Arne Elofsson. Can correct protein models be identified? *Protein Science*, 12:1073–1086, 2003.

[33] Felts A, Gallicchio E, Wallqvist A, and Levy R. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the opls all-atom force field and the surfacegeneralized born solvent model. *Proteins*, 48:404–422, 2002.

[34] Dominy B and Brooks C. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.*, 23:147–160, 2002.

[35] Hooft R W, Vriend G, Sander C, and Abola E E. Errors in protein structures. *Nature*, 381(6580):272, 1996.

[36] Shen M-Y and Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15:2507–2524, 2006.

[37] MJ Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[38] Eisenberg D, Luthy R, and Bowie J U. Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 277:396–404, 1997.

[39] S Tosatto. The victor/frst function for model quality estimation. *J. Comp. Biol.*, 12:1316–1327, 2005.

[40] Pettitt C S, McGuffin L J, and Jones D T. Improving sequence-based fold recognition by using 3d model quality assessment. *Bioinformatics*, 21(17):3509–3515, 2005.

[41] Mereghetti P, Ganadu M L, Papaleo E, Fantucci P, and De Gioia L. Validation of protein models by a neural network approach. *BMC Bioinformatics*, 9, 2008.

[42] Krogh A. What are artificial neural networks? *Nature Biotechnology*, 26(2):195–197, 2008.

[43] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[44] Rosenblatt F. Perceptron simulation experiments. *Proceedings of the IRE*, 48:301–309, 1960.

[45] ML Minsky and SA Papert. *Perceptrons*. MIT Press, Cambridge, USA, 1969.

[46] Rumelhart David E, Hinton Geoffrey E, and Williams Ronald J. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[47] Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, USA, 1995.

[48] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, USA, 1999.

[49] Cybenko G V. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[50] Hertzberg R P and Pope A J. High-throughput screening: new technology for the 21st century. *Curr. Opin. Chem. Biol.*, 2000.

[51] Fernandes P B. Technological advances in high-throughput screening. *Curr. Opin. Chem. Biol.*, 2:597–603, 1998.

[52] Hernan A, Bliznyuk A A, and Gready J E. Combining docking and molecular dynamic simulations in drug design. *Medicinal Research Reviews*, 26(5):531–568, 2006.

[53] Kuntz I D, Blaney J M, Oatley S J, Langridge R, and Ferrin T E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 288:161–269, 1982.

[54] Morris G M, Goodsell D S, Halliday R S, Huey R, Hart W E, Belew R K, and Olson A J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:16391662, 1998.

[55] Rarey M, Kramer B, Lengauer T, and Klebe G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470489, 1996.

[56] Friesner R A, Banks J L, Murphy R B, Halgren T A, Klicic J J, Mainz D T, Repasky M P, Knoll E H, Shelley M, Perry J K, Shaw D E, Francis P, and Shenkin P S. Glide: A new approach for rapid, accurate docking and scoring 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47:1739–1749, 2004.

[57] Douglas B Kitchen, Hlne Decornez, John R Furr, and Jrgen Bajorath. Docking and scoring in virtual screening for drug discovery: nethods and applications. *Nature Reviews*, 3:935–949, 2004.

[58] Cvijovicacute, Djurdje, Klinowski, and Jacek. Taboo search: An approach to the multiple minima problem. *Science*, 267(5198):664–666, 1995.

[59] Marco E and Gago F. Overcoming the inadequacies or limitations of experimental structures as drug targets by using computational modeling tools and molecular dynamics simulations. *Chem. Med. Chem.*

[60] Sippl M J. Knowledge-based potentials for proteins. *Curr. Op. Str. Biol.*, 5:229–235, 1995.

[61] Hoffmann D, Kramer B, Washio T, Steinmetzer T, Rarey M, and Lengauer T. Two-stage method for protein-ligand docking. *J. Med. Chem.*, 42:44224433, 1999.

[62] Wang J, Kollman P A, and Kuntz I D. Flexible ligand docking: A multistep strategy approach. *Proteins*, 36:119, 1999.

[63] Alder B J and Wainwright T E. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27(5):1208–1209, 1957.

[64] Alder B J and Wainwright T E. Studies in molecular dynamics. i. general method. *J. Chem. Phys.*, 31(2):459–466, 1959.

[65] Stillinger F H and Rahman A. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60(4):1545–1557, 1974.

[66] McCammon J A, Gelin B R, and Karplus M. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[67] Mark E Tuckerman and Glenn J Martyna. Understanding modern molecular dynamics: Techniques and applications. *J. Phys. Chem. B*, 104:159–178, 2000.

[68] Daan Frenkel and Berend Smit. *Understanding molecular symulation.* Academic Press, London, UK, 2002.

[69] Verlet L. Computer experiments on classical fluids : I. thermodynamics properties of lennard-jones molecules. *Phys. Rev.*, 159, 1967.

[70] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, 1982.

[71] T Lazaridis and M Karplus. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.*, 10:139–145, 2000.

[72] Feig M and Brooks C L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, 14(2):217–224, 2004.

[73] Fogolari F, Brigo A, and Molinari H. The poisson-boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Molecular Recognition*, 15(6):377–392, 2002.

[74] D Qiu, PD Shenkin, FP Hollinger, and WC Still. The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *J. Phys. Chem. A*, 101:3005–3014, 1997.

Chapter 2

A neural network approach for protein models validation

The world is the totality of facts, not of things.
Ludwig Wittgenstein (1889 - 1951)

2.1 Introduction

THe very large and continuously increasing amount of data obtained by genome sequencing makes the development of reliable computational methods capable to infer protein structures from sequences a crucial step for functional annotation of proteins. In fact, functional annotation is often strictly dependent on the availability of structural data, which in turn are still difficult to obtain experimentally. As a consequence, efforts and progresses in high throughput X-ray and NMR methods need to be accompanied by computational techniques suitable for three-dimensional structure predictions, such as homology modelling, fold recognition or ab-initio methods[52, 34, 49, 18, 32, 57, 56], which are intrinsically characterized by different levels of accuracy.

In parallel to the development and improvement of prediction methods, reliable and accurate evaluation tools are necessary to check the quality of computational protein models [31, 19]. Moreover, in the context of protein structure

refinement, which has been recently identified as one of the bottlenecks limiting the quality and usefulness of protein structure prediction [52], it has been noted that improvements in the selection of the most native-like model from an ensemble of closely related alternative conformations can be crucial. The increasing importance of the field of quality assessment methods is demonstrated by the introduction of a dedicated section in the latest CASP edition (CASP7) [51].

To evaluate protein structures, several different scoring functions have been developed, which can be classified into different categories depending on the principles and on the structural features considered in the evaluation. Physical scoring (energy) functions aim to describe the physics of the interaction between atoms in a protein and are generally parameterized on molecular systems smaller than proteins [22]. Knowledge-based scoring functions are designed by evaluating the differences between some selected features of a random protein model and the characteristics of a real protein structure [46, 48, 28, 50, 29]. Learning-based functions can be developed by training algorithms to discriminate between correct and incorrect models [55]. Independently by the category, scoring functions are generally tested by examining their capability to detect the native structure among a set of decoys [41], which can be generated in several different ways [35, 45, 44].

It is important to note that the performance of learning-based functions are generally strongly dependent on the specific aim for which they were developed, and consequently on the training set used. As an example, ProQ, a neural network based method developed to predict the quality of protein models [55], was specifically designed to discriminate between correct and wrong models, i.e. to recognize folds that are not compatible with a protein sequence. In fact, ProQ was recently combined successfully with the Pcons [25] fold recognition predictor and ranked as one the best methods in a recent survey of quality assessment methods [51]. Other reliable and extensively used computational methods used to validate the quality of protein structures are PROSA [47], ERRAT [7], Verify3D [53, 26], PROCHECK [20], what-if [11], PROVE [36] and victor/FRST [50].

In the present contribution, we present a computational method (Artificial Intelligence Decoys Evaluator: AIDE) that is able to reliably and consistently discriminate between correct and incorrect protein models. In particular, the quality of the protein structure is evaluated with neural networks using as input 15 structural parameters, which include solvent accessible surface, hydrophobic contacts and secondary structure content. In the first section of the paper,

the neural network structure and the training procedure are presented and discussed. In the second section, the performance of the neural network is evaluated, compared to available methods, and critically discussed.

2.2 Methods

2.2.1 Protein datasets

The 4state-reduced set is an all-atom version of the models generated by Park & Levitt [35] using a four-state off-lattice model. The fisa and fisa-casp3 sets contain decoys of four small alpha-helix proteins. In these sets main chains were generated using a procedure of fragment insertion based on simulated annealing: native-like structures were assembled from a combination of fragments of known unrelated protein structures characterized by similar local sequences, using Bayesian scoring functions [45]. The side chains of fisa and fisa-casp3 were modeled with the software package SCWRL [30]. The hg-structal is a set of hemoglobin models generated by homology modelling. The lmds subset [35],[10] was produced by Keasar and Levitt by geometry optimizations carried out using a complex potential that contains a pairwise component, as well as cooperative hydrogen bonds terms. The Rosetta all-atom decoys were generate with the ROSETTA method developed by David Baker [44]. The molecular dynamics set of decoys was generated by molecular dynamics (MD) simulations carried out in vacuum with the software GROMACS 3.2 [24, 6]. Each protein structure was submitted to 100 ps of simulation using the OPLS force field [15]. MD simulations were performed in the NVT ensemble at 600 K, using an external bath with a coupling constant of 0.1 ps [5]. The LINCS algorithm [12] was adopted to constrain bond lengths of heavy atoms, allowing us to use a 2 fs time step. Van der Waals and Coulomb interactions were truncated at 8 Å, while long-range electrostatics interactions were evaluated using the particle mesh Ewald summation scheme [54]. The Van der Waals radii were increased to 4 Å for all atoms, in order to speed up the unfolding process [21]. Snapshots from the trajectory have been extracted every 0.4 ps, collecting 250 misfolded structures for each protein, with a backbone RMSD (root mean square deviation between the initial structure and each snapshot) ranging from 0 to about 10 Å. In addition to these decoy datasets, the CASP5 [2], CASP7 [3] and LiveBench2 [27] sets were also included. The complete dataset contains 62819 protein models build on 193 proteins.

2.2.2 Training-set and test-set

The dataset was splitted into two disjoint sets : a training-set and a test-set. The training-set includes only proteins belonging to the LiveBench2 and CASP7 decoys sets (13693 model structures built on 96 different proteins). The test-

set includes the lmds, CASP5, hg_structal, MD, Rosetta and 4state-reduced datasets (49126 models build on 97 proteins).

2.2.3 Parameters-Descriptors used in the neural network

The relative solvent accessible surface ($rSAS$) was computed as it follows :

$$rSAS_{hydrophobic} = \frac{SAS_{hydrophobic}}{SAS_{total}}$$

$$rSAS_{hydrophilic} = \frac{SAS_{hydrophilic}}{SAS_{total}}$$

where the residues A, L, V, I, P, F, M, W were considered as hydrophobic and the SAS_{total} is the total solvent accessible surface computed using NACCESS [13].

The secondary structure was evaluated with the DSSP program [16], in which the typical 8-state DSSP definition was simplified according to the following rules : H and G to helix, E and B to strand and all other states considered as coil, in agreement with PSIPRED definition [14].

The fraction of secondary structure (SS) is defined as :

$$SS = \frac{n_{ss}}{N} * 100,$$

where n_{ss} is the number of residues located in well-defined secondary structure elements, and N is the number of protein residues.

The secondary structure for each decoy was also compared with the corresponding secondary structure predicted by PSIPRED. Accordingly, the relative consensus secondary structure (SSc) was defined as the ratio:

$$SSc = \frac{n_c}{N} * 100,$$

where n_c is the number of residues located in corresponding secondary structure elements according to DSSP definition and PSIPRED secondary structure prediction.

Generally, native structures are characterized by hydrophobic residues clustered in buried regions. Therefore, the number of contacts between hydrophobic residues was chosen as a possible relevant parameter to discriminate among correct and incorrect models. According to our definition, a contact is present if

the distance between two residues is greater than 2.5 Å and lower or equal to 5 Å [40]. Given n hydrophobic residues, the number of hydrophobic contacts (Q) is normalized relative to the number of all possible contacts:

$$rQ = \frac{2Q}{n(n-1)!},$$

Moreover, to keep into account the stereochemical quality of the model, some PROCHECK parameters were considered (table 2.1).

Parameter
Percentage of residue in Ramachandran plot core regions
Percentage of residue in Ramachandran plot allowed regions
Percentage of residue in Ramachandran plot generously allowed regions
Percentage of residue in Ramachandran plot disallowed regions
Number of bad contacts
G-factor for dihedral angles
G-factor for covalent bonds
Overall G-factor

Table 2.1: PROCHECK parameters used in AIDE. The G-factor, which is a log-odds score based on the observed distributions of stereochemical parameters, provides a measure of how "normal", or alternatively how "unusual", a given stereochemical property is.

2.2.4 Model accuracy measures

Quality of protein models was evaluated by means of five different descriptors, using the crystal structure as reference: RMSD on the backbone atoms, TM-score [58], GDT-TS [38], LG-score [38] and MaxSub [43].

RMSD was computed on the backbone atoms after superposing the model structure on the crystal structure, using the program CE [42].

TM-score was developed to evaluate the topology similarity of two protein structures [58]. TM-score values fall into the interval $[0, 1]$. Scores equal or below 0.17 indicate that the prediction has a reliability compared to a random selection from the PDB library.

GDT-TS gives an estimation of the largest number of residues that can be found in which all distances between the model and the reference structure are

shorter than the cutoff D. The number of residues is measured as a percentage of the length of the target structure. The values of GDT-TS fall into the interval $[0-1]$, with a GDT-TS of 1 corresponding to perfect superposition.

The LG-score represents the significance (P-value) of a score (S_str [23]) associated to the best subpart of a structural alignment between the model and the correct structure. The value is measured by using a structural P-value ranging from 0 to 1, with a value of 0 corresponding to optimal superposition.

MaxSub is calculated from the largest number of residues that superimpose well over the reference structure, and produces a normalized score that ranges between 0 and 1. A MaxSub value of 1 is associated to perfect superposition.

2.2.5 Neural network

Four layers feed-forward neural networks were used, with fifteen neurons in the input layer, two neurons in two hidden layers and one neuron in the output layer. A linear activation function was chosen for all neurons.

For each accuracy measure chosen to evaluate proteins quality (RMSD, TM-score, GDT-TS, LG-score and MaxSub) a different neural network was trained.

The inverse of the Pearson correlation coefficient (CC) between the true and the predicted data was used as performance function.

$$\{CC\}^{-1} = \left\{ \frac{\left(\boldsymbol{t} - \mu_t\right)^T \left(\boldsymbol{y} - \mu_y\right)}{(M-1)\sigma_t \sigma_y} \right\}^{-1} \tag{2.1}$$

where $\boldsymbol{t}$ is the vector of predicted values for each decoy, $\boldsymbol{y}$ is the vector of true values, μ_t, σ_t, μ_y, σ_y are the averages and the standard deviations of predicted and true values, respectively, and M is the number of decoys.

Optimization of neural networks was carried out using the attractive-repulsive particle swarm optimization algorithm (AR-PSO) [37], which is a modification of the original PSO method [17] [8]. PSO is a stochastic population-based optimization approach which explores the hyper-dimensional parameters space of a population of candidate solutions named particles. Particles fly over the solution space looking for the global optimum. Each particle retains an individual memory of the best position visited and a global memory of the best position visited by all the particles.
A particle calculates its next position combining information from its last movement, the individual memory, the global memory and a random component.

The PSO updating rule is described as follow :

$$\begin{cases} \boldsymbol{v}_{i,t+1} = \mu \boldsymbol{v}_{i,t} + c_1(\boldsymbol{w}_{i,t}^{best} - \boldsymbol{w}_{i,t}) + c_2(\boldsymbol{w}_t^{global} - \boldsymbol{w}_{i,t}) \\ \boldsymbol{w}_{i,t+1} = \boldsymbol{w}_{i,t} + \boldsymbol{v}_{i,t+1} \end{cases} \tag{2.2}$$

in which $\boldsymbol{w}_{i,t+1}$ represents the position vector of the particle i at time t (i.e. the neural network weights), $\boldsymbol{w}_{i,t}^{best}$ is the best position identified by the particle i so far (i.e. the neural network weights associated with the best performance value) and $\boldsymbol{w}_t^{global}$ is the best position identified among all the particles. The vector $\boldsymbol{v}$ represents the particles velocity, which is computed as the difference between two positions and assuming unitary time.

The term $(\boldsymbol{w}_{i,t}^{best} - \boldsymbol{w}_{i,t})$ represents the individual memory component and $(\boldsymbol{w}_t^{global} - \boldsymbol{w}_{i,t})$ is the global one. These two terms are rescaled by the random coefficients c_1 and c_2, respectively. The μ coefficient is used to rescale the velocity.

Starting particle positions and velocities were initialized at random. To reduce the problem of premature convergence to relative minima, the Attractive-Repulsive modification has been introduced [37]. This modification defines a measure of global diversity (D) among the particles as:

$$D = \frac{1}{S} \sum_{i=1}^{S} \sqrt{\sum_{j=1}^{N} (w_{ij} - \bar{w}_j)^2} \tag{2.3}$$

where S is the number of particles in the swarm, N is space dimension (the number of networks weights) and $\bar{w}_j$ is the average of the parameter j among the particles.
If D falls below a minimal threshold (t_{min}) the update rule is inverted as follow

$$\begin{cases} \boldsymbol{v}_{i,t+1} = \mu \boldsymbol{v}_{i,t} + (-1)c_1(\boldsymbol{w}_{i,t}^{best} - \boldsymbol{w}_{i,t}) + (-1)c_2(\boldsymbol{w}_t^{global} - \boldsymbol{w}_{i,t}) \\ \boldsymbol{w}_{i,t+1} = \boldsymbol{w}_{i,t} + \boldsymbol{v}_{i,t+1} \end{cases} \tag{2.4}$$

causing the particles to spread in the phase space. If D reaches a maximal threshold (t_{max}) the update rule is restored as in the standard PSO method. We choose $t_{min} = 0.1$ and $t_{max} = 5.0$.
The parameters $c1$, $c2$ and μ were set as in the original PSO method as $c1 = c2 \in [0.0, 2.0]$ and $\mu = 0.7298$. The maximum number of iterations was set to 10000. A population size of 5 particles was chosen. It should be noted that standard training algorithms such as gradient descent back-propagation, Levenberg-Marquardt and BFGS, led to poorer results when compared to the particle swarm optimization (data not shown).

2.2.6 Statistical analysis

The following statistical parameters were used: Pearson correlation coefficient, already described in the neural network section, fraction enrichment (F.E.) and Z_nat.

Fraction enrichment (F.E.) is defined as the fraction of the top 10% conformations featuring best structural resemblance to the native structure among the top 10% best scoring conformations.

Z_{nat} is the Z-score of the X-ray structure compared to the ensemble of decoys structures. It is computed using the equation 2.5 :

$$Z_{nat} = \frac{score_{native} - \mu_{decoys}}{\sigma_{decoys}} \tag{2.5}$$

Higher Z_{nat} values correspond to higher capacity to discriminate between the native structure and the corresponding decoys.

The Receiver Operating Characteristic (ROC) graph is a plot of all sensitivity/specificity pairs resulting from continuously varying the decision threshold over the range of results observed. The sensitivity or true positive fraction is reported on the y-axis, while the x-axis represents the 1-specificity or true negative fraction. A test with perfect discrimination (no overlap between the two distribution of results) has a plot curve that passes through the upper left corner, where both specificity and sensitivity are 1.00. The hypothetical plot of a test with no discrimination between the two groups is a 45° line going from the lower left to the upper right corner. Qualitatively, the closer the plot is to the upper left corner, the higher the overall accuracy of the test.

2.3 Results and Discussion

The evaluation of the quality of protein structures is generally carried out calculating a score which is a function of a set of parameter values computed for the protein model under study. In our computational procedure, the description of the relation between the parameters space and the scoring values is obtained using neural networks, because of their ability to describe complex non-linear relationships among data.

2.3.1 Selection of protein parameters related to structure quality

Among the possible parameters that can be computed for a protein structure, we have selected some properties that are expected to be related to structure quality: solvent accessible surface of hydrophobic and hydrophilic residues, secondary structure content, the fraction of secondary structure content of the model fitting with that predicted by PSIPRED [14], number of hydrophobic contacts, and selected PROCHECK parameters [20](see Methods for details). It should be noted that other possibly relevant parameters, such as the number of hydrogen bonds, have not been used due to intrinsic difficulties in the normalisation of their values.

2.3.2 Selection of the parameters used to evaluate structure similarity

A key issue for evaluating the quality of a predicted protein structure is the measure of its "distance" relative to the "real" structure, experimentally obtained by X-ray diffraction or NMR. Since AIDE has been developed to evaluate protein models that are often characterized by the correct fold but may differ for local details, the backbone root mean square deviation (RMSD) of the protein model relative to the X-ray structure can be considered a suitable measure of structure similarity [9]. In fact, it is well known that the proper evaluation of the quality of protein structures can be a non-trivial task, often depending on the methods used to generate protein models. Therefore, several other measures of protein structure similarity have been formulated, the most commonly used being: GDT-TS [38], LG-score [38],TM-score [58] and MaxSub [43], which have also been adopted in the present work.

2.3.3 Selection and optimization of the neural networks

A preliminary evaluation of the relative importance of each parameter in the description of structure quality was obtained using a linear model built with the M5-prime attribute selection algorithm [4], as implemented in Weka 3.4.2 [39]. A different linear model was computed for each accuracy measure. The training-set and and the test-set used for obtaining and evaluating the linear model are the same used for the neural network. The obtained linear models are listed below (parameters abbreviations are described in table 2.2):

- RMSD: $1.9305SS - 0.2326DH_G - 0.351COV_G + 0.9225OVER_G - 0.1815CORE - 0.1591ALL - 0.1899GENALL - 0.1215DISALL - 4.1309HB_{SAS} - 12.1232HY_{SAS} - 0.0362SSc - 0.0382rQ - 0.0022L + 30.4336$
  Excluded parameters : nBC
  Pearson correlation coefficient : 0.30

- GDT_TS: $-0.3586SS - 0.0006nBC - 0.0434DG_G - 0.036COV_G + 0.0463OVER_G - 0.0018ALL + 0.0041DISALL - 1.5662HB_{SAS} + 0.0025SSc - 0.0088rQ - 0.0001L + 1.2403$
  Excluded parameters : CORE, GENALL, HY_{SAS}
  Pearson correlation coefficient : 0.32

- TM-score: $-0.3937SS - 0.0006BC - 0.0369DG_G - 0.0168COV_G - 0.002ALL + 0.0026DISALL - 1.6329HB_{SAS} + 0.0027SSc - 0.0163rQ + 0.0003L + 1.2718$
  Excluded parameters : CORE, GENALL, HY_{SAS}, $OVER_G$
  Pearson correlation coefficient : 0.31

- MaxSub: $-0.3937SS - 0.0006BC - 0.0369DG_G - 0.0168COV_G - 0.002ALL + 0.0026DISALL - 1.6329HB_{SAS} + 0.0027SSc - 0.0163rQ + 0.0003L + 1.2718$
  Excluded parameters : CORE, GENALL, HY_{SAS}, $OVER_G$
  Pearson correlation coefficient : 0.31

- LG-score: $1.0516SS + 0.0473DG_G + 0.1107COV_G + -0.2139OVER_G + 0.0014ALL + -0.0161DISALL + -0.8566HB_{SAS} + -1.0837HY_{SAS} + -0.0015rQ + 0.0328SSc + -0.0008L + 1.2131$
  Excluded parameters : nBC, CORE, GENALL, GENALL, L
  Pearson correlation coefficient : 0.28

SS	Fraction of secondary structure
CORE	Percentage of residue in Ramachandran plot CORE
ALL	Percentage of residue in Ramachandran plot allowed regions
GENALL	Percentage of residue in Ramachandran plot generously allowed regions
DISALL	Percentage of residue in Ramachandran plot disallowed regions
nBC	Number of bad contacts
DH_G	G-factor for dihedral angles
COV_B	G-factor for covalent bonds
$OVER_G$	Overall G-factor
HB_{SAS}	hydrophobic relative solvent accessible surface
HY_{SAS}	hydrophilic relative solvent accessible surface
SSc	Secondary structure consensus
rQ	Relative number of hydrophobic contacts
L	Number of residues

Table 2.2: Linear models parameters abbreviation.

Analysis of the linear models revealed that the secondary structure content and the solvent accessible surface have the highest importance in all models. Moreover, results show that it is not possible to exclude any parameter since non negligible weights are associated to all selected parameters, when the accuracy measures chosen are considered as a whole. The neural networks forming the core of AIDE are four layers feed-forward neural networks with fifteen neurons (corresponding to the selected parameters) in the input layer, two hidden layers formed by two neurons each, and one neuron in the output layer. A linear activation function was chosen for all neurons. Indeed, different combinations of hidden layers (one or two) and different numbers of hidden neurons per layer (from two to ten nodes per layer) were tested. In addition, we tested also different activation functions of the neurons (sigmoid, log-sigmoid and linear functions). It turned out that, among the different combinations, the neural network featuring two hidden layers formed by two neurons gave the best results. In fact, an increase in the number of neurons led to poorer performances, probably due to the increased difficulties in the optimization procedure arising from the augmented network complexity. To carry out the optimization of neural networks, we have implemented the attractive-repulsive particle swarm optimization algorithm (AR-PSO) [37], as explained in Methods. Training of the neural networks using more conventional approaches (Back-propagation [a],Levenberg-Marquardt [b]), led to slightly lower performances (table 2.3). This may be due to the greater exploration ability that characterize the PSO methods.

	LM	BP	PSO
GDT_TS	0.36	0.38	0.45
LG-score	0.35	0.48	0.51
MaxSub	0.49	0.39	0.51
RMSD	0.38	0.33	0.42
TM-score	0.43	0.49	0.49

Table 2.3: For each assessment measure (GDT_TS, LG-score, MaxSub, RMSD and TM-score) the same neural network was trained using different algorithms : Back-propagation (BP), Levenberg-Marquardt (LM), and Particle Swarm Optimzation(PSO). The performance of the neural networks was computed as Pearson correlation coefficient on the overall test-set.

[a]Back-propagation parameters used for training : Training epochs : 100, Minimum performance gradient : 10^{-10}, Learning rate : 0.01

[b]Levenberg-Marquardt parameters used for training : Training epochs : 100, Minimum performance gradient : 10^{-10}, μ : 10^{-3}, μ increment : 10, μ decrement : 0.1

AIDE was trained and tested on datasets of all-atoms protein decoys for which the three-dimensional structures are available. Since it is known that methods used for building decoys may introduce some systematic bias, it is important to benchmark a scoring function on different decoy sets in order to assess its generality. The overall dataset used in the present study is composed by an ensemble of widely used all-atom datasets containing models of different proteins (4state-reduced, fisa, fisa-casp3, rosetta all-atoms, CASP5, CASP7, Livebench2, lmds, and hg-structal [35, 45, 44, 2, 3, 27]), plus a molecular dynamics set that was generated in our laboratory from X-ray structures (see Methods). After computation of the structural parameters to be inserted in the neural networks, the overall dataset was subdivided into a training and a test set, which were composed by 13693 and 49126 structures, respectively. The training-set includes only the proteins belonging to the LiveBench2 and CASP7 decoy sets (13693 model structures built on 96 different proteins). The test-set includes the lmds, CASP5, hg_structal, MD, Rosetta and 4state-reduced subsets (49126 models build on 97 proteins). The LiveBench2 and CASP7 decoy sets were chosen as training sets because they contain models build with different methods and of different protein size, ranging from 20 to 500 residues. No protein contained in the training set is present also in the test set. Then, a population of 50 neural networks was trained starting from different initializations of the structural parameters. The network featuring the best performance (the highest correlation coefficient on the training set) was selected as the working network in AIDE. A different neural network was trained for each measure of structure similarity chosen to evaluate proteins quality (RMSD, TM-score, GDT-TS, LG-score and MaxSub). Therefore, five different versions of AIDE were obtained from the training procedure, referred to in the following as AIDE RMSD, AIDE TM-score, AIDE GDT-TS, AIDE LG-score and AIDE MaxSub.

2.3.4 Assessment of AIDE performance

The performances of the different version of AIDE have been compared to results obtained from widely used methods developed to evaluate protein models quality.

The performances of the different methods were evaluated using a test-set which includes lmds, CASP5, hg_structal, MD, Rosetta and 4state-reduced subsets. The LiveBench2 and CASP7 sets were already used for training AIDE and therefore were not used in the comparative evaluation.

The Pearson correlation coefficient, Z_{nat} and fraction enrichment (F.E.),

which give indications about a method ability to assign good scores to good models, have been computed and results are collected in tables 2.4,2.5,2.6.

Analysis of Pearson correlation coefficients (table 2.4) shows that, according to this statistical indicator, the different AIDE versions behave quite similarly. Most importantly, average AIDE performances are similar or slightly better than those obtained by two state-of-the-art methods such as ProQ [55] and Victor [50]. It is also noteworthy that the performance of AIDE changes significantly moving through the different subsets forming the test-set. In particular, very high correlation coefficients are obtained with the MD and hg_structal datasets (correlation coefficient in the range 0.61-0.89 and 0.48-0.73, respectively), whereas low values of Pearson coefficients are associated to the CASP5 dataset (0.15-0.38). Relatively different values of Pearson correlation coefficients are obtained also with ProQ and Victor. In particular, and differently from AIDE, low correlation coefficients are obtained by ProQ for the Rosetta subset, and by Victor for the fisa subset (table 2.4).

	lmds	4state_reduced	CASP5	fisa	MD	hg_structal	ROSETTA	average
AIDE RMSD	0.39	0.42	0.15	0.63	0.61	0.69	0.27	0.45
AIDE TM-score	0.39	0.32	0.38	0.48	0.89	0.70	0.43	0.51
AIDE GDT-TS	0.45	0.34	0.28	0.58	0.77	0.73	0.44	0.51
AIDE LG-score	0.52	0.31	0.22	0.29	0.77	0.48	0.38	0.42
AIDE MaxSub	0.39	0.34	0.36	0.55	0.73	0.70	0.40	0.49
ProQ LG-score	0.20	0.62	0.48	0.18	0.81	0.80	0.06	0.45
ProQ MaxSub	0.15	0.48	0.39	0.14	0.77	0.76	0.05	0.39
Victor GDT-TS	-0.29	-0.53	-0.29	-0.05	-0.78	-0.75	-0.23	-0.41

Table 2.4: For each dataset belonging to the test-set the Pearson correlation coefficient between the predicted and the computed values is reported. The performance of AIDE is compared to that of ProQ and Victor/FRST validation softwares.

The factors responsible for such non-homogeneous performances of the methods, when applied to different datasets, could not be unrevealed and might require further dissection of the test-set. In light of these results and observations it can be concluded that, even if the overall performances of AIDE, ProQ and Victor are similar, these methods can behave very differently on protein models obtained using different approaches, suggesting that the combined use of AIDE, ProQ and Victor could be useful to properly evaluate the quality of a protein structure. Analysis of F. E. values (table 2.5) shows again quite similar overall performances of AIDE, ProQ and Victor. However, the average F. E. values obtained using ProQ are consistently higher (by 5-10%) relative to the corresponding values obtained with Victor and AIDE. A more detailed analysis of F. E. values obtained from the different subsets composing the test set highlights some interesting trends. F. E. values obtained from the lmds and

fisa subsets are consistently lower than the average. Moreover, AIDE and ProQ versions trained using different parameters to evaluate structure similarity can give quite different results.

	lmds	4state_reduced	CASP5	fisa	MD	hg_structal	ROSETTA	average
AIDE RMSD	15.20	42.58	37.10	25.00	48.19	43.67	20.80	33.22
AIDE TM-score	1.84	31.18	34.84	19.50	72.29	44.82	26.49	32.99
AIDE GDT-TS	2.07	32.68	29.86	11.50	72.28	50.57	25.29	32.03
AIDE LG-score	25.80	34.40	31.67	17.52	67.47	42.53	29.96	35.62
AIDE MaxSub	3.22	33.54	35.29	10.00	73.49	44.82	24.12	32.06
ProQ LG-score	18.30	54.78	39.59	12.50	72.45	74.71	13.30	40.80
ProQ MaxSub	1.95	52.84	45.24	12.00	65.86	67.84	43.69	41.34
Victor GDT-TS	14.40	42.57	28.50	4.0	63.85	54.02	11.61	31.27

Table 2.5: The 10%-fraction enrichment is shown for each dataset belonging to the test-set. The performance of AIDE is compared to that of ProQ and Victor/FRST validation softwares.

The latter observation is particularly evident for the lmds subset. It is also interesting to note that the best performances on the different subsets forming the test set are often obtained by different methods. As an example, the best F. E. values for the fisa subset are obtained using AIDE, whereas the best values for the hg_structal subset are obtained with ProQ, further suggesting that the combined use of the different methods can be a good strategy to obtain a more confident evaluation of the quality of a protein structure. Z_{nat} allows to evaluate how (and if) the different methods distinguish the native (X-ray) structure from the ensemble of its models (table 2.6). In this case it was possible to extend the comparison to other methods widely used to evaluate protein structures quality (Errat, Prosa II and Verify 3D). Only the lmds and 4state_reduced subsets have been used in this comparison because these are the only datasets in common among all the compared methods for which data are available. Analysis of Z_{nat} values reveals that ProQ and Victor have better performances in this statistical test, whereas AIDE results are generally comparable to those obtained with Errat, Prosa II and Verify 3D. Notably, very low Z_{nat} scores are obtained using AIDE RMDS and AIDE LG-score on the 4state_reduced subset.

It should be noted that Z_{nat} and F.E. do not give information about the ability of a method to assign low scores to bad models, i. e. these statistical indicators do not allow to check if a method is confusing different classes. To explore this issue we have qualitatively compared AIDE and ProQ performances, superposing the ROC plots (see Methods) computed on the test-set for each different performance function (figure 1). According to this analysis, ProQ MaxSub exhibits the greatest overall accuracy, whereas AIDE GDT-TS has the lowest

	lmds	4state_reduced
AIDE RMSD	2.4	0.5
AIDE TM-score	3.4	2.9
AIDE GDT-TS	3.5	3.1
AIDE LG-score	2.0	1.6
AIDE MaxSub	3.1	3.1
ProQ LG-score	3.9	4.4
ProQ MaxSub	1.8	3.5
Victor GDT-TS	3.5	4.4
Errat	3.1	2.5
Prosa II	2.5	2.7
Verify 3D	1.4	2.6

Table 2.6: Comparison of Z_{nat} values obtained using AIDE and other protein structure validation softwares. ProQ values have been obtained from Ref. [55].

accuracy. Considering the different AIDE versions, a clear distinction can be observed when comparing the overall accuracy of AIDE RMSD and AIDE MaxSub relative to AIDE LGscore, AIDE GDT-TS and AIDE TMscore (figure 1). Notably, a similar difference was not evident when considering the correlation coefficients or the fraction enrichment test. It is also important to note that AIDE LGscore behaves very similarly to ProQ LGscore until about 60% of sensitivity, whereas at higher sensitivity levels AIDE outperforms ProQ LGscore. These observations further corroborate the hypothesis that the combined use of ProQ and AIDE should give improved results in the evaluation of the quality of three-dimensional protein models.

2.3.5 The web interface of AIDE

The availability of five different AIDE versions gives a nice picture of the overall performance of the method. However, the overloading of output information can become a drawback for the user interested only in the most relevant results. In fact, the analysis of AIDE performance has shown that the five different versions of AIDE are generally characterised by similar behaviour (see table 2.4,2.5,2.6). To better evaluate the degree of correlation among different AIDE versions we have carried out a principal component analysis on the Pearson correlation matrix of the descriptors chosen to evaluate models quality. This analysis reveals a strong correlation between TM-score, GDT-TS and MaxSub. The different clustering of TM-score, GDT-TS and MaxSub relative to RMSD and LG-score is mainly due to the inverse relationship between the two families (figure 2.1 table 2.7).

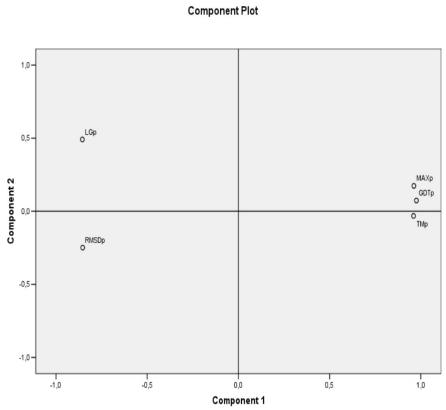


Figure 2.1: Loading plot of the accuracy parameters correlation matrix obtained by principal component analysis. Given the accuracy parameters matrix, where each row represents a different model and the columns are the descriptors used as quality measure (GDT_TS, LG-score, MaxSub, RMSD and TM-score), the correlation matrix (see table 2.7) was computed and analyzed by principal component analysis. Only the first two principal components are plotted.

Therefore, two (GDT-TS and the MaxSub) of these highly correlated parameters have been excluded from the output of the AIDE program available on the Internet [1] (figure 2.2). Moreover, to help the user in the evaluation of AIDE results, we have defined a threshold for each predicted parameter, in or-

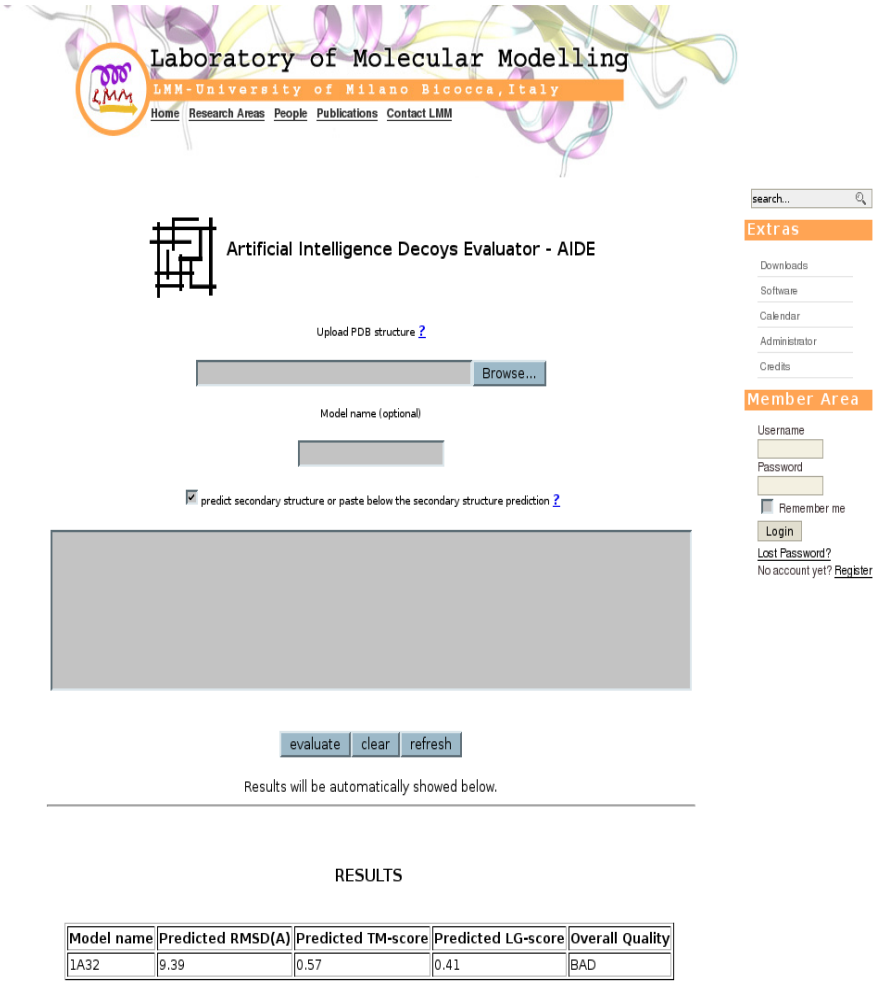


Figure 2.2: AIDE web interface: "http://linux.btbs.unimib.it/cgi-bin/aide.cgi".

der to discriminate between incorrect and correct models. In particular, correct models should have TM-score ≥ 0.31, RMSD ≤ 4.96Å and LG-score ≤ 0.35. These thresholds were chosen using a dataset of manually assessed models composed by some CASP5 targets belonging to the new fold and fold recognition categories. According to the visual evaluation of Aloy and coworkers [33], the models were divided into three class: class 2 ("excellent") when the overall fold is correct, class 1 ("good") when the model is considered partway to the correct fold, and class 0 for all the other models. For each model, the TM-score, LG-score and RMSD were computed (figure 2.3(a),2.3(b),2.3(a)) and the average value for the models belonging to the "excellent" class was used as threshold.

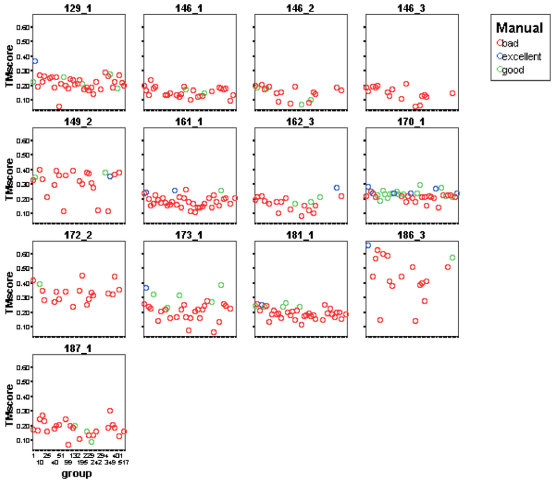
To further evaluate the classification ability using the chosen thresholds, the sensitivity and the specificity based on the ROC plots were also computed, figure 2.5.

2.4 Conclusions

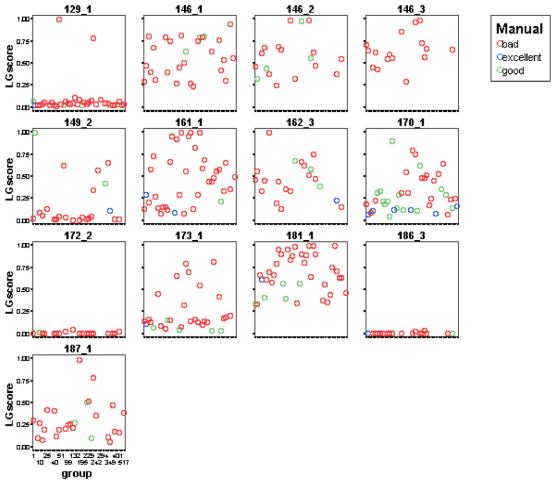
In this paper we have presented AIDE, a neural network system which is able to evaluate the quality of protein structures obtained by prediction methods. AIDE differs from other evaluation methods mainly for : *i.* a different choice of the parameters used to describe the protein structure, *ii.* a different choice of the parameters related to structure quality, *iii.* a novel strategy used to optimize the neural networks. AIDE overall performances are comparable to recently published state of the art methods, such as ProQ [55] and Victor [50]. However, detailed comparative analysis of results obtained using AIDE, ProQ and Victor reveals that the three methods have different and often complementary ability to properly assess the quality of protein structures. This observation suggests

	GDT_TS	LG-score	MaxSub	RMSD	TM-score
GDT_TS	1.00	-0.77	-0.98	-0.76	0.95
LG-score	-0.77	1.00	0.71	0.68	-0.79
MaxSub	-0.98	0.71	1.00	0.78	-0.93
RMSD	-0.76	0.68	0.78	1.00	-0.73
TM-score	0.95	-0.79	-0.93	-0.73	1.00

Table 2.7: Pearson correlation matrix of predicted accuracy parameters for the test-set.



(a)



(b)

Figure 2.3: Manual assessment of different models of 13 targets of CASP5 belonging to the category of "new fold" and "fold recognition". Each model has been classified into one of the following three classes : "excellent", "good" and "bad", and showed in the figure as blue, green and red circles, respectively [33]. Each target is represented into a subpanel different panel, where the horizontal axes indicates the model number and the vertical axes is the TM-score (a), LG-score (b).

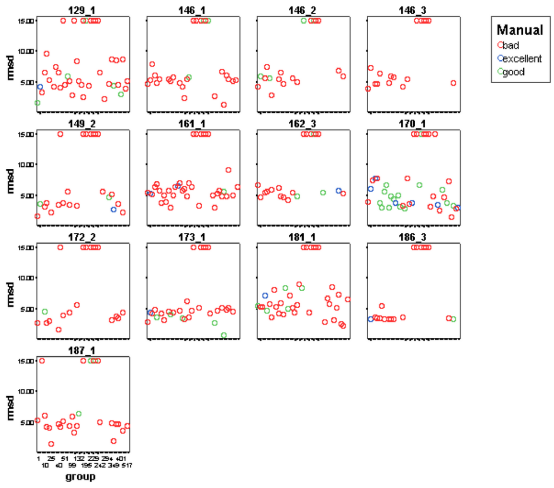


Figure 2.4: Manual assessment of different models of 13 targets of CASP5 belonging to the category of "new fold" and "fold recognition". Each model has been classified into one of the following three classes : "excellent", "good" and "bad", and showed in the figure as blue, green and red circles, respectively [33]. Each target is represented into a subpanel different panel, where the horizontal axes indicates the model number and the vertical axes is the RMSD.

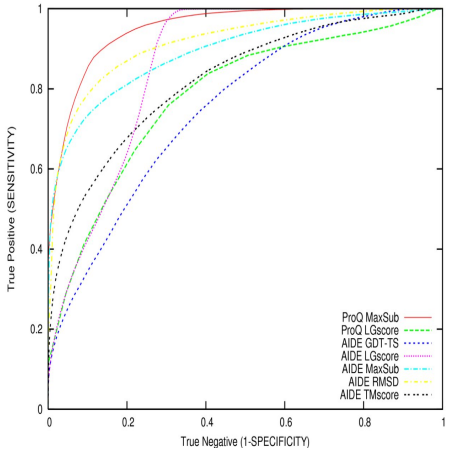


Figure 2.5: Sensitivity and specificity of AIDE TM-score, AIDE RMSD and AIDE LG-score, as obtained from the ROC curves at the chosen threshold.

that the combined use of AIDE, ProQ and Victor could increase the reliability in the evaluation of protein structures quality. AIDE is presently available on the Internet [1].

Bibliography

[1] Aide : Artificial intelligence decoys evaluator.

[2] Casp5, http://predictioncenter.genomecenter.ucdavis.edu/casp5/.

[3] Casp7, http://predictioncenter.org/casp7.

[4] *Induction of model trees for predicting continuous classes*, In Proc. Poster Papers Europ. Conf. Machine Learning, 1997.

[5] Berendsen H J C, Postma J P M, Dinola A, and Haak J R. Md with coupling to an external bath. *J. Phys. Chem.*, 81:3684–3690, 1984.

[6] Berendsen H J C, van der Spoel D, and van Drunen R. Gromacs: A message passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[7] Colovos C and Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci*, 2:1511–1519, 1993.

[8] Eberhart R C and Kennedy J. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micromachine and Human Science; Nagoya, Japan*, pages 39–43, 1995.

[9] Eramian D, Shen M-Y, Devos D, Melo F, Sali A, and Marti-Renom M A. A composite score for predicting errors in protein structure models. *Protein Sci*, 15(7):1653–1666, 2006.

[10] Boris Fain, Yu Xia, and Michael Levitt. Design of an optimal chebyshev-expanded discrimination function for globular proteins. *Protein Sci*, 11:2010–2021, 2002.

[11] Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph*, 8:52–56, 1990.

[12] B Hess, H Bekker, H J C Berendsen, and Fraaije J G E M. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

[13] S J Hubbard and J M Thornton. Naccess computer program. *Department of Biochemistry and Molecular Biology, University College London*, 1993.

[14] D T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[15] WL Jorgensen and J Tirado-Rives. The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J.A.C.S.*, 110:1657–1666, 1988.

[16] W Kabsch and C Sander. Dictionary of protein secondary-structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[17] J Kennedy and RC Eberhart. Particle swarm optimization. *Proc. IEEE Int'l. Conf. on Neural Networks, IV, 1942???1948. Piscataway, NJ 1995 : 1942-94.*

[18] Andrzej Kolinacuteski and Janusz M Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins: Structure, Function, and Bioinformatics*, 61 Suppl 7:84–90, 2005.

[19] A Kryshtafovych, C Venclovas, K Fidelis, and J Moult. Progress over the first decade of casp experiments. *Proteins: Structure, Function, and Bioinformatics*, 61 Suppl 7:225–267, 2005.

[20] R A Laskowski, M W MacArthur, D S Moss, and J M Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[21] T Lazaridis and M Karplus. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.*, 288:477–487, 1999.

[22] T Lazaridis and M Karplus. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.*, 10:139–145, 2000.

[23] M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structure comparison. *PNAS*, 95:5913–5920, 1998.

[24] E Lindahl, B Hess, and D van der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Biol.*, 7:306–317, 2001.

[25] J Lundstrom, L Rychlewski, J Bujnicki, and A Elofsson. Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci*, 10:2354–2362, 2001.

[26] R Luthy, JU Bowie, and D Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[27] Bujnicki J M, Elofsson A, Fischer D, and Rychlewski L. LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, Suppl 5:184–191, 2001.

[28] F Melo and Feytmans. Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, 267:207–222, 1997.

[29] F Melo, R Sanchez, and A Sali. Statistical potentials for fold assessment. *Protein Science*, 11:430–448, 2002.

[30] Bower MJ, Cohen F E, and Dunbrack R L. Prediction of protein side-chain rotamer from a backbone dependent rotamer library: a new homology modelling tool. *J. Mol. Biol.*, 267:1268–1282, 1997.

[31] J Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15:285–289, 2005.

[32] J Moult, K Fidelis, B Rost, T Hubbard, and A Tramontano. Critical assessment of methods of protein structure prediction (casp)–round 6. *Proteins: Structure, Function, and Bioinformatics*, 61 Suppl 7:3–7, 2005.

[33] Aloy P, Stark A, Hadley C, and Russell RB. Prediction wihout templates: new fold, secondary structure, and contacts in casp5. *Proteins*, 53 Suppl 6:436–456, 2003.

[34] Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim D E, Meiler J, Misura K M, and Baker D.

[35] B Park and M Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.

[36] J Pontius, J Richelle, and SJ Wodak. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.*, 264:121–136, 1996.

[37] J Riget and S Vesterstrom. A diversity-guided particle swarm optimizer - the arpso. 2002.

[38] Cristobal S, Zemla A, Fischer D, Rychlewski L, and Elofsson A. A study of quality measures for protein threading models. *BMC Bioinformatics*, 2:5, 2001.

[39] Garner S. *WEKA: The Waikato Environment for Knowledge Analysis.* University of Waikato, University of Waikato, Hamilton, New Zealand, 1995.

[40] William J Salerno, Samuel M Seaver, Brian R Armstrong, and Ishwar Radhakrishnan. Monster: inferring non-covalent interactions in macromolecular structures from atomic coordinate data. *Nucleic Acids Res*, 32:566–568, 2004.

[41] R Samudrala and M Levitt. Decoys r us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9:1399–1401, 2000.

[42] IN Shindyalov and PE Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11:739–747, 1998.

[43] N Siew, A Elofsson, L Rychlewski, and D Fischer. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 2000.

[44] KT Simons, R Bonneau, I Ruczinski, and D Baker. Ab initio protein structure prediction of casp iii targets using rosetta proteins. *Proteins*, Suppl 3:171–176, 1999.

[45] KT Simons, C Kooperberg, ES Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268:209–225, 1997.

[46] MJ Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[47] MJ Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[48] MJ Sippl. Knowledge based potential for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[49] J Soonming, K Eunae, S Seokmin, and Youngshang P. Ab inition folding of helix bundle proteins using molecular dynamics simulations. *J.A.C.S.*, 125:14841–14846, 2003.

[50] S Tosatto. The victor/frst function for model quality estimation. *J. Comp. Biol.*, 12:1316–1327, 2005.

[51] Anna Tramontano. An account of the seventh meeting of the worldwide critical assessment of techniques for protein structure prediction. *FEBS Journal*, Vol. 274 Iss. 7:1651–1654, 2007.

[52] M Tress, I Ezkurdia, O Grana, G Lopez, and Valencia A. Assessment of predictions submitted for the casp6 comparative modeling category. *Proteins: Structure, Function, and Bioinformatics*, 61 Suppl 7:27–45, 2005.

[53] Bowie J U, Luthy L, and Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[54] Essman U, Perela L, Berkowitz M L, Darden T, LeeH, and Pederson L G. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577–8592, 1995.

[55] Bjorn Wallner and Arne Elofsson. Can correct protein models be identified? *Protein Science*, 12:1073–1086, 2003.

[56] Jinbo Xu. Fold recognition by predicted alignment accuracy. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 2(2):157–165, 2005.

[57] Jinbo Xu, Libo Yu, and Ming Li. Consensus fold recognition by predicted model quality. In *APBC*, pages 73–83, 2005.

[58] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, 2004.

Chapter 3

The sweet taste receptor and sweeteners

Please prepare for docking procedure.
Kerry Conran. (Film, 2004)

3.1 Introduction

The sense of taste gives us important information about the nature and quality of food, and of all the basic taste qualities, sweetness is the most universally liked. The human appetite for refined sugar and for sweet foods and drinks has been so strong that it has influenced the course of human history, and the recent and sharp rise in the consumption of sugar may be unprecedented. In the human taste bud, some cells express sweet receptors and respond to sweetness. Inside the taste receptor cell, two proteins combine to create a sweet receptor [24]. These proteins are the taste receptor family 1, proteins 2 and 3, named t1r2 and t1r3 [24].

3.1.1 The sweet taste receptor

The sweet taste receptor (SR) is a G protein coupled receptor (GPCR) similar to the dimeric mGluR1 (metabotropic glutamate receptors) receptor. Both belong to class C of GPCRs, which includes several metabotropic glutamate receptors, sweet and umami (monosodium glutamate) taste receptors, the Ca^{2+} sensing

receptor, the γ-aminobutyric acid type B receptor, and pheromone receptors. Class C receptors have an extracellular domain composed of a Venus fly trap domain (VFTD) containing the active site, a seven helices transmembrane domain (7TMD), and a cysteine-rich domain. The mGluR1 is homodimeric while the SR is an heterodimeric receptor composed by the taste receptors t1r2 and t1r3 [24]. While the ligands of mGluR1 are either glutamate or closely related molecules, the ligands able to activate the sweet taste receptor vary widely in chemical constitution, ranging from sugars to amino acids, peptides, proteins, and several other classes of organic compounds. The very fact that sweeteners cover a particularly wide range of chemical constitution hints that at least some of them may interact with parts of the SR different from the two likely cavities corresponding to the Glu hosting cavities of mGluR1, either in the N-terminal domain or in the transmembrane helices. For example the C-terminal transmembrane domain of t1r3 is required for recognizing cyclamate and the sweet taste inhibitor lactisole [10]. In figure 3.1 it is possible to see the mapped interactions between the t1rs receptors and some of its ligands.

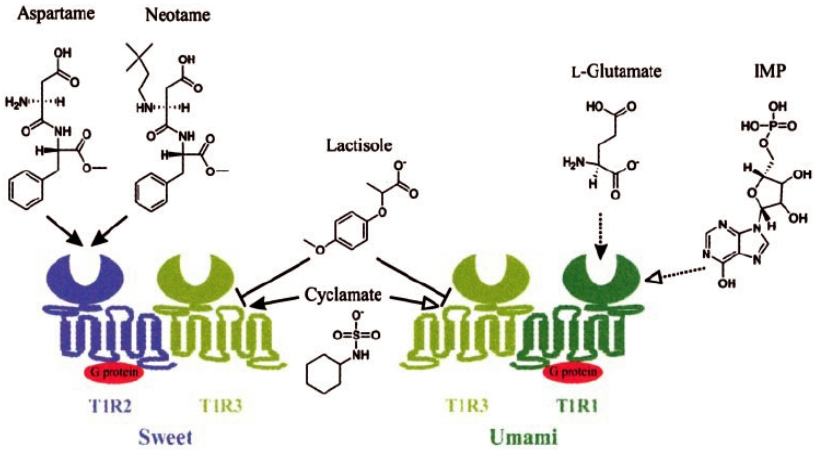


Figure 3.1: Model for the sweet and umami taste receptor structure function relationships. Filled arrows indicate direct activation, open arrows indicate enhancement, and bar heads indicate inhibition. Solid lines indicate proposed mechanisms based on experimental evidence; broken lines indicate mechanisms based on our speculations.

The extracellular domain of the sweet taste receptor should exists, as for the mGluR1, as a mixture of the free form I and the free form II[4, 9]. The two forms are characterized by a different orientation of the two monomers which are composed. As all the 3D models of the SR were made using the mGluR1 crystal structure we first briefly describe it. The mGluR1 was crystallized in three dimeric conformations: the free form I (PDB ID: 1ewt), the free form II (PDB ID: 1ewv) and the complexed form (PDB ID: 1ewk) with the glutamate bound on both monomers. The complexed form is nearly identical to the free form II but for the ligands. Each monomer is composed by two subdomains, named LB1 and LB2 (figure 3.2). The monomers exist in two different conformers (open and close) depending on the spatial orientation of the subdomains LB1 and LB2. The free form I contains two open conformers, whereas the free form II, as well as the complexed form, contains an open and a closed conformer (figure 3.2). The complexed form is also named active, because it is able to activate the transmission signal [13, 4].

The extracellular domain of the SR is able to bind various kind of small molecular weight sweeteners, as well as sweet macromolecules[24]. The complexed (active) form is stabilized either upon binding of the small molecular weight molecules or by the sweet macromolecules, and both activate long lasting signal transmission.
Depending on the localization the sweet taste receptor can have different functional roles. Indeed, beside to be expressed in the mouth where the receptor is involved into the sweet taste perception, it is also localized on the enteroendocrine cells of the gut. Here, the stimulation of the t1r2 t1r3 receptors activates intracellular signaling elements, including α-gustducin, and causes the release of GLP-1 and GIP hormones. These hormones stimulate the expression of SGLT1 in enterocytes, which, in turn, increase the absorption of glucose from the intestinal lumen [3, 14].

3.1.2 Sweeteners

A sugar substitute is a food additive that emulates the effect of sugar or corn syrup in taste, but usually has low caloric power. Some sugar substitutes are natural and some are synthetic or artificial sweeteners.
The most diffused artificial sweeteners are : acesulfame potassium, nutrinova, aspartame, neotame, saccharin, and sucralose. Since their discovery, the safety

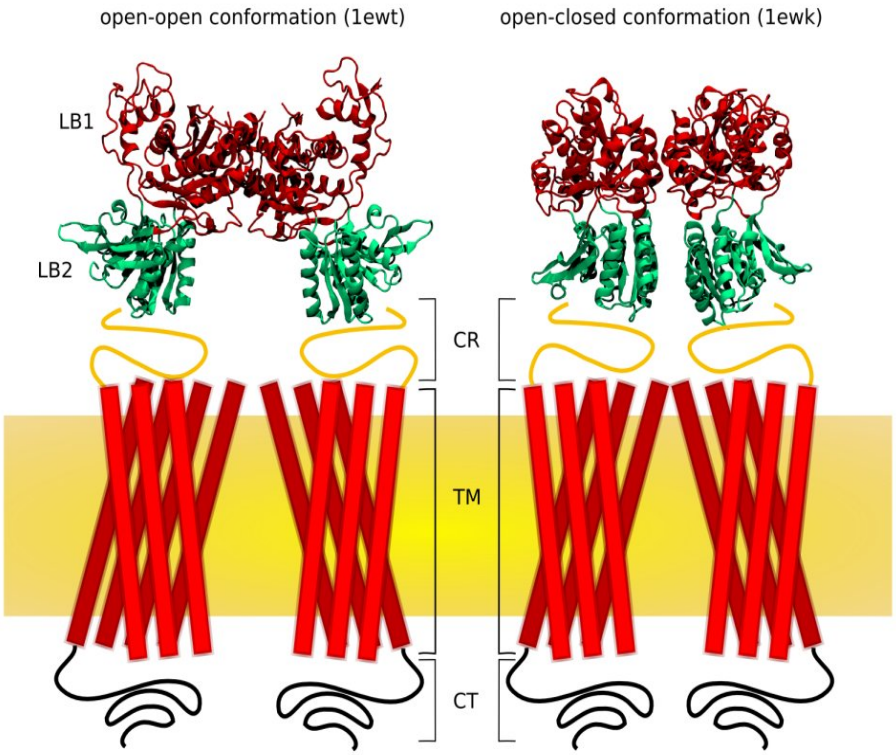


Figure 3.2: Crystal structure of the ligand-binding core of metabotropic glutamate receptor subtype 1. Metabotropic Glu receptors (mGluRs) are essential for the development and function of the mammalian central nervous system and exist as homodimers. The receptor is characterized by a large extracellular domain divided into a ligand-binding region (LB1-LB2) and a cysteine-rich region (CR) that links the ligand binding region to the transmembrane region (TM). The crystal structure of the LBRs of the mGluR1 homodimer, composed of the two subdomains LB1 and LB2, is shown in the non-bound form (left) and bound to its natural agonist, glutamate (Glu) (right). The LBRs of the homodimer can have either an open or a closed conformation. The ligand-free dimer shows either an open-open resting or a closed-open active conformation. Binding of the agonist stabilizes the closed-open conformation, which is characterized by a change in the relative orientation of the protomers. Thus, the LB2 subdomains of the LBR protomers come closer to each other at approximately 25Å, and this change may trigger the active state of the receptor.

of artificial sweeteners has been controversial. After the increased attention on the diffusion of the obesity in the United States, more people are choosing to use these products. These choices may be useful for those who cannot tolerate sugar in their diets (e.g., diabetics). However, scientists disagree about the relationship between sweeteners and many disease such as lymphomas, leukemias, cancers of the bladder and brain, chronic fatigue syndrome, Parkinson's disease, Alzheimer's disease, multiple sclerosis, autism, and systemic lupus [19]. Natural sweeteners can be small molecules such as fructose, glycerol, sorbitol, stevioside or small proteins such as brazzein, thaumatin, monellin, pentadin [4, 22].

Stevioside

Stevioside is a natural sweetener extracted from leaves of Stevia rebaudiana (Bertoni), that is a perennial shrub of the Asteraceae (Compositae) family native to certain regions of South America (Paraguay and Brazil). Stevioside, the main sweet component in the leaves of Stevia rebaudiana (Bertoni) tastes about 300 times sweeter than sucrose (0.4% solution). Stevioside is one of the sweet components that can be extracted from the leaves of Stevia rebaudiana; the other compounds present, but in lower concentration, are: steviolbioside, rebaudioside A, B, C, D, E, F and dulcoside A. Their content varies depending on the cultivar and growing conditions [5]. All the stevioside related compounds contained into the leaves are shown in table 3.1.

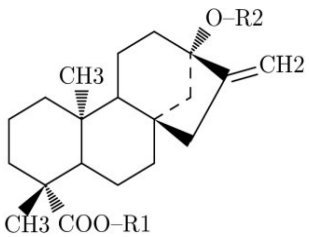


Figure 3.3: Backbone structure of the steviol glycosides compounds.

The Stevia plant, its extracts, and stevioside have been traded almost all over the world and have been used to sweeten hundreds of diabetic products, particularly soft drinks. In particular Japanese people are considered to be the greatest consumers of this sweetener. The important property to be non-caloric sweeteners makes of these molecules a particular interesting products for food industry.

Compound name	**R1**	**R2**
Steviol	H	H
Steviolbioside	H	β-Glc-β-Glc(2→1)
Stevioside	β-Glc	β-Glc-β-Glc(2→1)
Rebaudioside A	β-Glc	β-Glc-β-Glc(2→1) | β-Glc(3→1)
Rebaudioside B	H	β-Glc-β-Glc(2→1) | β-Glc(3→1)
Rebaudioside C	β-Glc	β-Glc-α-Rha(2→1) | β-Glc(3→1)
Rebaudioside D	β-Glc-β-Glc(2→1)	β-Glc-β-Glc(2→1) | β-Glc(3→1)
Rebaudioside E	β-Glc-β-Glc(2→1)	β-Glc-β-Glc(2→1)
Rebaudioside F	β-Glc-β-Glc(2→1)	β-Glc-β-Xyl(2→1) | β-Glc(3→1)
dulcoside A	β-Glc	β-Glc-α-Rha(2→1)

Table 3.1: List of the steviol glycosides contained in the Stevia leaves. The substituents attached to the steviol glycoside backbone shown in figure3.3 are listed.

This plant has been reported to be hyploglycemiant, hypo-tensor, diuretic and cardiotonic [7, 20]. In Brazil, it has been successfully used as the most suitable sweetener for diabetic people [5, 20]. Moreover the stevioside has been found to be an immunostimulator as evidenced by the increase in B- and T-cell mediated humoral and DTH response, respectively. It also has been shown to enhance macrophage function and substantially modulate the T and B cell proliferation [20].
In some countries, especially USA and Europe, the alimentary employment of these natural sweeteners is forbidden due to the pressure and the lobbying led by the powerful industry of artificial sweeteners [20].

The binding site of stevioside and related compounds is not known and only few studies dealing with the stevioside binding site investigation are reported [16, 15, 13]. From these studies, the location of the stevioside binding site can not be clearly evinced, we can only infer that it has different potential binding sites on the sweet taste receptor t1r2-t1r3. In particular the transmembrane or extracellular domains of t1r2 and of t1r3 are all possible candidates. Here we have made an *in-silico* study with the aim to identify the most probable binding site for the stevioside and related compounds. Moreover an atomic-level characterization of the bound complexes have been done.

3.2 Methods

3.2.1 Modelling of the sweet taste receptor

The modelling of the sweet taste receptor was performed considering the transmembrane and the extracellular domains separately. Indeed, because of the completely different nature of the two domains, the modelling follows a different procedure.

3.2.2 Homology modelling of extracellular domain

Following the procedure used by Temussi et al. ([4, 22, 9]) we have modelled all possible receptor models based on the human sequence, either the free form I and the complexed form (identical to the free form II). Moreover, given that there is no a preferred monomer for t1r2 or t1r3 we have built t1r2 and t1r3 on each monomer, obtaining the models listed in table 3.2.

Template	**Model**	**Notes**
1ewt, chain A	t1r2 on 1ewtA	Free form I (Resting), open monomer
1ewt, chain B	t1r2 on 1ewtB	Free form I (Resting), open monomer
1ewt, chain A	t1r3 on 1ewtA	Free form I (Resting), open monomer
1ewt, chain B	t1r3 on 1ewtB	Free form I (Resting), open monomer
1ewk, chain A	t1r2 on 1ewkA	Complexed form (Active), closed monomer
1ewk, chain B	t1r2 on 1ewkB	Complexed form (Active), open monomer
1ewk, chain A	t1r3 on 1ewkA	Complexed form (Active), closed monomer
1ewk, chain B	t1r3 on 1ewkB	Complexed form (Active), open monomer

Table 3.2: List of models for the sweet taste receptor obtained using the available crystal structures of the glutamate receptor mGluR1.

The sequence alignment was generated using ClustalW [8], producing alignments with a sequence identity lower than 30% on average. Hence the models were built using Modeller version 9v4 [2]. and were refined using the molecular dynamics refinement procedure included into Modeller. The model validation was performed using AIDE [17] and PROSA [21].

3.2.3 Homology modelling of transmembrane domain

Modelling of the transmembrane domain requires a more sophisticated procedure. Following the work of Peihua Jiang and coworkers [16] the models were obtained using the bovine rhodopsin as template (PDB ID: 1f88). The sequence alignment between the bovine rhodopsin and the t1r2 was done using clustalW [8] obtaining an alignment with a sequence identity of 24%. With this alignment an initial model was constructed using Modeller [2], thus the model was submitted to a refinement procedure: the model was minimized in vacuum by 2000 steps of conjugate gradient using NAMD [18]. Then it was inserted into a phospholipid bilayer composed by 1-Palmitoyl-2-oleoyl-sn-Glycero-3-phosphoethanolamine (POPE) [12], which was solvated by water molecules. This system was minimized by 2000 steps of conjugate gradient and then subjected to 200ps of molecular dynamics simulation at constant pressure (1 atm) and temperature (310K) using periodic boundary conditions. A smoothing function starting at 10Å and ending at 12Å was used for the non-bonded interactions. At the end of the molecular dynamics simulation the system was subjected to another 2000 steps of conjugate gradient minimization. All model refinements were carried out using NAMD with the CHARMM 27 force field.

Omologues sequences of the transmembrane domain of t1r2 were identified using BLAST. Then a multiple alignement was done using t-coffee [6].

3.2.4 Building the sweeteners

All the sweeteners were built using MOE version 2007a [11]. The molecules were minimized in the MMFF94s force field using the conjugate gradient algorithm until it reaches an rms of the gradient equal to 0.01.
Before submitting the ligands to the docking procedure, each one were treated with another energy minimization using Delos. Delos integrates a simulated annealing optimization algorithm that it more suited to identify the global energy minimum of complex molecules compared to the conjugate gradient.

3.2.5 *In-silico* docking

All docking procedure was performed using Delos software [1]. For the extracellular domain the binding pocket was identified considering the position of the glutamate bound to the complexed form of the mGluR1. All the models were superposed to the corresponding chain of the crystal structure of mGluR1 and the binding area was defined as a box of 30Å side centered on the center of mass

of the bound glutamate.
Hence, each ligand was put into the binding box and submitted to a semiflexible docking search using the simulated annealing algorithm starting from a temperature of 3000K linearly decreasing till 3K using a temperature step of 0.5K. Each simulated annealing run was repeated 100 times. The position corresponding to the lowest binding energy was saved.

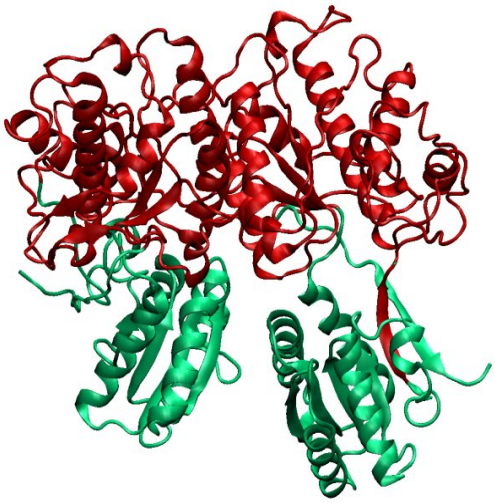
3.3 Results and Discussion

3.3.1 Modelling of the sweet taste receptor

The modelling of the sweet taste receptor was performed considering separately the transmembrane and the extracellular domains. Indeed, because of the completely different nature of the two domains, the modelling follows a different procedure which is described in Methods, section 3.2.1.
The final models obtained for the extracellular domain are depicted in figure 3.4 and 3.5.
All the extracellular models were validated using two different methods of models quality assessment; the first is AIDE [17], which is a learning-based approach estimating different measures of models correctness (see chapter 2 for a detailed explanation of the method). The AIDE quality assessment results were shown in table 3.3. The second method used is PROSA-web [21, 23], which is based on a knowledge-based potential and it gives a statistical value of models quality.
Using either AIDE or PROSA we have obtained good results for all the models, especially for the t1r3 ones (table 3.3,3.4).
The transmembrane domain was modelled and refined as described in section 3.2.3. The refined final system dipped into a solvated lipid bilayer is shown in figure 3.6(a), for clarity in figure 3.6(b) only the refined model without the membrane and water is shown.

3.3.2 *In silico* docking of stevioside and others sweeteners

The binding free energies [kJ/mol] of all the compounds are summarized in table 3.5, the missing values indicate that the binding free energy are positive (indicated as >0 in the table), or that the docking procedure is not terminated because of too many clashes encountered, i.e. the molecule is too bulky for the pocket.
The transmembrane domain of the receptor is able to bind mostly of the tested compounds. Moreover looking at the free energy values we see that large compounds, such as the rebaudioside B, E, F, dulcoside A, and steviolbioside, fit better than smaller sweeteners (steviol,aspartame,sucrose, and saccharin).



(a) t1r2 on 1ewk A - t1r3 on 1ewk B

(b) t1r3 on 1ewk A - t1r2 on 1ewk B

Figure 3.4: Models of the dimeric N-terminal domain of the sweet taste receptor t1r2-t1r3 built on the open-closed form of the metabotropic glutamate receptor mGluR1 (PDB ID 1ewk). Both the possible models are shown: a. t1r2 on 1ewk A - t1r3 on 1ewk B and b. t1r3 on 1ewk A - t1r2 on 1ewk B.

(a) t1r2 on 1ewt A - t1r3 on 1ewt B

(b) t1r3 on 1ewt A - t1r2 on 1ewt B

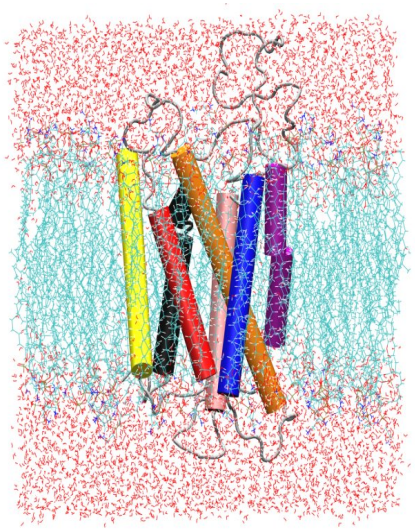
Figure 3.5: Models of the dimeric N-terminal domain of the sweet taste receptor t1r2-t1r3 built on the open-open form of the metabotropic glutamate receptor mGluR1 (PDB ID 1ewk). Both the possible models are shown: a. t1r2 on 1ewt A - t1r3 on 1ewt B and b. t1r3 on 1ewt A - t1r2 on 1ewt B.

Model	AIDE validation			
	RMSD(Å)	TM-score	LG-score	Overall Quality
t1r2 on 1ewkA	5.72	0.68	0.22	GOOD
t1r2 on 1ewkB	6.23	0.65	0.23	GOOD
t1r2 on 1ewtA	6.45	0.64	0.21	GOOD
t1r2 on 1ewtB	6.07	0.67	0.21	GOOD
t1r3 on 1ewkA	3.57	0.67	0.18	EXCELLENT
t1r3 on 1ewkB	3.76	0.67	0.18	EXCELLENT
t1r3 on 1ewtA	3.70	0.70	0.17	EXCELLENT
t1r3 on 1ewtB	2.98	0.69	0.17	EXCELLENT

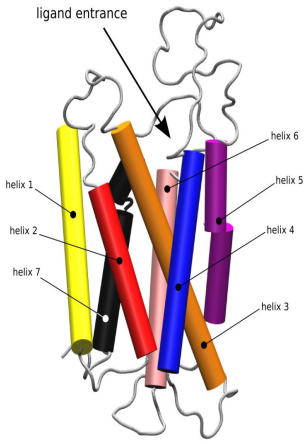
Table 3.3: AIDE extracellular domain models validation. Aide gives three measures of accuracy and an averall indicator computed combining the three measures. The best models structure should have RMSD=0.0, TM-score=1.0 and LG-score=0.0. The RMSD has only the lower bound which is 0.0 but the upper bound is not fixed. Both the TM-score and the LG-score range from 0.0 to 1.0.

Model	Z-score PROSA
t1r2 on 1ewkA	-6.37
t1r2 on 1ewkB	-6.70
t1r2 on 1ewtA	-6.59
t1r2 on 1ewtB	-6.27
t1r3 on 1ewkA	-4.79
t1r3 on 1ewkB	-5.36
t1r3 on 1ewtA	-6.79
t1r3 on 1ewtB	-6.98

Table 3.4: PROSA extracellular domain models validation. The web interface of prosa was used to evaluate each model. The z-score indicates overall model quality. Comparing the z-score of the model with the z-score of experimentally determined protein chains of similar size in current PDB (precomputed) it is possible to assess the accuracy of a given model. In this case the z-score of experimental determined structures of similar size (300 aminoacids) ranges from -11 to -4; outside this range the protein model is more probable to be incorrect.



(a) Transmembrane model of t1r2 dipped into a solvated POPE bilayer



(b) Refined transmembrane model of t1r2

Figure 3.6: a. Refined model of the transmembrane portion of t1r2 dipped into a solvated lipidic bilayer. b. Refined model of the trasnsmembrane portion of t1r2, the entrance of the ligand is indicated with an arrow.

Compound	TM	N-terminal on 1EWT		N-terminal on 1EWK		Exp
		t1r2A/t1r3B	t1r3A/t1r2B	t1r2A/t1r3B	t1r3A/t1r2B	
steviol	-8.72	x/x	-11.90/x	x/-11.77	x/x	x
stevioside	>0	x/-14.26	x/x	x/x	x/x	x
dulcosideA	-11.63	x/x	x/x	x/-12.4142	x/x	x
rebaudiosideA	>0	x/x	x/x	x/x	x/x	x
rebaudiosideB	-12.96	x/-15.78	x/x	x/x	x/x	x
rebaudiosideC	>0	x/x	x/x	x/x	x/x	x
rebaudiosideD	>0	x/x	x/x	x/x	x/x	x
rebaudiosideE	-13.10	x/x	x/x	x/x	x/x	x
rebaudiosideF	-13.98	x/x	x/x	x/x	x/x	x
steviolbioside	-14.02	x/x	x/x	x/-11.68	x/x	x
aspartame	-10.85	x/-15.31	-15.22/x	-20.06/-13.91	x/x	-9.70
sucrose	-8.20	x/-11.64	-11.21/x	x/-9.43	x/x	-6.71
saccharin	-5.69	-5.81/-11.65	-12.21/-11.65	-15.88/-9.60	-8.44/-6.74	-9.66
neotame	x/x	x/x	x/x	x/x	x/x	-12.12

Table 3.5: Binding free energies [kJ/mol] of stevioside and others common sweeteners to different portions of the Human models of the sweet taste receptor t1r2-t1r3. The x indicate that too many clashes were encountered during the docking procedure, i.e. the ligand doesn't fits into the binding pocket. Binding free energies values greater than 0 are indicated as > 0. The last column show the experimentally determined values of binding free energies obtained from ref. 9.

3.3.3 The extracellular domain binding pocket

Looking at the binding free energies (table 3.5), we could see that the t1r3 is able to bind most of the compounds. In particular, when t1r3 is build onto the chain B of either 1ewk or 1ewt it is able to bind bulky compounds such as stevioside, rebaudioside B, steviolbioside or dulcoside A, whose volume is reasonably larger compared to the small sweeteners such as aspartame, saccharin and sucrose (see table 3.6 to compare some stuctural properties of the studied compounds). The largest compounds (rebaudioside A, B, C, D, E, and F) are all unable to bind any model but the rebaudioside B that binds the t1r3 chain B (built on the inactive open-open conformation).
The residues and the interactions involved into the ligand binding are shown in figure 3.8; only the compounds with the lowest free energy values were analyzed. The aspartame and saccharin are bound very strongly to the closed form of t1r2 (t1r2 built on the closed monomer close-open form of mGluR1), which is completely unable to bind all the other ligands. The closed form of the monomer can bind only small ligands as the binding pocket is smaller compared to the open form, this is clearly visible in figure 3.7 where the binding cavity of the open and the closed monomer are shown. This is confirmed looking at binding free energies values of table 3.5 where it is possible to see that the closed forms (t1r2 on 1ewk A and t1r3 on 1ewk A) are mostly unable to bind any but the smallest compounds.
Some aminoacids are involved in the interaction with many different ligands. In particular the E230 and Y147, are shared by all the complexes but the saccharin, whereas the L233 is common to all the ligands except for the aspartame and the saccharin. In the open monomer these aminoacid are located at the entrance of the binding cavity, and may be participate in the substrate recognition. In the closed form the LB1 subdomain closes over the LB2, the residues that are located at the entrance of the binding pocket get buried inside the protein. However, in the case of aspartame, Y147 (146 in t1r2) and E230 (233 in t1r2) still holds a role in the ligand interaction (see figure 3.8d).
The active form of the receptor is the open-closed form, it is not clear if the receptor shifts from the inactive open-open form to the active open-closed form upon binding to the ligand, or the ligand only stabilizes the active open-closed conformation. Moreover it's not yet clear if the ligand should bind to both monomers in order to activate the signal transmission.
The only compounds that bind to the closed form are aspartame and saccharin, which size is not far from the glutamate, whereas all the other are able to bind the open monomer only.

This entails different possible scenarios for the activation of the signal transmission. In case of small compounds, such as aspartame or saccharin, we have two possible instances: the first is that small compounds stabilize the open-closed form binding either one or both monomers. The second possibility is that they bound to the inactive open-open conformation inducing a conformational change to the active form.
Taking into account large compounds we have again two possible scenarios: the first case is that the compounds bind to the chain B of the inactive open-open form leading to a conformational switch of the active open-close form. They can only bind the chain B of the dimer as this chain holds an open conformation in the active open-closed form. The second case implies that the large ligands directly bind to the chain B (open form) of the active open-closed form stabilizing it.
It is worth to note that we are always dealing with a semiflexible docking procedure, that avoids the receptor structure to adapt its conformation during the docking process. This problem becomes particularly relevant for large compounds because they require a wide pocket to fit them in. As mentioned in the introduction, we need some ways to treat the flexibility of the receptor. As we have 8 different receptors, and many different ligands, the molecular dynamics simulations is not feasible due to the excessive computational costs. Hence we planned to apply a normal mode analysis that allows the identification of the most important protein motions in a computational cost-effective way.

Compound	diameter	dipole	rgyr	ASA+	ASA_H	vol
Aspartame	11	3.30	3.36	223.34	309.34	280.00
Dulcoside A	17	1.05	5.23	145.17	508.93	579.62
Neotame	13	0.50	3.82	181.96	426.00	382.12
RebaudiosideA	23	1.66	7.23	338.76	597.68	849.62
RebaudiosideB	18	0.94	5.64	223.49	537.02	714.12
RebaudiosideC	18	1.88	5.48	208.19	556.89	706.75
RebaudiosideD	26	1.76	8.25	441.80	613.56	981.50
RebaudiosideE	25	2.28	7.76	387.94	590.31	848.25
RebaudiosideF	18	1.30	5.66	219.83	537.96	691.12
Steviol	9	0.39	3.34	94.493	380.44	327.62
Steviolbioside	17	1.32	5.01	217.47	442.98	587.12
Stevioside	22	1.31	6.44	276.06	527.91	714.62
Sucrose	10	1.22	3.27	163.18	174.02	279.50
Saccharine	5	0.82	2.22	135.36	160.52	142.50

Table 3.6: Structural properties of the docked compounds. The diameter is a 2D topological descriptor related to the size of the molecule, rgyr is the radius of gyration [Å], ASA+ is the solvent accessible area positively charged, ASA_H is the hydrophobic solvent accessible area [$Å^2$], and vol is the volume of the molecule [$Å^3$].

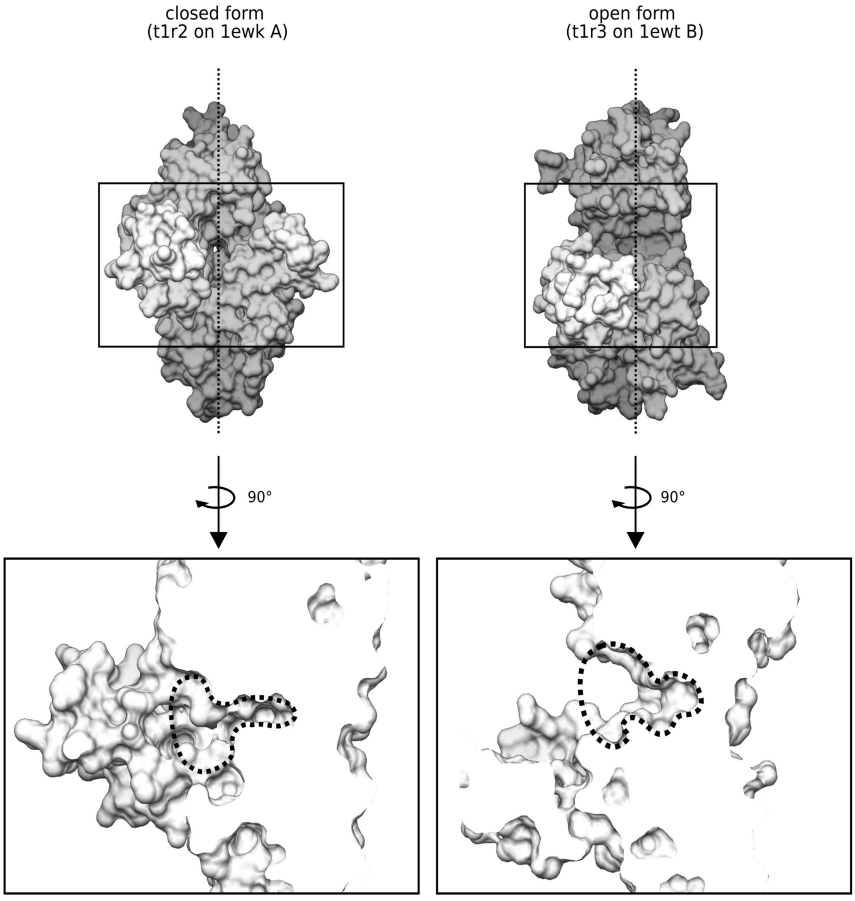
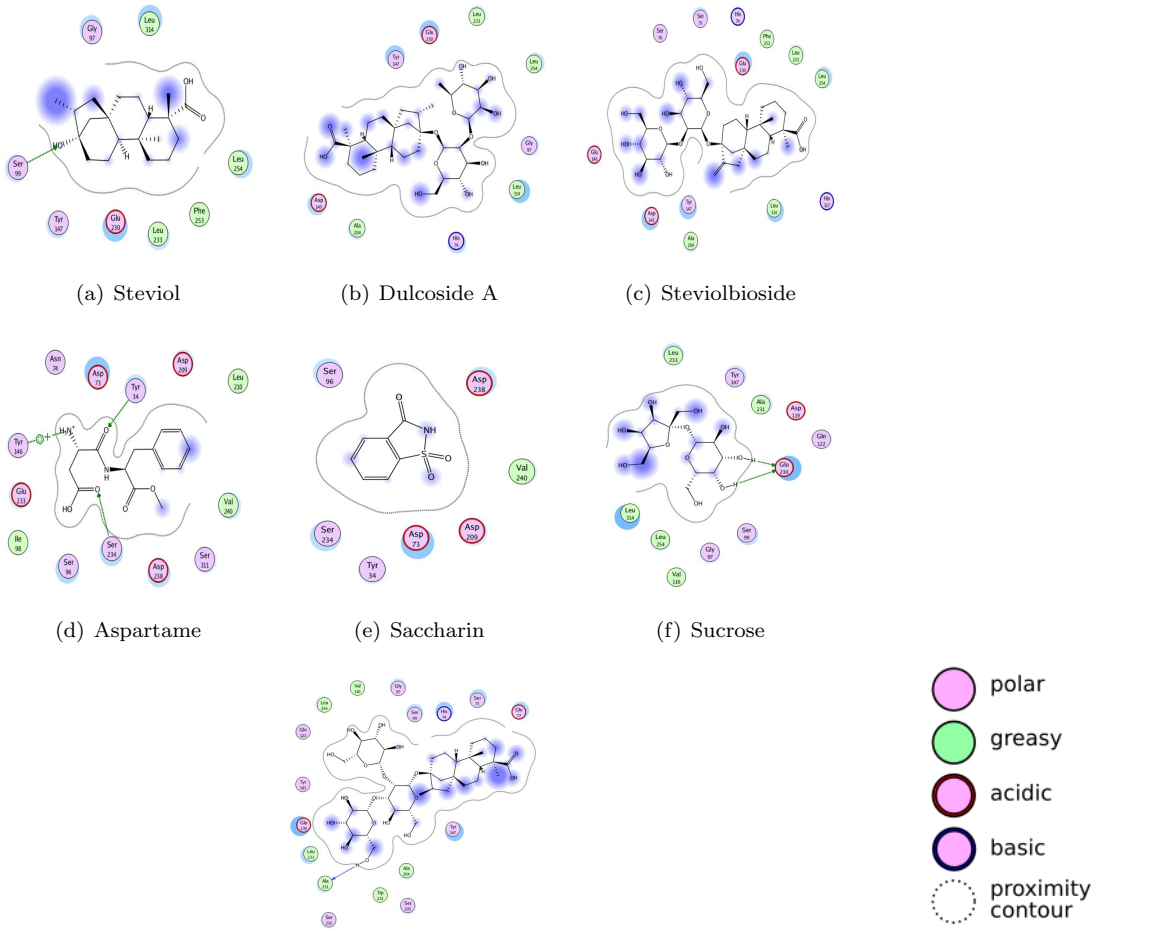


Figure 3.7: Van der Waals surface representation of the closed (t1r2 on 1ewk A) and open (t1r3 on 1ewt B) monomers. For a better visual comparison of the internal cavity of the two proteins, a section of each protein was made at the intersecting plane indicated with a fine dashed line. Moreover a qualitative profile of each cavity is drawn.



(a) Steviol (b) Dulcoside A (c) Steviolbioside

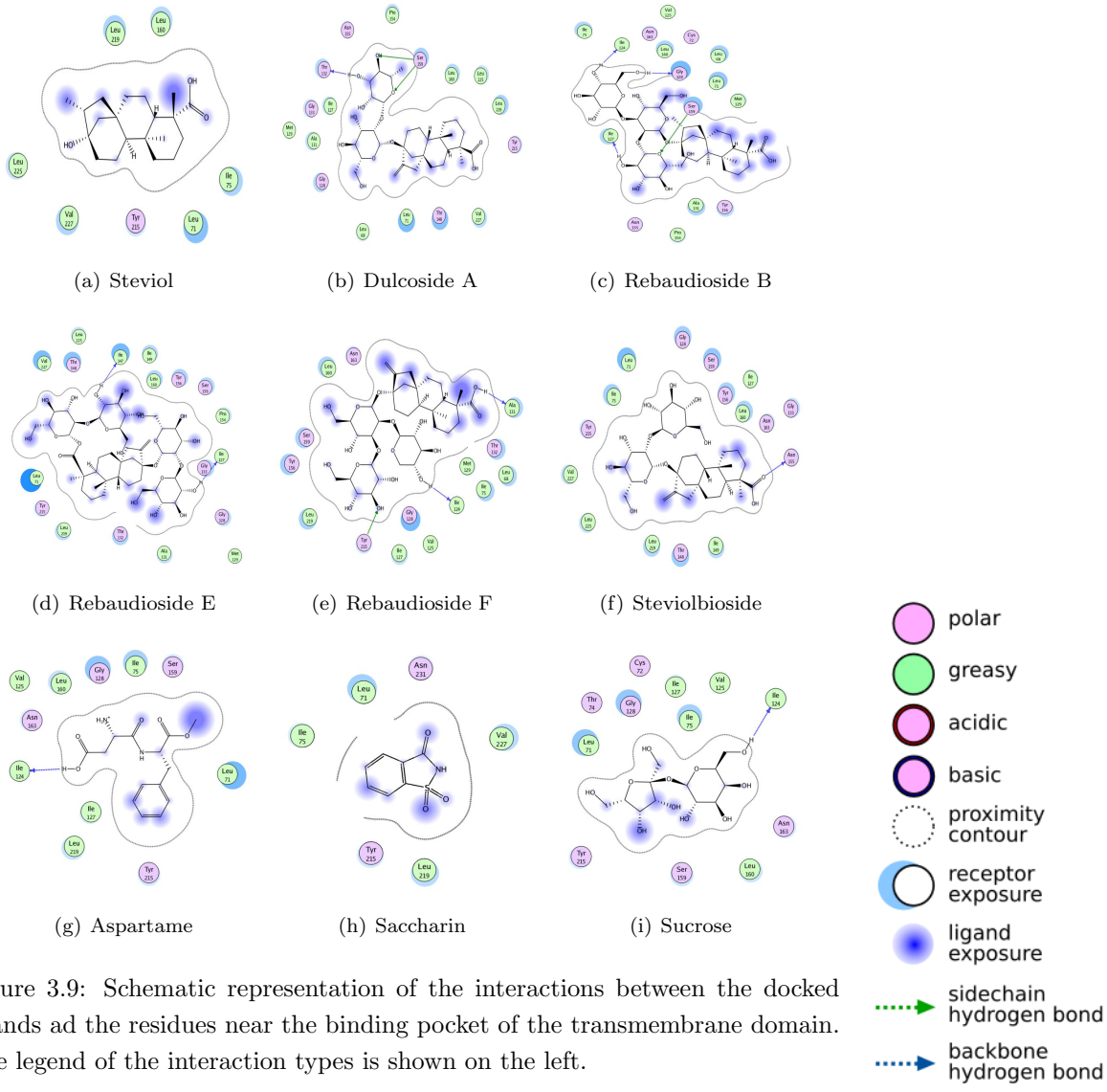
(d) Aspartame (e) Saccharin (f) Sucrose

(g) Rebaudioside B

Figure 3.8: Schematic representation of the interactions between the docked ligands ad the residues near the binding pocket of the extracellular portion of the sweet taste receptor. Only the compounds with the lowest free energy value are depicted. The legend of the interaction types is shown on the left.

3.3.4 The transmembrane domain binding pocket

Analyzing the interactions formed by the bound ligands with the residues of the binding pocket of the transmembrane portion (figure 3.9), we can see that the backbone oxygen of the isoleucine 124 (I124) is often involved as hydrogen bond donor. In particular it binds an oxydril of the rebaudioside B, rebaudioside F, aspartame, and sucrose. The I124 is located at the end of the fourth helix and the backbone carbonyl group is oriented toward the binding pocket. The backbone carbonyl of the isoleucine 127 (I127) also seems to play an important role as it is involved in hydrogen bonding of some ligands (rebaudioside B and rebaudioside E).
I124 and I127 are not strongly conserved residues, as depicted in the multiple alignment in figure 3.11. These residues may be involved in substrate recognition (that usually varies among different proteins) and are not conserved into paralogues protein. Indeed, plotting the conservation index (shown on the alignment) onto the 3D structure of the modelled TM domain of the receptor (figure 3.10), it is possible to see that the highest conserved residues are located into the cytoplasmic region (bottom of the protein) and not near the ligand binding site (top side of the protein). This implies that the binding site is composed by residues that are specific for a given species, i.e. the substrates that bind the human receptor may not be able to bind the receptor of other species.



(a) Steviol (b) Dulcoside A (c) Rebaudioside B

(d) Rebaudioside E (e) Rebaudioside F (f) Steviolbioside

(g) Aspartame (h) Saccharin (i) Sucrose

Figure 3.9: Schematic representation of the interactions between the docked ligands ad the residues near the binding pocket of the transmembrane domain. The legend of the interaction types is shown on the left.

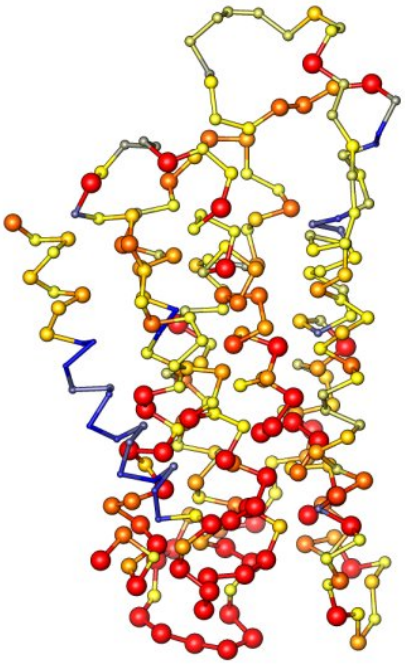


Figure 3.10: The conservation index is mapped on the C_α carbons of the t1r2 receptor. Lower values of conservation index are coloured in blue, intermediate values in yellow and higher values in red. The upper part of the protein is the extracellular portion while the lower part is the cytoplasmatic portion.

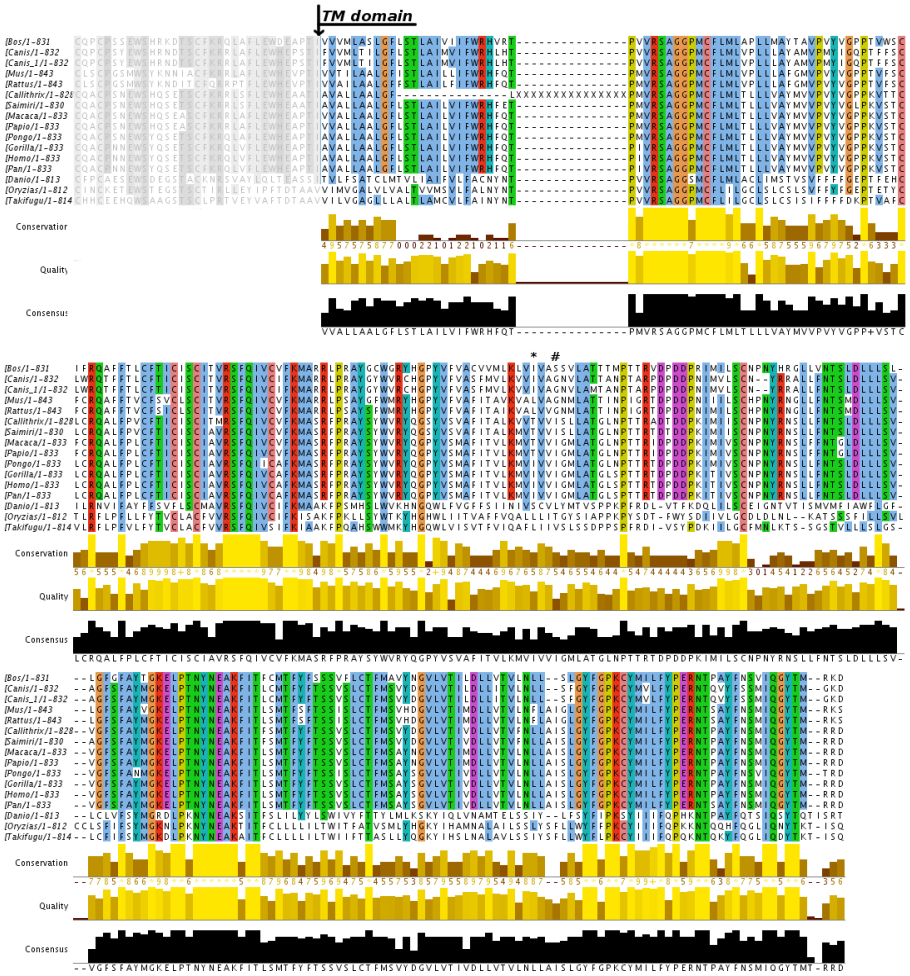


Figure 3.11: Multiple alignment of the transmembrane portion of t1r2. Homologues sequences were identified using BLAST. Then a multiple alignment was made by means of T-coffe. Residues are coloured according to the clustalX colouring scheme. Only the residues conserved more that 70% are coloured. The I124 and the I127 are indicated with * and # respectively.

3.4 Conclusions

In this work we have studied the interaction between the a new class of natural sweeteners, the steviol glycosides, and the human sweet taste receptor t1r2-t1r3 by means of a detailed computational study. As the crystal structure of the sweet taste receptor has not been experimentally determined we have predicted it before proceeding to the interaction study. A three-dimensional model of the transmembrane portion and different 3D models of the extracellular domains has been obtained. The interaction between 10 different steviol glycosides and four small artificial sweeteners and all the predicted models has been studied. Form this analysis we found that the binding site located on the extracellular portion is more suited for small compounds such as the aspartame or the saccharin, while larger compounds tend to prefer the transmembrane binding site. It is worth to notice that the protein was kept fixed during the docking procedure, this approximation may be responsible for the partial disagreement of the computational predicted free energies and the experimental ones (table 3.5). We are currently introducing the flexiblity of the protein by means of the use of the normal mode analysis. Moreover as the steviol glycosides have not any experimental free energies to be compared to, we are planning to include a new set of experimental values for the them.

Bibliography

[1] Delos, http://www.delos-bio.it/.

[2] Sali A and Blundell T L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815, 1993.

[3] Sciafani A. Sweet taste signaling in the gut. *PNAS*, 104(38):14887–14888, 2007.

[4] Temussi P A. Why are sweet proteins sweet ? interaction of brazzein, monellin and thaumatin with the t1r2-t1r3 receptor. *FEBS Letters*, 526:1–4, 2002.

[5] Geuns J M C. Stevioside. *Phytochemistry*, 64:913–921, 2003.

[6] Notredame C, Higgins D J, and Heringa J. T-coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[7] Jianguo Chen, Per Bendix Jeppesen, Iver Nordentoft, and Kjeld Hermansen. Stevioside improves pancreatic beta-cell function during glucotoxicity via regulation of acetyl-coa carboxylase. *Am. J. Physiol. Endocrinol. Metab.*, 292:1906–1916, 2007.

[8] Higgins D, Thompson J, Gibson T, Thompson J D, Higgins D G, and Gibson T J. Clustal w: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680.

[9] Morini G, Bassoli A, and Temussi P A. From small sweeteners to sweet proteins: anatomy of the binding sites of the human t1r2-t1r3 receptor. *J. Med. Chem.*, 48:5520–5529, 2005.

[10] Xu H, Staszewski L, Tang H, Adler E, Zoller M, and Li X. Different functional roles of t1r subunits in the heteromeric taste receptors. *PNAS*, 101(39):14258–14263, 2004.

[11] Chemical Computing Group Inc. *Montreal, Quebec, Canada*, 2007.

[12] Bond P J and Sansom M S P. Membrane protein dynamics versus environment: Simulations of ompa in a micelle and in a bilayer. *J. Mol. Biol.*, 329:1035–1053, 2003.

[13] Naoki K, Yoshimi S, Yuji T, Toshihiro S, Masaki Y, Takashi K, Shigetada N, Hisato J, and Kosuke M. Structural basis of glutamate recognition by a dimeric metabotropic glutamate. *nature*, 407:971–977, 2000.

[14] Kellett G L, Brot-Laroche E, Mace O J, and Leturque A. Sugar absorption in the intestine: The role of glut2. *Annu. Rev. Nutr.*, 28:35–54, 2008.

[15] Winnig M, Bufe B, and Meyerhof W. Valine 738 and lysine 735 in the fifth transmembrane domain of rtas1r3 mediate insensitivity towards lactisole of the rat sweet taste receptor. *BMC Neuroscience*, 6(22):1–7, 2005.

[16] Jiang P, Cui M, Zhao B, Liu Z, Snyder L A, Benard L M J, Osman R, Margolskee R F, and Max M. Lactisole interacts with the transmembrane domains of human t1r3 to inhibit sweet taste. *J. Biol. Chem.*, 280(15):15238–15246, 2005.

[17] Mereghetti P, Ganadu M L, Papaleo E, Fantucci P, and De Gioia L. Validation of protein models by a neural network approach. *BMC Bioinformatics*, 9, 2008.

[18] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipotand Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *J. Comp. Chem.*, 26:1781–1802, 2005.

[19] Whitehouse C R, Boullata J, and McCauley L A. The potential toxicity of artificial sweeteners. *American Association of Occupational Health Nurses*, 56:251–259, 2008.

[20] Irum Sehar, Anpurna Kaul, Sarang Bani, Harish Chandra Pal, and Ajit Kumar Saxena. Immune up regulatory response of a non-caloric natural sweetener, stevioside. *Chemico-Biological Interactions*, in press, 2008.

[21] MJ Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[22] Tancredi T, Pastore A, Salvadori S, Esposito V, and Temussi P A. Interaction of sweet proteins with their receptor. a conformational study of peptides corresponding to loops of brazzein, monellinand thaumatin. *Eur. J. Biochem.*, 271:2231224.

[23] Wiederstein and Sippl MJ. Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35:407–410.

[24] Li X, Staszewski L, Xu H andDurick K, Zoller M, , and Adler E. Human receptors for sweet and umami taste. *PNAS*, 99(7):46924696, 2002.

Chapter 4

Relationship between dynamical properties and function : the psychrophilic enzymes

Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed uponit.
Isaac Newton (1643 - 1727)

4.1 Introduction

Extremophiles are mainly microorganisms experimenting unusual environmental conditions compared to organisms living roughly under atmospheric pressure, at temperatures close to the average temperature on earth which is about 15°C, using O_2 as a source of electron acceptor and metabolising substrates at pH values close to neutrality [7]. The earth's surface is, however, dominated by low temperature environments, made up of extremely cold parts such as the Arctic and the Antarctic, moderately cold parts such as mountain regions and a huge, cold and stable ecosystem, namely the marine waters which cover 70% of the earth's surface and display, below 1000 m, temperatures not

exceeding 5°C independently of the latitude [7, 15]. All these environments are permanently cold habitats exerting on ectothermic populations a highly selective pressure [7]. They have been colonised, despite their extreme character, by largely diversified organisms which have developed adaptation strategies enabling them not only to survive but also to grow successfully, like true **psychrophiles** [7]. Microorganisms are the most abundant psychrophiles in terms of species diversity and are likely to be the most abundant in terms of biomass [14].

Psychrophiles
Organism living permanently at temperatures close to the freezing point of water, that is devoid of temperature regulation and that is unable to develop at mesophilic temperatures [14].

4.1.1 Cold-adaptation and Biotechnology

Psychrophilic enzymes are not only of extraordinary interest at the fundamental level to investigate the thermodynamic stability of proteins, but also to understand the relationship between stability, flexibility and catalytic activity. The knowledge of these relationships could help in site-direct mutagenesis experiments to obtain the enzyme with the desired characteristics.

4.1.2 Strategies in Cold Adaptation

The immediate consequences of the low temperatures are a low heat-content (enthalpy) and a reduction in the amplitudes and frequencies of atomic motions as well as of molecular motions. Psychrophilic organisms have to face and overcome a variety of challenges to survive, such as to avoid freezing of the intracellular fluid, maintenance of membrane fluidity and permeability, and probably the most important factor, to cope with the reduction in chemical reaction rates induced by low temperatures [28].

Cold Adapted Enzymes

Enzymes of psychrophilic organisms have to cope with the reduction of the reaction rate due to the low temperature. A strategy could be to increase the enzyme concentration: this is energetically expensive and, therefore, it was reported only in few cases. The solution found during the evolution was to develop a repertoire of specific enzymes able to carry out the biological function in these extreme environment. Psychrophilic enzymes have three basic features [29]:

- to compensate for the slow reaction rates at low temperatures, psychrophiles synthesize enzymes with an up to tenfold higher specific activity in this temperature range. In fact this is the main physiological adaptation mechanism to cold at the enzyme level;

- the temperature for apparent maximal activity for cold-active enzymes is shifted towards low temperatures, reflecting the weak stability of these proteins which are prone to unfolding and inactivation at moderate temperatures;

- the adaptation to cold is frequently not perfect. The specific activity of the psychrophilic enzyme at low temperatures, although very high, remains generally lower than that of the mesophilic enzymes at 37°C.

The adjustment of the conformational flexibility of proteins at the environmental temperatures is achieved by reducing the number or strength of weak interactions stabilising the folded and biologically active conformation. Furthermore, some psychrophilic proteins reduce the hydrophobicity of the core clusters or expose a larger hydrophobic surface to the solvent: both induce an entropy-driven destabilization by weakening the hydrophobic effect on folding [28, 14, 13, 32, 30]. These properties result in a poor structured conformation characterized by a low conformational stability and a marked heat-lability of the activity in enzyme catalysts.

Biocatalysis at Low Temperatures

The rate of reactions is described by the Arrhenius equation:

$$k_{cat} = Ake^{-E_a/RT} \tag{4.1}$$

where k_{cat} is the enzyme reaction rate at a given temperature, which is expressed as the number of substrate molecules that are transformed by one molecule of enzyme per unit of time (it is also known as the turnover number), A is the preexponential factor, k is the dynamic transmission coefficient (generally assumed to be 1) and R is the universal gas constant ($8.314 Jmol^{-1}K^{-1}$).

According to this equation, the reaction rate increases with an increase in absolute temperature (T) and a decrease in activation energy (E_a) [32]. The majority of cold-adapted enzymes have a higher k_{cat} and K_m than their thermostable counterparts, with the exception of enzymes that work at a substrate concentration close to K_m. At very low temperatures (0°-4°C), insufficient kinetic energy is available in the system to overcome reaction barriers: psychrophilic organisms have evolved several strategies to adapt at these conditions. The majority of cold-adapted enzymes are characterised by a shift in apparent T_{opt} (optimum temperature of activity) to a low temperature with a concomitant decrease in stability. They also tend to exhibit a high reaction

rate by decreasing the activation free energy ($\Delta G^{\#}$), which is the barrier between the ground state (substrate) and the transition state ($ES^{\#}$). The $\Delta G^{\#}$ is composed of two components:

$$\Delta G^{\#} = \Delta H^{\#} - T\Delta S^{\#} \tag{4.2}$$

where $\Delta H^{\#}$ is the change in activation enthalpy, $\Delta S^{\#}$ is the change in activation entropy and T is the absolute temperature.

According to transition-state theory, k_{cat} is related to temperature and thermodynamic activation parameters by the following equation:

$$k_{cat} = (k_B T/h)e^{-\Delta G^{\#}/RT} \tag{4.3}$$

where k_B is the Boltzman constant ($1.38 * 10^{-23} JK^{-1}$) and h is the Plank constant ($6.63 * 10^{-34} Js$).

In almost all cold-adapted enzymes studied the trend is to have a low $\Delta H^{\#}$ and the result is that the reaction rate tends to be less dependent on temperature and a high reaction rate (k_{cat}) is maintained at low temperature. Joining the latter equations is useful to consider the effect of $\Delta S^{\#}$ and $\Delta H^{\#}$ on k_{cat}

$$k_{cat} = (k_B T/h)e^{-\{(\Delta H^{\#}/RT)+(\Delta S^{\#}/R)\}} \tag{4.4}$$

To increase k_{cat} at low temperatures, either $\Delta S^{\#}$ has to increase or $\Delta H^{\#}$ has to decrease. In cold-adapted enzymes studied until now a decrease in $\Delta H^{\#}$ is generally observed when compared to mesophilic counterparts. The difference in activation entropy between an enzyme from a psychrophile and a mesophile ($\Delta(\Delta S^{\#})_{p-m}$) is always negative whatever the sign of the activation entropy [8]. Enthalpy-entropy compensation implies that a decrease in $\Delta H^{\#}$ accompanied by a decrease in $\Delta S^{\#}$ produces an overall small change in $\Delta G^{\#}$ [32]. Thus, in an enzymatic reaction catalyzed by cold-adapted enzymes, the decrease of the $\Delta H^{\#}$ can be considered as the main adaptative parameter. The corresponding decrease in activation energy is achieved structurally by a decrease in the number of enthalpy-driven interactions that have to be broken to reach the transition state. This indicate a lower stability of the psychrophilic enzymes and hence also a greater flexibility at or near the catalitic site [8].

4.1.3 Flexibility in Cold-Adapted Enzymes

Cold-adapted enzymes show low stability and high activity at low temperature that implies a flexible enzyme structure. Interestingly, amino acids involved in

catalysis are conserved between cold and thermostable homologs and therefore gain in flexibility should involve other regions [32]. Enzyme catalysis generally involves the movement of all or of a particular region of the enzyme, enabling the accommodation of the substrate. The ease of such molecular movement may be one of the determinants of catalytic efficiency. Therefore, optimizing a function of an enzyme at a given temperature requires a proper balance between two of the opposing factors: structural rigidity (allowing the retention of a specific three-dimensional conformation at the physiological temperature) and flexibility (allowing the protein to perform its catalytic function). At room temperature, a thermophilic enzyme would therefore be stable but poorly active: this is due to an increase in the molecular edifice rigidity induced by the low thermal energy in the surroundings, thus preventing essential movement of residues. Since a thermophilic protein is in general related with the rigidity of the structure, a psychrophilic one, at the opposite end of the temperature scale, should be characterised by an increase of the plasticity or flexibility of appropriate parts of the molecular structure in order to compensate with the lower thermal energy provided by the low temperature habitat [8]. Psychrophilic enzymes can increase their flexibility from a general reduction in strength of intramolecular forces (*global flexibility*) or from weakened interactions in one or a few important regions of the structure(*localized flexibility*). Instead, a gain in flexibility in some specific regions has been demonstrated in the cases of serine-proteases [9], uracil DNA glycosidase [21], A4-lactate dehydrogenase [2]: in these enzymes the flexibility is increased in small areas that affect the mobility of adjacent active-site structures. Flexibility is difficult to evaluate with experimental methods. For example, B-factors evaluated from a largest set of X-ray structure is used as a static index of flexibility. Dynamic flexibility is measured by dynamic fluorescence quenching and proteolytic nicking [32]. The flexibility of a protein, especially that related to activity and/or stability, remains a difficult parameter to determine experimentally, as the increase in flexibility can be limited only to a small but crucial part of the protein [8]. Molecular dynamics (see section 1.4.2) is, at now, a suitable technique that can be used to calculate all atoms protein motions during time: it permits to simulate the behaviour of the studied system at atomic level, so it can be used to evaluate protein flexibility [9, 24].

4.1.4 Activity-Stability-Flexibility relationship

The high activity and low stability of cold-adapted enzymes underlie a general principle of activity-stability trade-off. Since there is a conservation in the

amino acids involved in cold and thermostable homologs, the cause of flexibility must reside in other parts of the enzyme.
The available data regarding cold-adapted enzymes indicates that a high specific activity is almost always associated with a low thermostability. In general, thermophilic enzymes characterised by a high thermostability are poor catalysts at room temperature. The thermostability derived from the pronounced rigidity of the molecular edifice is thought to impair interaction between substrate and enzyme, leading to a weak specific activity. By contrast, flexibility or plasticity of the molecular structure would enable greater complementarity at a low energy cost, thus explaining the high specific activity of cold-adapted enzymes.

To shed light on the molecular features responsible for cold adaptation in psychrophilic enzyme, similarly to previous works in which has been carried out a comparison between mesophilic and thermophilic enzymes [18], we have performed comparative molecular dynamics studies [9] between mesophilic and psycrophilic variants belonging to two different enzymatic families: the serine-proteases and the uracil-DNA glycosilases. In particular, using multiple molecular dynamics simulations, an accurate sampling of the near-native conformations was obtained. To efficiently and accurately sample the potential energy surface, multiple MD simulations were carried out in explicit solvent at 283 and 310 K, close to the optimal growth temperatures for the organisms, collecting 0.1 μs trajectories. For every enzyme, the resulting ensemble was analyzed estimating and comparing the near-native free energy landscapes considering different reaction coordinates. In particular, the principal components of the trajectories, the radius of gyration and the root mean square deviation were used as collective coordinates. Moreover the configurational entropy, estimated using the formula for a quantum mechanical (QM) oscillator [17] and the potential energy contributions, were also computed.

4.2 Methods

4.2.1 Molecular dynamics simulations

MD simulations were performed using the 3.3 version of the GROMACS software (www.gromacs.org), using GROMOS96 force field. The X-ray structures of two mesophilic and psychrophilic proteins were used as starting points for the MD simulations. In particular, a mesophilic (from *Sus scrofa*, PDB ID: 1lvy) and psychrophilic (from *Salmo salar*, PDB ID: 1elt) elastases and a mesophilic (from Homo sapiens, PDB ID: 1akz) and psychrophilic (from *Gadus morhua*, PDB ID: 1okb) uracil-DNA-glycosilases were selected.
The name of each enzyme were abbreviated as following :

- mPE: mesophilic Porcine Elastase
- pSE: psychrophilic Salmon Elastase
- mHUDG: mesophilic Human Uracil-DNA-Glycosilases
- pCUDG: psychrophilic Cod Uracil-DNA-Glycosilases

Protein structures, including the crystallographic water molecules and calcium ions for elastases, were soaked in a dodecahedral box of SPC water molecules and simulated using periodic boundary conditions. For UDG all the histidine with one exception (His148) were considered as neutral in the simulations, as explained in ref. [25]. Productive MD simulations were performed in the NPT ensemble at 283 and 310 K, using an external bath with a coupling constant of 0.1 ps. Pressure was kept constant (1 bar) by modifying box dimensions and the time-constant for pressure coupling was set to 1 ps [6]. The LINCS [16] algorithm was used to constrain bond lengths, allowing the use of a 2 fs time step. Long range electrostatic interactions were calculated using the Particle-mesh Ewald (PME) [34] summation scheme. Van der Waals and Coulomb interactions were truncated at 1.0 nm. The non-bonded pair list was updated every 10 steps and conformations were stored every 2 ps. To improve the conformational sampling, ten 12 ns simulations were carried out for each protein system at 283 and 310 K, respectively, initializing the MD runs with different initial atomic velocities taken from a Maxwellian distribution. In the following, MD trajectories collected for the same system but characterized by different initial velocities are referred to as replica 1 to replica 10.
The root mean square deviation (rmsd), which is a crucial parameter to evaluate the equilibration of MD trajectories, was computed for mainchain atoms using

the starting structure of the MD simulations as reference. The analysis of MD trajectories have been carried out discarding the first 2 ns for each simulation in order to ensure stable values for mainchain rmsd.
Investigation of elastase structures during the simulation time shows that the coordination of the calcium ions, which is important for both function and stability of elastases, is maintained throughout the simulations in agreement with data previously reported [9]. Moreover, potential and total energy of the system, as well as the protein gyration radius are constant throughout the simulations (data not shown). For each protein, the stable region of the ten replicas at the same temperature were joined in a concatenated trajectory, which is representative of different directions of sampling around the starting structure.

4.2.2 Essential dynamics analysis

The all-atoms covariance matrix (C) was calculated on the equilibrated portions of the trajectories. In particular C was calculated considering both the concatenated trajectories and the single replicas for each system at both 283 and 310 K:

$$C = cov(x) = \overline{(\mathbf{r}^N - \overline{\mathbf{r}^N})(\mathbf{r}^N - \overline{\mathbf{r}^N})^T} \tag{4.5}$$

where $\bar{\cdot}$ is the average and $\mathbf{r}^N$ is the vector of the atomic position.
After removal of the translational and rotational degrees of freedom (fitting each structure onto the initial one), the matrix C was calculated and then diagonalized to obtain the eigenvectors and eigenvalues, which give information about correlated motions throughout the protein. To define the dimensionality of the essential subspace, the fraction of total motion described by the reduced subspace was considered and computed as the sum of the eigenvalues relative to the included eigenvectors, describing the amount of variance retained by the reduced representation of the total space. A measure of the similarity of a MD trajectory to random diffusion is the cosine content (c_i) of the pi principal component [5]:

$$c_i = \frac{1}{T}\left(\int cos(it)p_i(t)dt\right)(p_i(t)dr)^{-1} \tag{4.6}$$

where T is the total simulation time and p_i is the i-th principal component.
c_i is an absolute measure that can be extracted from covariance analysis and ranges between 0 (no cosine) and 1 (a perfect cosine). It has been demonstrated that insufficient sampling can lead to high ci values, representative of random motions. The evaluation of cosine contribution for first eigendirections is sufficient to give a reliable idea of the protein behaviour [5]. When the cosine

content of the first few PCs is close to 1, the largest scale motions in the protein dynamics resemble diffusion, and can not be interpreted in terms of characteristic features of the energy landscape [26].
The analysis of the sampling convergence can be performed computing the root mean square inner product (RMSIP, equation 4.7) as a measure of similarity between subspaces defined by their basis vectors [1]:

$$RMSIP = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\eta_i^A \eta_j^B\right) \tag{4.7}$$

Where η_i^A and η_j^B are the eigenvectors of the spaces to be compared.
Usually the RMSIP is computed onto the first 10 eigenvectors [12, 1]. The statistical significance of the observed RMSIP value was tested by simulating an empirical distribution of RMSIP data under the null hypothesis of no relationship between both spaces [12, 4]. In particular, the dependence of the RMSIP by the dimensionality of the spaces can be expressed as $RMSIP(K) = \frac{\sqrt{D}}{\sqrt{K}}$, where D and K indicate the number of eigenvectors considered in the computation of RMSIP and the dimensionality of the two spaces, respectively.
Comparing these results with the RMSIP distribution of our systems it is possible to assess the significance of our RMSIP [12, 4].

4.2.3 Analysis of the potential energy

The Gromacs potential energy function was decomposed into the following contributions: angle, proper dihedral, improper dihedral, Coulomb 1-4, Coulomb short-range protein-protein, Coulomb long-range protein-protein, Lennard-Jones 1-4, Lennard-Jones short-range protein-protein, Lennard Jones short-range protein-solvent, Coulomb short-range protein-solvent and, Coulomb long-range protein-solvent.
The long-range electrostatic interactions during the molecular simulation were computed using the PME scheme previously described. Because of the non pair-additive property of the PME algorithm, to obtain the long-range contributions of the protein we have re-computed the electrostatic interactions on the trajectories without PME, and increasing the cut-off of the electrostatics interaction to 1.4 nm.

4.2.4 Free energy landscapes

Given a reaction coordinate[a] q, the probability of finding the system in a particular state is proportional to $exp^{-\frac{G(q_\alpha)}{kT}}$, where $G(q_\alpha)$ is the free energy of that state.
The free energy landscape can be computed from $G(q_\alpha) = -kTlnp(q_\alpha)$. Where k is the Boltzmann constant, T is the temperature of the simulation and $P(q_\alpha)$ is an estimate of the probability density function obtained from a histogram of the data. To ensure that for the lowest free energy minimum the $G = 0$ we have subtracted the maximum probability $G(q_\alpha) = -kTlnp(q_{alpha}) - P^{max}(q)$. Considering two different reaction coordinates ,the two-dimensional free energy landscapes were obtained from the joint probability distributions of the considered variables [18]. In particular, the reaction coordinates investigated were: the rmsd calculate on the C_α atoms using the starting structure of the molecular dynamics simulation as reference; the radius of gyration (R_g) computed on the C_α atoms and the cartesian principal components derived by ED analysis, as previously described.

4.2.5 Cluster analysis

Clustering on cartesian coordinates was performed computing the root mean square distance matrix calculated for C_α atoms between each pairs of structure using GROMACS. Therefore, the complete linkage algorithm implemented in Matlab was applied onto this distance matrix obtaining a dendrogram.

4.2.6 Configurational entropy

The configurational entropy was computed using the formula for a quantum mechanical (QM) oscillator (equation 4.8, as suggested by Andricioaei and Karplus [17].

$$S_{qm} = k \sum_{i=1}^{3N-6} \left[\frac{\alpha_i}{exp^{\alpha_i} - 1} - ln\left(1 - exp^{-\alpha_i}\right) \right] \tag{4.8}$$

[a] An *order parameter* is a variable chosen to describe the degree of order in the system, or, even more generally it is a variable chosen to describe changes in the system. In the free energy context, order parameters are collective variables that are used to describe transformations from the initial to the final state. An order parameter may (does not necessarily have to) correspond to the path along with the transformation takes place in nature. In this case, it would be called *reaction coordinate*, or *reaction path*.

where $\alpha = \frac{\hbar\omega}{kT}$, $\hbar = \frac{h}{2\pi}$, ω is the frequency of the oscillator, and h is the Planck's constant.

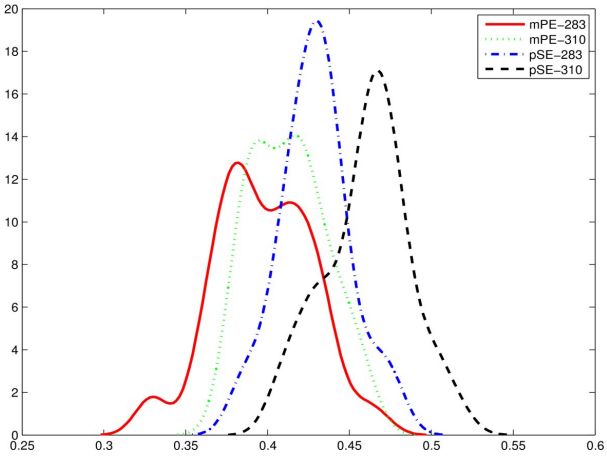
The frequency ω is connected to the variance through equipartition theorem $m\omega\langle x^2\rangle = kT$. The entropy of the harmonic oscillator is an upper bound for the true entropy of the system [19]. Several approximations are used into the computation of the entropy : *i.* every degree of freedom is treated as a quantum harmonic oscillator; *ii.* the equipartition theorem is used to connect the classical variance to the frequency of a quantum harmonic oscillator. This relation holds for $\hbar\omega \ll kT$, which is a good approximation given that the high-frequency of motion for which fails will contribute little to the entropy; *iii.* absence of supra-linear correlation between different coordinates.

4.3 Results and Discussion

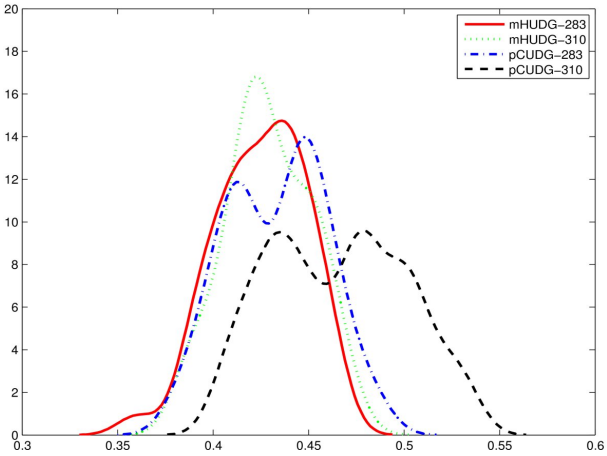
4.3.1 Evaluation of the conformational sampling

Molecular dynamic simulations of mesophilic and psychrophilic elastases and uracil-DNA-glycosylases were carried out at 283 and 310 K (see Materials and Methods). After concatenation of the equilibrium portions of the trajectories, the resulting MD ensembles consisted of 0.1 μs trajectories for each system at both temperatures. In order to gain insights into the configurations visited by the system and to evaluate the conformational sampling essential dynamics analysis (see Materials and Methods) was carried out, with particular attention to the direction of motion along the first eigenvectors. In fact, the first three eigenvectors are sufficient to describe a consistent part of the total motion, and the subspace defined by them could be used as the three-dimensional (3D) reference subspace to analyze protein dynamics. The projections of simulation frames in the 3D-reference subspace shown a wide sampling of the conformational space with re-sampling of similar conformations in our simulations, indicating that the essential subspace is well explored when concatenated trajectories are considered.
To further evaluate the sampling efficiency, we have also computed the cosine content (c_i) of the principal components of protein motion, which is a measure of the similarity of the trajectories to random diffusion. It turned out that single simulations are often characterized by relatively large c_i in the first eigenvectors, and therefore partly describe a random diffusion motion, while the corresponding concatenated trajectories have lower or null cosine content and therefore adequately represent essential and significant motions.
Another measure used for the convergence assessment is the root mean square inner product (RMSIP). For each protein system (mPE, pSE, mCUDG, pHUDG) the RMSIP was computed comparing all replicas each other, obtaining a distributions of RMSIP values (figure 4.1). The average values of the distributions are summarized in table 4.1

To compute the expected RMSIP of two unrelated spaces we have obtained an empirical distribution of RMSIP values considering the first 10 eigenvectors of random (normal distributed) orthogonal matrices of different size (from 10 to 500). We found that expected RMSIP depends on the dimensionality of the considered space and on the subset of eigenvectors included in the RMSIP calculation, see figure 4.2).



(a) ELASTASE



(b) UDG

Figure 4.1: RMSIP distributions estimated using a kernel density estimation as implemented in Matlab. As kernel smoother, the normal distribution was chosen.

System	μ^a_{RMSIP}	σ^a_{RMSIP}	$\mu^b_{RMSIP^r}$	$\sigma^b_{RMSIP^r}$	n. $atoms^b$	n. $eigenvalues^b$
mPE, 283	0.39	0.02	0.04	0.01	2349	7047
mPE, 310	0.41	0.02				
pSE, 283	0.43	0.02	0.04	0.01	2259	6777
pSE, 283	0.46	0.03				
mHUDG, 283	0.42	0.02	0.04	0.01	2334	7002
mHCDG, 310	0.43	0.02				
pCUDG, 283	0.43	0.03	0.04	0.01	2297	6891
pCUDG, 310	0.41	0.03				

Table 4.1: Average of the RMSIP values (μ_{RMSIP}) and standard deviation (σ_{RMSIP}). a. Average and the standard deviation computed on the RMSIP distribution for every simulated system. b. Expected values and standard deviation of of empirical RMSIP distribution of unrelated spaces.

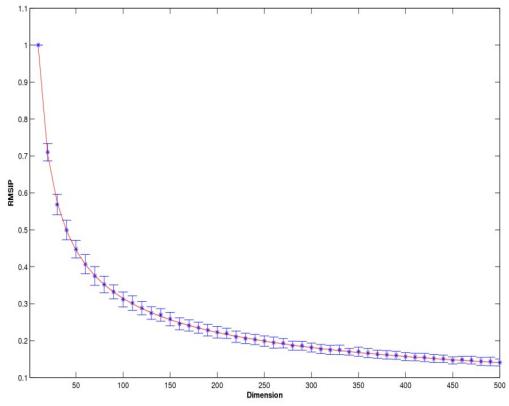


Figure 4.2: RMSIP averages and standart deviations plotted versus the dimensionality of the random orthogonal matrices.

Considering the dimensionality of our system (ranging from a minimum of 2259 atoms to 2349 atoms), using the previously identified relationship we found that if the set of principal components were completely unrelated, the expected RMSIP value would be 0.038 (table 4.1). In particular, the average RMSIP of our systems (table 4.1) is sufficiently distant from this random reference, allowing to conclude that a good conformational sampling was achieved by our simulations. The average RMSIP obtained for our simulations is sufficiently distant from this random reference and presents a sufficient overlap between different replicas of the same system, allowing to conclude that a proper conformational sampling was reached by our simulations.
In conclusion, all the indexes analyzed by ED indicate a wide conformational sampling by our trajectories and allow to confidently analyzed the free energy landscape of the mesophilic and psychrophilic enzymes under analysis.

4.3.2 Free energy landscapes

The study of free energy landscape could give an accurate picture of the protein structural properties around the native-state [18, 12]. However, in order to understand the behaviour of complex systems it is necessary to project it onto low dimensional subspace of physically meaningful coordinates. The choose of the reaction coordinates is a system-related task and is usually driven by the features that have to be depicted. In particular only the degree of freedom directly related to the properties of interest should be included in the analysis to prevent masking of important informations [3].
Since many different free energy landscapes can be obtained using different combinations of collective variables, we have selected as reaction coordinates the following properties: the radius of gyration (Rg), the root mean square deviation (rmsd), and the cartesian principal components (See Methods for details). The cartesian principal components analysis, in particular, is a frequently used method for obtaining collective coordinates for projecting the configurational free energy landscape of proteins [20, 22, 27].
We would like to point out that this kind of free energy landscapes, which lack barrier in-formation, reflect the overall shape of the free energy surface and not necessarily its details [23, 22]. The absence of barriers, is somehow compensated by the empty space between the sampled regions: poorly or unsampled regions often correspond to high energy regions [23]. Moreover, because of the large dimensionality reduction due to the projection onto few collective coordinates,

these maps may represent an incomplete description of the free energy profile of the protein. This lack of information can be partially complemented by means of a cluster analysis, which allows to represent the geometrical relationship in a multidimensional way [22].
In light of the above observations, we carried out the analysis of the FEL using different reaction coordinates and integrated it by a structural clusters analysis, which was performed considering the cartesian coordinates (Materials and Methods).
Since the two model systems (elastases and uracyl-DNA glycosylases) show a different behaviour for most part of the analyzed aspects, we prefer to discuss them separately.

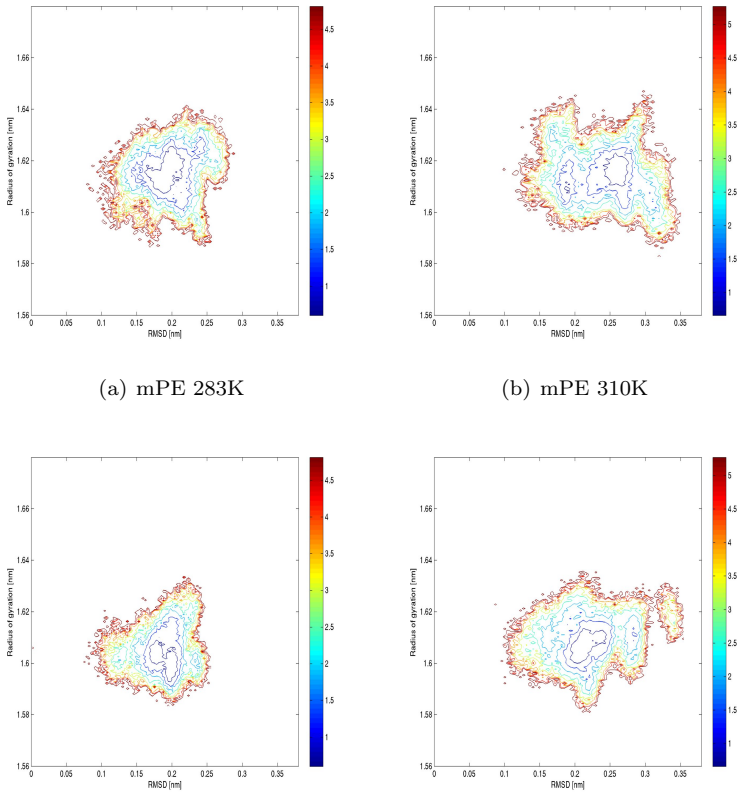
4.3.3 Mesophilic and psychrophilic elastases

When rmsd and Rg were used as collective coordinates, the psychrophilic elastase show a more compact FEL compared to the mesophilic homologous at 283K (figure 4.3a, 4.3c). At 310 K, the landscape of the mesophilic enzyme splits into two relative minima whereas the psychrophilic retains a single minimum (figure 4.3b, 4.3d).

If the cartesian principal components (figure 4.4) were adopted as reaction coordinates, the same trend of rmsd vs Rg coordinates can be highlighted: the mesophilic enzyme samples a wider conformational space at 283 K (figure 4.4a), and, more clearly, at 310K (figure 4.4b). Moreover it is possible to observe that the psychrophilic enzyme samples a greater number of relative minima at both temperatures (figures 4.4c, 4.4d). The distribution of the configurations into many minima implies that the surface is shallow but more rugged than the mesophilic surface.

This is clearly visible from the probability density plots 4.5, in which the mesophilic enzyme depicts few highly populated relative minima, whereas the psychrophilic surface is characterized by many few populated relative minima.

These findings are in agreement with the observation that the porcine elastase exhibits an higher global flexibility compared to the psychrophilic elastase, as indicated by previous works [9, 10]. In particular, the conformations sampled by the psychrophilic enzyme are very similar among each other. Global reaction coordinates (global RMSD and radius of gyration) are in fact unable to distinguish among different conformational sub-states (figures 4.3c, 4.3d). The near-native free energy landscape of the psychrophilic enzyme seems to be a narrow shallow basin whit a rugged bottom composed by many local minima



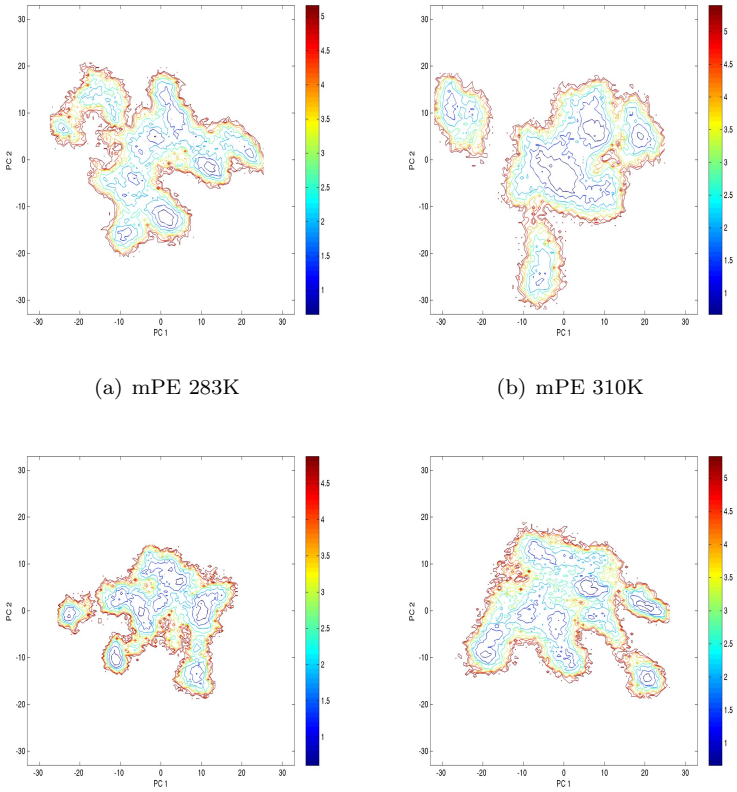
(a) mPE 283K

(b) mPE 310K

(c) pSE 283K

(d) pSE 310K

Figure 4.3: Contour plot representation of the free energy landscapes for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels). The reaction coordinates are the C-alpha RMSD and the radius of gyration. See Methods section for details.



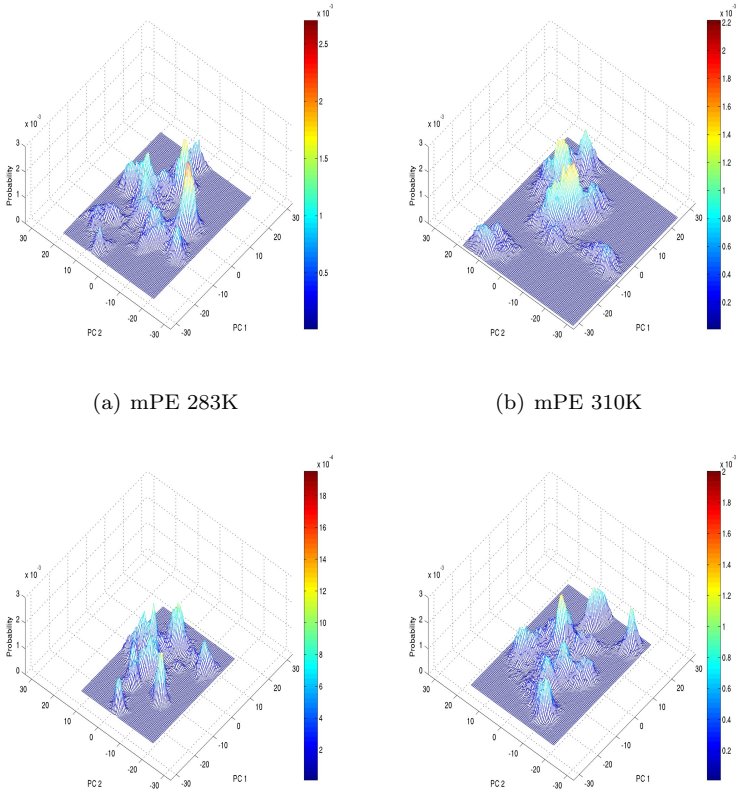
(a) mPE 283K

(b) mPE 310K

(c) pSE 283K

(d) pSE 310K

Figure 4.4: Contour plot representation of the free energy landscapes for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels). The free energy landscape is projected onto the first two principal components of the all-atoms mass-weighted covariance matrix of the combined trajectory.



(a) mPE 283K

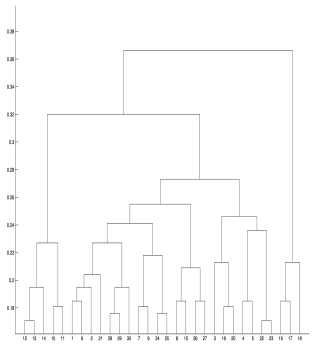
(b) mPE 310K

(c) pSE 283K

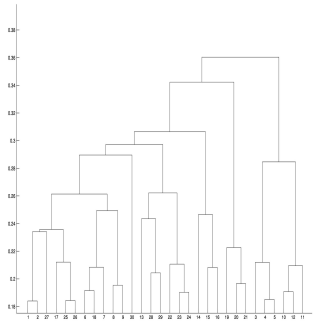
(d) pSE 310K

Figure 4.5: Representation of the 2D-probability plots for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels). The free energy landscape is projected onto the first two principal components of the all-atoms mass-weighted covariance matrix of the combined trajectory.

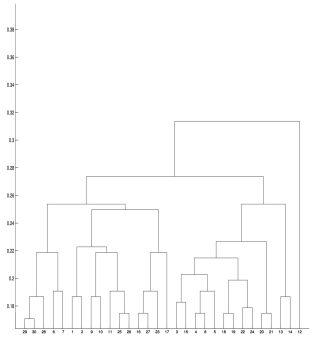
separated by low energy barriers. On the other hand, the mesophilic enzyme has a free energy landscape that shows a funnel-like shape where the conformations are organized in hierarchical fashion. These results are confirmed by cluster analysis: it is clearly visible the pronounced hierarchical structure of the clustering dendrogram of the mesophilic enzyme (figure 4.6a, 4.6b) compared to the flatter structure of the clustering dendrogram referred to the psychrophilic enzyme (figure 4.6c, 4.6d).



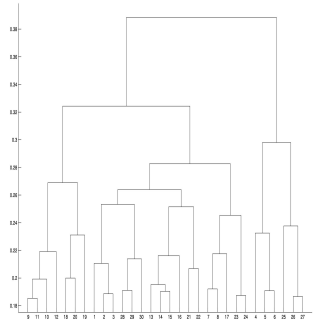
(a) mPE 283



(b) mPE 310



(c) pSE 283

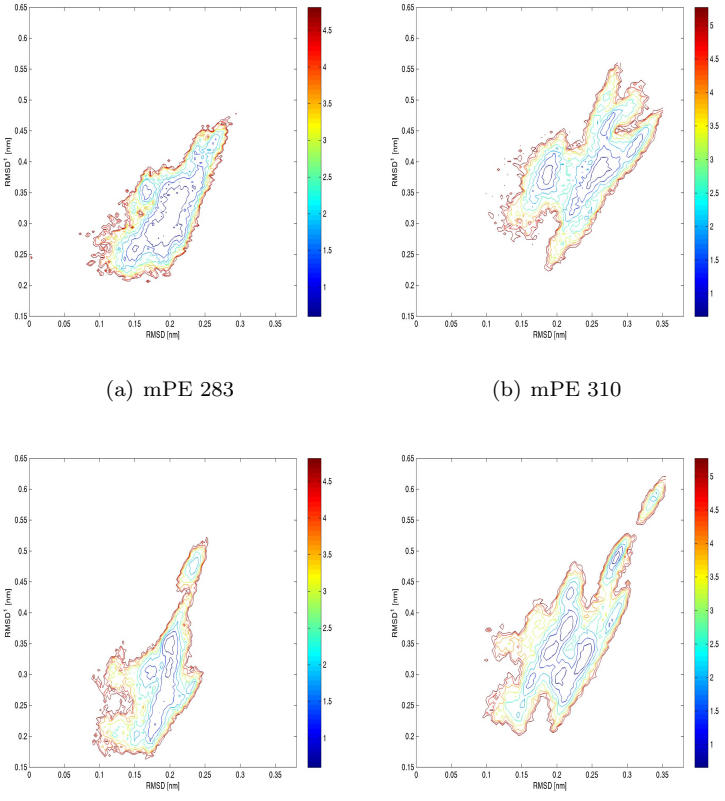


(d) pSE 310

Figure 4.6: Dendrograms of the complete linkage cluster analysis of the C-alpha RMSD distance matrix for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels) are shown.

Some portions of the structure of the psychrophilic enzyme are associated to a greater flexibility compared to the equivalent regions onto the mesophilic

protein [9]. Projecting the free energy landscape onto the rmsd of these regions (figure 4.7) it is possible to discriminate between different sub-states of the psychrophilic proteins which are not observable considering the global RMSD.



(a) mPE 283 (b) mPE 310

(c) pSE 283 (d) pSE 310

Figure 4.7: Contour plot representation of the free energy landscapes for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels). The reaction coordinates are the C-alpha RMSD and the local C-alpha RMSD, computed only on the higher flexibility residues. See Methods section for details.

The higher global flexibility of the porcine elastase can be quantitatively assessed estimating the configurational entropy.

System	n. eigenvalues	T [K]	Entropy (S) [J/Kmol]	S/(num.eigen.) [J/Kmol]
pSE	6777	283	48714.1	7.19
pSE	6777	310	53542.5	7.91
mPE	7047	283	52098.5	7.40
mPE	7047	310	57220.0	8.13
pCUDG	6891	283	51028.1	7.41
pCUDG	6891	310	55225.4	8.02
pHUDG	7002	283	52202.2	7.46
pHUDG	7002	310	57571.3	8.23

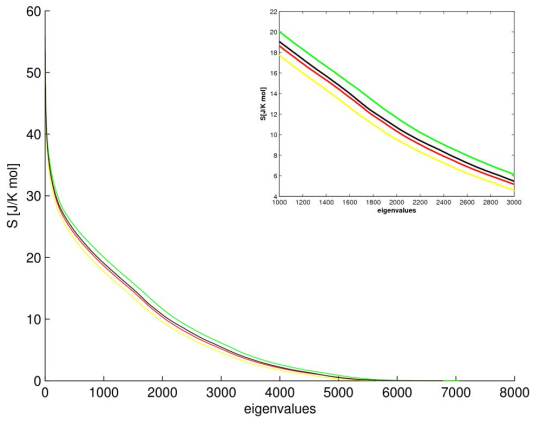
Table 4.2: Configurational entropy estimated using the Karplus [17] relationship. The last column represent the entropy normalized by the number of eigenvalues.

Configurational entropy

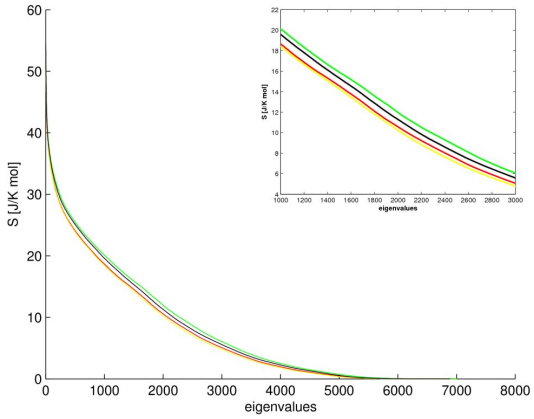
From the free energy landscapes and the conformational cluster analysis it is possible to conclude that the mesophilic enzyme (mPE) exhibits a higher conformational flexibility at both 283 and 310 K. The configurational entropy of the mesophilic enzyme has been estimated to quantify its higher flexibility. Plotting the entropy values for each eigenvalue it is possible to see that the first eigenvalues, associated to the highest entropy, are similar among the different systems (figure 4.8).

Also the last eigenvalues, to which correspond low entropy values are similar into all systems. The main differences are associated to the middle eigenvalues, in particular the curve of the mesophilic enzyme is always over the psychrophilic one (figure 4.8a). Detailed results, summarized in table 4.2, show that the entropy of the mPE enzyme is higher than the SE enzyme at both temperatures.

Because of the number of atoms of mPE and pSE enzyme are different, to compare their entropy values we have divided the total entropy for the number of eigenvalues (three folds the number of atoms). The entropy (per atom) differences between the mPE and pSE system are : 0.20 J/Kmol at 283 K and 0.22 J/Kmol at 310K. Moreover the entropy difference due to the increase of the temperature is lightly higher for the mesophilic enzyme= 0.73 J/Kmol compared to the psychrophilic enzyme= 0.71 J/Kmol. These small differences between the entropy values of the cold- and warm-adaped enzyme are due to the high similarities in the entropy values associated to the first eigenvalues, that have the highest weight.



(a) ELASTASE



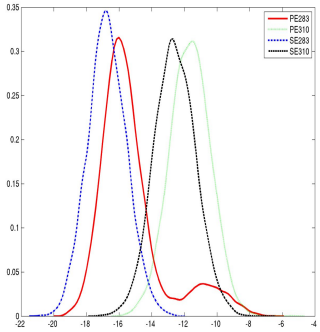
(b) UDG

Figure 4.8: Representation of the entropy per eigenvalues for the elastases (top panel) and for the UDG (bottom panel). Psychrophilics enzyme at 283K and at 310K is shown in yellow and in red, respectively. The mesophilic enzyme at 283K and at 310K is shown in black and in green, respectively. For clarity the central region, corresponding to highest entropy differences, is magnified.

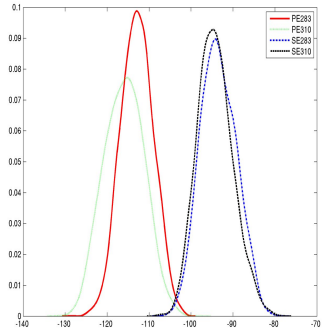
Analysis of the potential energy contributions

To easily understand the influence of the single contributions we grouped it into two intra-protein terms and two protein-solvent terms. The intra-protein terms include a packing energy and an electrostatic energy term. The packing energy is the sum of Lennard-Jones 1-4, Lennard-Jones short range, angle, proper dihedral and improper dihedral contributes. The intra-protein electrostatic term contain the Coulomb 1-4 interactions, the short-range and the long-range electrostatic interactions. The two protein-solvent terms are the Lennard-Jones term and the protein-solvent electrostatic contribute composed by the sum of the short-range and long-range protein-solvent Coulomb interactions. The average values of the packing energy are proportional to the temperature of the simulation and not to the protein, i.e. at 283K the mPE shows similar distribution to pSE at 283K and mPE at 310K show the same distribution of pSE at 310K (figure 4.9a). The intra-protein electrostatic energy show different behaviour into the two systems, in particular the mesophilic enzyme (mPE) has lower values of the electrostatic terms. There is no relationship between the intra-protein electrostatic and the temperature of the simulations.(figure 4.9b)

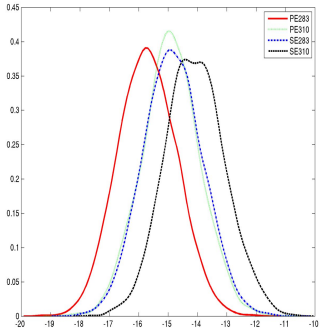
The protein-solvent Lennard-Jones term shows overlapping distributions, so is completely uninformative (figure 4.9c). The protein-solvent electrostatic contribute has partially overlapping distributions (figure 4.9d), the highest gap is between pSE at 238K and mPE at 310K. In particular the psychrophilic enzyme shows lower values compared to the mesophilic enzyme.



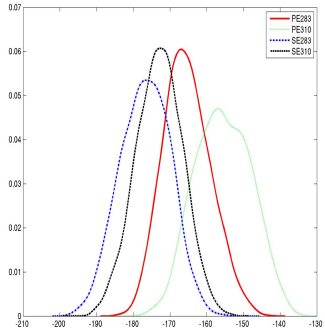
(a) Packing



(b) Electrostatics Protein-Protein



(c) Lennard-jones Protein-Solvent



(d) Electrostatics Protein-Solvent

Figure 4.9: Probability distribution of different potential energy contributions [cal/g] of elastases simulations. a. Packing energy: sum of Lennard-Jones 1-4, Lennard-Jones short range, angle, proper dihedral, and improper dihedral. b. Electrostatics Protein-Protein: Coulomb 1-4 interactions, the short-range and the long-range electrostatic interactions. c. Electrostatics Protein-Solvent: short-range and long-range protein-solvent Coulomb interactions. d. Lennard-Jones Protein-Solvent: Lennard-Jones protein-solvent interactions.

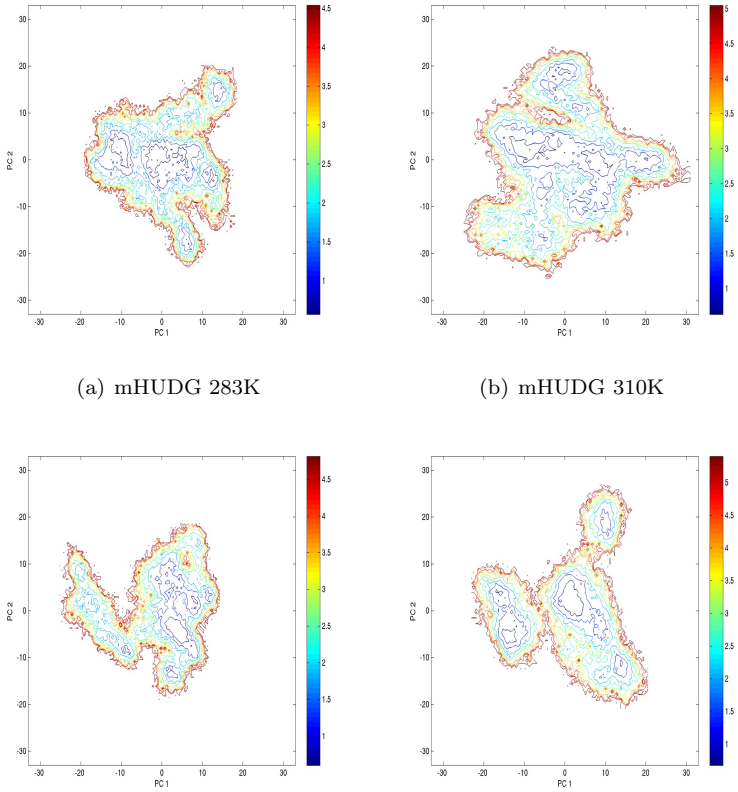
4.3.4 Uracil-DNA glycosylase (UDG)

Differently from the elastases the near-native free energy landscapes of the uracil-DNA glycosylases result more smoothed and flatter. Projecting the near-native free energy landscape onto the RMSD and radius of gyration, the landscape results into a single roughly smoothed well. Hence these global collective coordinates are unable to distinguish between different configurational sub-states and are then useless for the characterization of the shape of the near-native free energy landscape. Considering the projection onto the first two eigenvectors as collective coordinates, it is possible to distinguish different relative minima, the collective coordinates obtained from the PCA allow a finer separation of the different conformations of the protein (figure 4.10). The mesophilic enzyme show a shallow rugged surface at either 283 K and 310 K, with many local minima (figure 4.10 a,b). Moreover moving from 283 K to 310 K the conformations spread over a large range. On the contrary, the free energy landscapes of the psychrophilic enzyme is more compact and deeper (figure 4.10 9 c,d). In particular at higher temperature (310K) the conformations split into well defined higly populated clusters. Looking at the probability plots (figure 4.11), it is possible to see that the psychrophilic enzyme displays some highly populated regions (figure 4.11) which are non visible into the mesophilic enzyme (figure 4.11 a,b) resulting in a deeper free energy landscape (figure 4.11 c,d).

As in the elastases the free energy landscapes of the psychrophilic enzyme is more rugged, i.e. the conformations group into separated clusters. On the contrary, the main difference here, is that the landscape of mesophilic enzyme loses the accentuated funneled shape becoming shallow and rugged. This is clearly visible looking at the dendrograms obtained from the clustering of the conformations (figure 4.12). The shape of the dendrograms is quite similar for both the enzymes and it is not possible to identify a strong hierarchical distribution of the conformations.

Configurational entropy

Considering the configurational entropy the UDG enzymes show the same behaviour of the elastases, indeed, as shown in figure 10a, the curve of the mesophilic enzyme (mHUDG) is above the psychrophilic one (pCUDG) at both 283 K and 310K. In this case the difference of the entropy between the two systems at 283K is very low (0.05 J/K mol), whereas at 310 K is similar to the value of the elastase (0.21 J/K mol). Moreover, as for the elastases, the entropy difference due to the increase of the temperature is lightly higher for the mesophilic



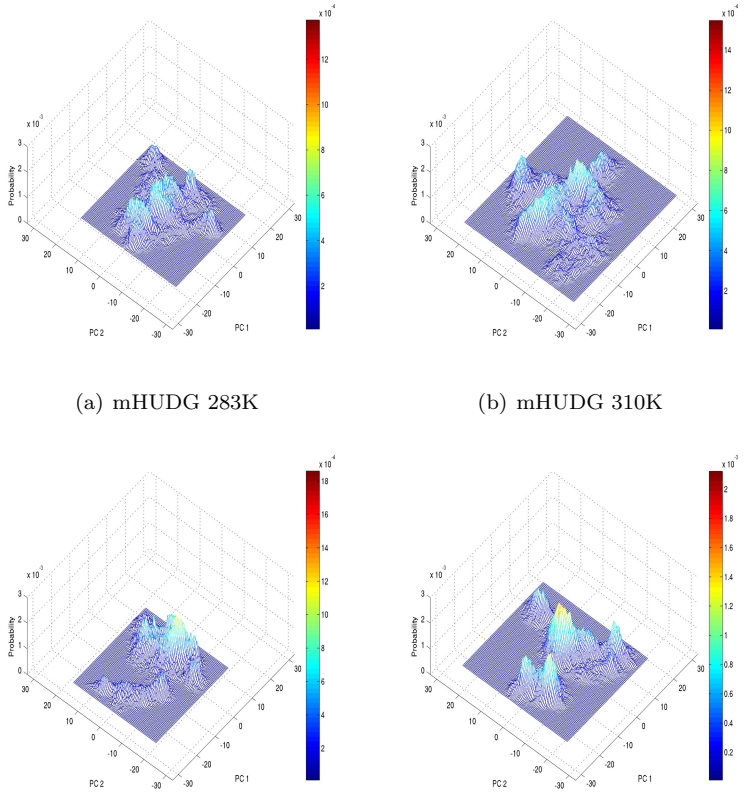
(a) mHUDG 283K

(b) mHUDG 310K

(c) pCUDG 283K

(d) pCUDG 310K

Figure 4.10: Contour plot representation of the free energy landscapes for the psychrophilic UDG (bottom panels), and for the mesophilic UDG (top panels) at 283K (left panels) and at 310K (right panels). The free energy landscape is projected onto the first two principal components of the all-atoms mass-weighted covariance matrix of the combined trajectory.



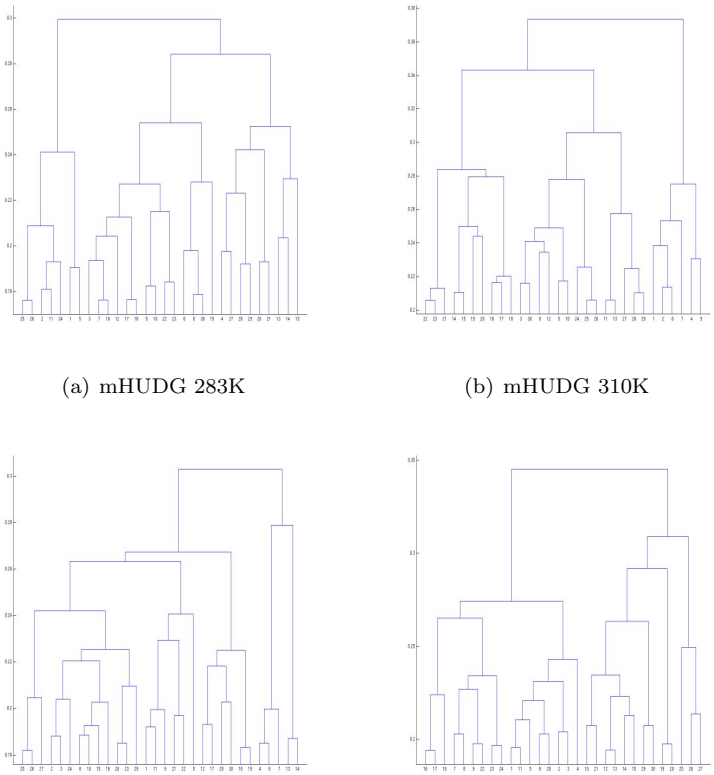
(a) mHUDG 283K

(b) mHUDG 310K

(c) pCUDG 283K

(d) pCUDG 310K

Figure 4.11: Representation of the 2D-probability plots for the psychrophilic UDG (bottom panels), and for the mesophilic UDG (top panels) at 283K (left panels) and at 310K (right panels). The free energy landscape is projected onto the first two principal components of the all-atoms mass-weighted covariance matrix of the combined trajectory.



(a) mHUDG 283K

(b) mHUDG 310K

(c) pCUDG 283K

(d) pCUDG 310K

Figure 4.12: Dendrograms of the complete linkage cluster analysis of the C-alpha RMSD distance matrix for the psychrophilic elastase (bottom panels), and for the mesophilic elastase (top panels) at 283K (left panels) and at 310K (right panels) are shown. For clarity the tree is truncated to a minimum cut-off value (CUTOFF).

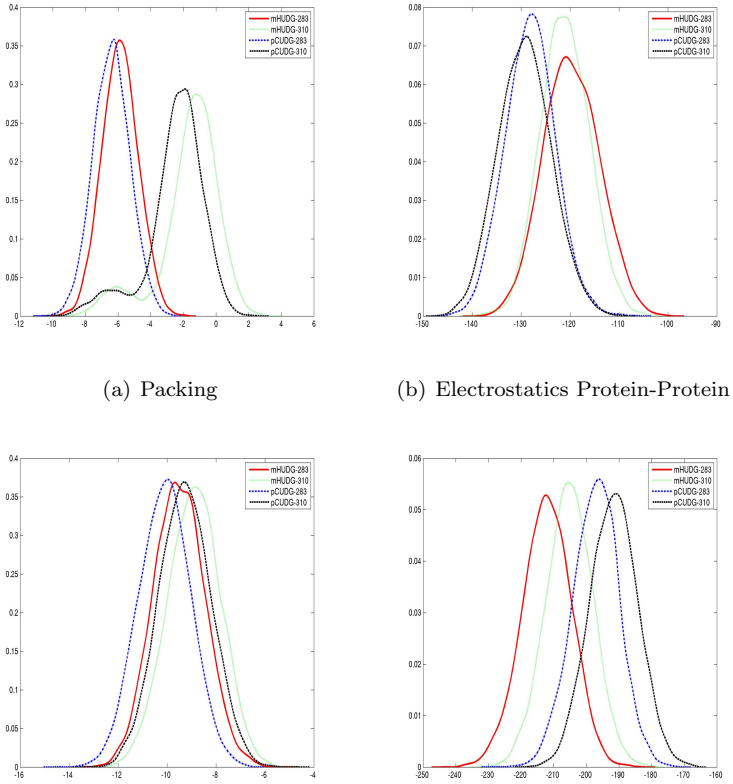
enzyme (0.76 J/K mol) compared to the psychrophilic enzyme (0.61 J/K mol). As depicted in figure 4.8b, the differences in entropic content are associated to the middle eigenvalues, whereas the first and the last eigenvalues show similar values as it happens for the elastases.

Analysis of the potential energy contributions

As well as the free energy surface and the entropy, the distributions of the potential energy terms reflect the high similarity between the psychrophilic and the mesophilic form of the UDG. Like the elastases the distributions of the packing energy are proportional to the temperature of the simulation and not to the different enzyme, i.e. at 283K the mHUDG shows similar distribution to pCUDG and at 310K mHUDG shows the same distribution of pCUDG (figure 4.13a). The intra-protein electrostatic energy shows different behaviour into the two systems, especially, contrary to that of elastases, i.e. the mesophilic enzyme (pCUDG) has lower values of the electrostatic terms (figure 4.13b). The protein-solvent Lennard-Jones term shows completely overlapping distributions, so it is completely uninformative as for the elastases (figure 4.13c). The protein-solvent electrostatic contribute has partially overlapping distributions (figure 4.13d), the highest gap is between mHUDG at 238K and pCUDG at 310K. Conversely to the behaviour of the elastases, the mesophilic enzyme shows lower values compared to the psychrophilic one. The averages and standard deviations of each energetic contributions are summarized in table 4.3.

System	$T[K]$	Ang^a	Dih^b	$iDih^c$	$E_{PP}^{SR\,d}$	$LJ_{PP}^{SR\,e}$	$E_{PP}^{LR\,f}$
mPE	283	33.78±1.26	12.50±0.57	11.13±0.54	-328.63±3.99	-89.53±0.85	-5.06±3.62
pPE	310	36.15±0.83	13.25±0.53	11.94±0.46	-331.59±4.84	-89.63±0.90	-5.29±3.73
mSE	283	34.29±0.80	12.84±0.51	10.97±0.45	-335.02±4.58	-90.88±0.88	1.19±3.26
mSE	310	36.61±0.87	13.51±0.53	11.87±0.47	-335.46±3.92	-90.57±0.89	0.33±3.46
mHUDG	283	34.37±0.79	15.35±0.55	10.80±0.42	-342.23±6.00	-72.24±0.72	-10.58±6.23
mHUDG	310	36.77±1.13	15.99±0.62	11.60±0.52	-344.41±6.24	-71.70±0.80	-9.45±6.60
pCUDG	283	34.02±0.78	15.15±0.53	10.73±0.42	-355.07±6.69	-72.36±0.78	-9.10±6.59
pCUDG	310	36.31±1.12	15.90±0.61	11.53±0.51	-355.28±6.76	-72.21±0.82	-9.49±6.76
System	$T[K]$	$E_{PP}^{14\,g}$	$LJ_{PP}^{14\,h}$	$E_{PS}^{LR\,i}$	$LJ_{PS}^{SR\,l}$	$E_{PS}^{SR\,m}$	
mPE	283	220.83±0.98	16.54±0.50	-155.46±5.66	-15.71±1.00	-10.14±3.35	
pPE	310	220.81±1.01	16.63±0.50	-146.74±6.23	-14.91±0.99	-8.88±3.40	
mSE	283	240.80±1.06	16.02±0.48	-161.16±6.01	-14.85±1.00	-15.26±3.25	
mSE	310	240.87±1.14	15.98±0.50	-157.43±5.01	-14.10±1.01	-15.51±3.39	
mHUDG	283	233.01±0.76	5.81±0.48	-200.08±6.77	-9.53±1.04	-11.95±3.90	
mHUDG	310	232.72±0.83	5.86±0.47	-193.68±6.89	-8.89±1.07	-11.65±3.87	
pCUDG	283	235.83±0.78	6.07±0.46	-189.63±6.65	-10.09±1.06	-7.57±3.30	
pCUDG	310	235.52±0.82	6.12±0.51	-184.10±7.02	-9.28±1.06	-7.43±3.37	

Table 4.3: Average values and standard deviation of the potential energy components [cal/g] of every systems. a. g96 angle, b. Proper dihedral, c. Improper dihedral, d. Intra-proteic Coulomb short-range, e. Intra-proteic Lennard-Jones short-range, f. Intra-proteic Coulomb long-range, g. Intra-proteic Coulomb 1-4, h. ntra-proteic Lennard-Jones 1-4, i. Protein-solvent Coulomb long-range, l. Protein-solvent Lennard-Jones short-range , m. Protein-solvent Coulomb short-range.



(a) Packing

(b) Electrostatics Protein-Protein

(c) Lennard-jones Protein-Solvent

(d) Electrostatics Protein-Solvent

Figure 4.13: Probability distribution of different potential energy contributions [cal/g] of UDG simulations. a. Packing energy: sum of Lennard-Jones 1-4, Lennard-Jones short range, angle, proper dihedral, and improper dihedral. b. Electrostatics Protein-Protein: Coulomb 1-4 interactions, the short-range and the long-range electrostatic interactions. c. Electrostatics Protein-Solvent: short-range and long-range protein-solvent Coulomb interactions. d. Lennard-Jones Protein-Solvent: Lennard-Jones protein-solvent interactions.

Conclusions

In this study a detailed comparative structural and energetic analysis of long molecular dynamic simulations of warm- and cold- adapted enzyme belonging to different families, have been carried out. This work has allowed to identify the structural and energetic features that characterize each enzyme and the common features shared by the enzyme of the two families. We would like to point out that, the critical changes for the thermal adaptation may be hidden by those produced by genetic drift and others effectors of natural selection [31], hence, it is not possible to assert with certainty, that the intra-family differences seen here between the mesophilic and psychophilic enzymes, are exclusively due as to the results of the cold adaptation . Up till this hypothesis holds, we can state that the two family of enzymes studied here, have developed partial distinct strategies to achieve the same goal, as already seen in many others cases [31]. Nevertheless, some properties show the same trend in the two families of enzymes and, moreover, they are in accord to the behaviour of other psychrophilic enzymes. Noteworthy, the cold adapted enzyme, of both the ELA and UDG, show a more rugged free energy landscape with separated energy basin, as can be clearly seen in the free energy landscapes projected onto the cartesian principal components (figure 4.4,4.10). This implies the existence of many metastable states, that cause the enzyme to assume for longer time not optimal conformations for the substrate binding which may result in higher K_m [33, 31] Moreover the higher flexibility, localized near the active site, can lead to rapid movement of the loop involved in the ligand binding and in turn augment the catalytic efficiency (higher values of k_{cat}) [33, 31]. Another common finding of the two families of enzyme is that the mesophilic counterpart shows a lightly higher global flexibility at both the temperatures, leading an increased entropic content. This result confirms what was already seen in previous works on these class of enzymes [9, 10, 11]. The main distinction in the two families of enzyme can be appreciated considering the potential energy components. Indeed, the intra-protein and the protein-solvent electrostatic contributions have an opposite behaviour in the two orthologous enzymes. Differences in the distributions of the electrostatic energy is supposed to be an important factor in thermal adaptation [31] This work allowed to shed light to some important structural, dynamical and energetic features of two different families of cold-adapted enzymes comparing them with its respective mesophilic counterpart. In particular we found that : i) the psychrophilic enzyme shows a rugged FEL with more meta-stable states, ii) confirming a previously founded result [9, 10, 11], the cold-adapted shows a lower global flexibility that is related to a lower configu-

rational entropy, and exhibits a larger flexibility localized on specific regions of the structure. iii) the two families show differences in the distributions of the electrostatics interactions which are related to their different behaviour.

Bibliography

[1] Amadei A, Ceruso M A, and Di Nola A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins*, 36:419–424, 1999.

[2] Fields P A and Somero G N. Hot spots in cold adaptation: Localized increases in conformational flexibility in lactate dehydrogenase a4 orthologs of antarctic notothenioid fishes. *PNAS*, 95:11476–11481, 1998.

[3] Hamprecht F A, Peter C, Daura X, Thiel W, and van Gunsteren W F. A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering and visualization. *J. Chem. Phys.*, 114(5), 2001.

[4] Leo-Maciasa A, Lopez-Romeroa P, Lupyanb D, Zerbinoa D, and Ortiz A R. Core deformations in protein families: a physical perspective. *Biophysical Chemistry*, 115:125–128, 2005.

[5] Hess B. Similarities between principal components of protein dynamics and random diffusion. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 62:8438–8448, 2000.

[6] Berendsen H J C, Postma J P M, Dinola A, and Haak J R. Md with coupling to an external bath. *J. Phys. Chem.*, 81:3684–3690, 1984.

[7] Gerday C, Aittaleb M, Arpigny J L, Baise E, Chessa J P, Garsoux G, Petrescu I, and Feller G. Psychrophilic enzymes: a thermodynamic challenge. *Biochim Biophys Acta*, 1342(2):119–131, 1997.

[8] Georlette D, Blaise V, Collins T, D'Amico S, Gratia E, Hoyoux A, Marx J C, Sonan G, Feller G, and Gerday C. Some like it cold: biocatalysis at low temperatures. *FEMS Microbiol Rev*, 28(1):25–42, 2004.

[9] Papaleo E, Riccardi L, Villa C, Fantucci P, and De Gioia L. Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family. *Biochim Biophys Acta*, 1764:1397–1406, 2006.

[10] Papaleo E, Olufsen M, De Gioia L, and Brandsdal B O. Optimization of electrostatics as a strategy for cold-adaptation: a case study of cold- and warm-active elastases. *J Mol Graph Model*, 26:93–103, 2007.

[11] Papaleo E, Pasi M, Riccardi L, Sambi I, Fantucci P, and De Gioia L. Protein flexibility in psychrophilic and mesophilic trypsins. evidence of evolutionary conservation of protein dynamics in trypsin-like serine-proteases. *FEBS Lett*, 582:1008–1018, 2008.

[12] Pontiggia F, Colombo G, Micheletti C, and Orland H. Anharmonicity and self-similarity of the free energy landscape of protein g. *Phys Rev Lett*, 98(4):048102, 2007.

[13] Feller G. Molecular adaptations to cold in psychrophilic enzymes. *Cell Mol Life Sci*, 60:648–662, 2003.

[14] Feller G and Gerday C. Psychrophilic enzymes: hot topics in cold adaptation. *Nat Rev Microbiol*, 1:200–208, 2003.

[15] Gianese G, Argos O, and Pascarella S. Structural adaptation of enzymes to low temperatures. *Protein Eng*, 14:141–148, 2001.

[16] B Hess, H Bekker, H J C Berendsen, and Fraaije J G E M. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

[17] Andricioaei I and Karplus M. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.*, 115:6289–6292, 2001.

[18] Tavernelli I, Cotesta S, and Di Iorio E E. Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *Biophysical Journal*, 85(4):2641–2649, 2003.

[19] Schlitter J. Estimation of absolute and relative entropies of macromolecules using the covariance-matrix. *Chem. Phys. Lett.*, 215:617–621, 1993.

[20] Ikeda K and Higo J. Free-energy landscape of a chameleon sequence in explicit water and its inherent alpha/beta bifacial property. *Protein Sci.*, 12(11):2542–2548, 2003.

[21] Miyazaki K, Wintrode P L, Grayling R A, Rubingh D N, and Arnold F H. Directed evolution study of temperature adaptation in a psychrophilic enzyme. *J Mol Biol*, 297:1015–1026, 2000.

[22] Hongxing Lei, Chun Wu, Haiguang Liu, and Yong Duan. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *PNAS*, 104(12):4925–4930, 2007.

[23] Becker O M. Principal coordinate maps of molecular potential energy surfaces. *Journal of computational chemistry*, 19 (11):1255–1267, 1998.

[24] Olufsen M, Smalas Arne O, Moe E, and Brandsdal B O. Increased flexibility as a strategy for cold adaptation: a comparative molecular dynamics study of cold- and warm-active uracil dna glycosylase. *J Biol Chem*, 280:18042–18048, 2005.

[25] Olufsen M, Arne O S, and Bjorn O B. Electrostatic interactions play an essential role in dna repair and cold-adaptation of uracil dna glycosylase. *J Mol Model*, 14:201–213, 2008.

[26] Gia G Maisuradze and David M Leitner. Free energy landscape of a biomolecule in dihedral principal component space: sampling convergence and correspondence between structures and minima. *Proteins*, 67(3):569–578, 2007.

[27] Kamiya N, Mitomo D, Shea J-E, and Higo J. Folding of the 25 residue abeta(12-36) peptide in tfe/water: temperature-dependent transition from a funneled free-energy landscape to a rugged one. *J. Phys. Chem. B*, 111(19):5351–5356, 2007.

[28] Smalas A O, Leiros H K, Os V, and Willassen N P. Cold adapted enzymes. *Biotechnol Annu Rev*, 6:1–57, 2000.

[29] D'Amico S, Claverie P, Collins T, Daphne G, Gratia E, Hoyoux A, Meuwis M-A, Feller G, and Gerday C. Molecular basis of cold adaptation. *Philos Trans R Soc Lond B Biol Sci*, 357(1423):917–925, 2002.

[30] D'Amico S, Collins T, Marx J C, Feller G, and Gerday C. Psychrophilic microorganisms: challenges for life. *EMBO Rep*, 7:385–389, 2006.

[31] Khawar S S and Cavicchioli R. Cold-adapted enzymes. *Annu Rev Biochem*, 75:403–433, 2006.

[32] Siddiqui K S, Poljak A, Guilhaus M, De Francisci D, Paul M G Curmi, Feller G, D'Amico S, Gerday C, Uversky V N, and Cavicchioli R. Role of lysine versus arginine in enzyme cold-adaptation: modifying lysine to homo-arginine stabilizes the cold-adapted alpha-amylase from pseudoalteramonas haloplanktis. *Proteins*, 64:486–501, 2006.

[33] George N Somero. Adaptation of enzymes to temperature: searching for basic "strategies". *Comp Biochem Physiol B Biochem Mol Biol*, 139(3):321–333, 2004.

[34] Essman U, Perela L, Berkowitz M L, Darden T, LeeH, and Pederson L G. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577–8592, 1995.

Chapter 5

Structural analysis of mutations

A curious aspect of the theory of evolution is that everybody thinks he understands it.
Jacques Monod (1910 - 1979)

5.1 Introduction

THe analysis of the protein structure can be useful when we are dealing with diseases which are dependent on mutations in a given protein. It help to identify how the mutation impairs the protein function and which is its impact on the disease. As examples, we report here three cases in which a structural study has been used to support biochemical and genetical data for the analysis of the impact of point mutations on the protein structure and function and its effect on the associated disease.
In particular, we have studied three different serious rare diseases which involve grave metabolic disorder associated to point mutations in mithocondrial proteins. In the following sections a short introduction of each case of study is given.

5.2 Ethylmalonic encephalopathy

Ethylmalonic encephalopathy (EE) (OMIM 602473) is an autosomal recessive disorder originally reported in Italian families and predominantly affecting children of Mediterranean or Arab descent. EE is characterised by psychomotor regression and generalised hypotonia, later evolving into spastic tetraparesis, dystonia, and eventually global neurological failure [2]. Magnetic Resonance Imaging (MRI) examination shows the presence of symmetrical and asymmetrical "patchy" lesions, distributed in the deep grey structures of the brain, including the brainstem, thalamus, and corpus striatum. The encephalopathy is typically accompanied by widespread lesions of the small blood vessels, causing showers of petechiae, especially during intercurrent infections, easy bruising, and orthostatic acrocyanosis. Chronic diarrhoea is another prominent feature of EE. The course is relentlessly progressive and usually leads to death within the first decade of life.
From a biochemical point of view, EE is characterised by persistent lactic acidaemia, a reduction in the activity of mitochondrial respiratory complex IV in skeletal muscle, and markedly elevated excretion of ethylmalonic and methylsuccinic acid in urine [25]. Ethylmalonic acid is believed to derive from the carboxylation of butyryl-coenzyme A (CoA), as a consequence of disorders of the mitochondrial b-oxidation of fatty acids, or from 2-ethylmalonic-semi-aldehyde, as a consequence of the catabolism of isoleucine [25].

5.2.1 ETHE1

The ETHE1 gene was identified as the responsible of the Ethylmalonic encephalopathy [26]. The name of this gene, previously known as HSCO (for hepatoma subtracted clone one), for its role in EE, has been changed to ETHE1. The product of this gene (Ethe1p) localized inside the mitochondrial matrix, and in particular a canonical mitochondrial leader peptide present at the N-terminus of the full-length ETHE1 protein, addresses the protein to the organelle and is cleaved off after internalization in the inner mitochondrial compartment through an energy-dependent process, presumably carried out by MPP [26].

The function of the Ethe1p mature protein is presently unknown. The Ethe1p is a phylogenetically conserved protein, sharing high homology with human Glyoxalase-II (Glyo-II) (β-lactamase fold). Besides Glyo-II, a BLAST search, with the human ethe1 predicted protein sequence as a probe, resulted in the identification of highly similar proteins in all metazoan species, in plants,

such as *Arabidopsis thaliana*, and in fungi, such as Saccharomyces cerevisiae 5.1. In contrast with the remaining portion of these protein sequences, the first 20-30 amino acid residues on the N-terminus appear to be poorly conserved.

The glyoxalase system catalyzes the conversion of toxic 2-oxoaldehydes into the corresponding 2-hydroxy acids. The main substrate seems to be methylglyoxal, which is formed as a by-product of glycolysis from triose phosphates through the action of triose-phosphate isomerase but also via other metabolic routes. As the first step of the glyoxalase system, methylglyoxal reacts spontaneously with reduced glutathione to form a hemithioacetal. Glyoxalase I converts hemithioacetal into S-D-lactoylglutathione, which is further metabolized to D-lactate and glutathione by Glyo-II [26]. Despite the similarity between the Ethe1p and the Glyo-II, Ethe1p failed to demonstrate a significant Glyo-II activity in isolated mitochondria, using D-lactoylglutathione as a substrate [26]. A likely possibility is that the ETHE1 protein could still be a mitochondrial metal-β-lactamase involved in the metabolism of an unknown substrate.

It has been recently suggested that subtle differences in the metal binding ligands of proteins characterized by the β-lactamase fold may be responsible for differences in metal binding properties among the different enzymes [19]. Therefore, enzymes featuring the metal-β-lactamase fold can bind several different metals and catalyze a broad number of different reactions. It has been proposed on the basis of crystal structure [4], NMR analysis [9] and EPR spectra [9] that the Glyo-II family has broad metal binding specificity and its members are suggested to accommodate mixed metal centers: Zn(II)-Zn(II), Fe(III)-Zn(II) or Fe(III)-Fe(II).
In a recent study a spectroscopic investigation of the crystal structure of Ethe1p from *Arabidopsis thaliana* demonstrates that Ethe1p binds one iron in a predominatly Fe(II) oxidation state [13]. This evidence suggests that the human Ethe1p is a novel, mononuclear Fe(II)-containing member of the β-lactamase fold superfamily. Anyway it should be noted that the sequence identity between the human and the *Arabidopsis* ethe1 is of 54%, hence the consideration on the *Arabidobsis* ehte1p may not completely transferred to the human one.

5.2.2 Genetic analysis of patients affected by EE

From genetic analysis many different mutations have been identified into the ETHE1 gene of the patient affected by EE. Tables 5.1 and 5.2 report all mutations so far described in the literature.

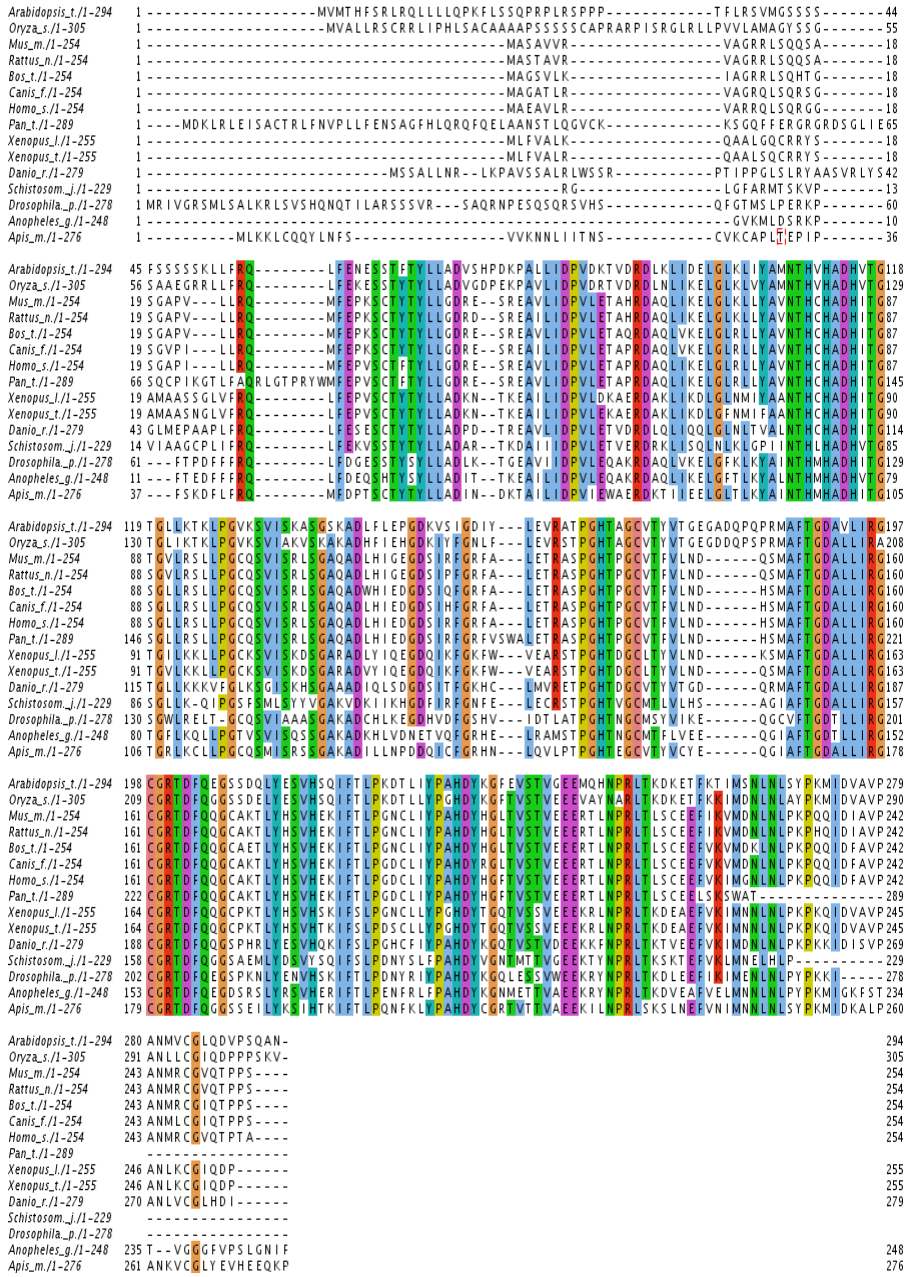


Figure 5.1: Multiple alignments of the ETHE1 protein sequences in different species. Protein sequences were search using BLAST, filtered with an identity threshold of 40% (only the proteins which share more than 40% of sequence identity with Ethe1p were kept) and aligned using CulstalW.

Gene mutation	Protein	Reference
(c.586GRA) + (c.586GRA)	**(p.D196N)+(p.D196N)**	[17]
(c.222_223insA) + (c.491CRA)	**(p.Y74fsX97)+(p.T164K)**	[17]
(c.164CRT)+(c.164CRT)	**(p.L55P)+(p.L55P)**	[17]
(c.487CRT)+(c.455CRT)	**(p.R163W)+(p.T152I)**	[17]
(g.del ex 4-7)+(g.del ex 4-7)	Not translated	[17]
(c.488GRA)+(c.488GRA)	**(p.R163Q)+(p.R163Q)**	[17]
(c.406ARG)+(c.488GRA)	**(p.T136A)+(p.R163Q)**	[17]
(c.505+1GRT)+(c.505+1GRT)	Splice exon-intron 4	[17]
(c.505+1GRT)+(c.505+1GRT)	Splice exon-intron 4	[17]
(g.del ex 4)+(g.del ex 4)	Not translated	[17]
(c.505+1GRT)+(c.505+1GRT)	Splice exon-intron 4	[17]
(g.del ex 4)+(g.del ex 4)	Not translated	[17]
(c.554TRG)+(g.del ex 4)	(p.L185R)/Not translated	[17]
(c.505+1GRT)+(c.505+1GRT)	Splice exon-intron 4	[17]
(c.487CRG)+(c.487CRG)	**(p.R163G)+(p.R163G)**	[8]
(c.delC66)+(c.delC66)	**(p.P22fsX32)+(p.P22fsX32)**	[3]
(c.487CRT)+(c.487CRT)	**(p.R163W)+(p.R163W)**	[12]
(g.-83delCGCCC)+(c.376-1GRT)	Not translated	[25]
(g.del ex 4)+(g.del ex 4)	Not translated	[25]
(g.del ex 4)+(g.del ex 4)	Not translated	[25]
(c.3GRT)+(c.3GRT)	**(p.M1I)+(p.M1I)**	[25]
(c.488GRA)+(c.488GRA)	**(p,R163Q)+(p.R163Q)**	[25]
(c.187CRT)+(c.482GRA)	**(p.Q63X)+(p.C161Y)**	[25]
(c.230delA)+(c.230delA)	(p.N77fsX144)+(p.N77fsX144)	[25]
(g.del ex 1-7)+(g.del ex 1-7)	Not translated	[25]
(c.406ARG)+(c.488GRA)	**(p.T136A)+(p.R163Q)**	[25]
(c.34CRT)+(c.34CRT)	**(p.Q12X)+(p.Q12X)**	[25]
(c.375+5GRA)+(c.375+5GRA)	Splice exon-intron 3	[25]
(c.487CRT)+(c.487CRT)	**(p.R163W)+(p.R163W)**	[25]

Table 5.1: List of Ethe1p mutations described in literature. The missense mutations are written in bold font.

Gene mutation	Protein	Reference
(c.604_605insG)+(c.604_605insG)	(p.V202fsX220)+(p.V202fsX220)	[26]
(c.3GRT)+(c.3GRT)	**(p.M1I)+(p.M1I)**	[26]
(g.del ex 4)+(g.del ex 4)	Not translated a	[26]
(c.487CRT)+(c.487CRT)	**(p.R163W)+(p.R163W)**	[26]
(c.406ARG)+(c.406ARG)	**(p.T136A)+(p.T136A)**	[26]
(c.222_223insA)+(440_450del11)	(p.Y74fsX97)+(p.H147fsX176)	[26]
(c.222_223insA)+(c.222_223insA)	(p.Y74fsX97)+(p.Y74fsX97)	[26]
(g.del ex 1-7)+(g.del ex 1-7)	Not translated	[26]
(g.del ex 4)+(g.del ex 4)	Not translated	[26]
(c.505+1GRT)+(c.505+1GRT)	Splice exon-intron 4	[26]
(c.375+5GRA)+(c.375+5GRA)	Splice exon-intron 3	[26]
(c.131_132delAG)+(c.488GRA)	(p.E44fsX102)+(p.R163Q)	[26]
(c.592_593insC)+(c.592_593insC)	(p.H198fsX220)+(p.H198fsX220)	[26]
(c.487CRT)+(c.487CRT)	**(p.R163W)+(p.R163W)**	[26]
(c.113ARG)+(c.554TRG)	**(p.Y38C)+(p.L185R)**	[26]
(c.487CRT)+(c.487CRT)	**(p.R163W)+(p.R163W)**	[26]
(c.505+1GRA)+(c.505+1GRA)	Splice exon-intron 4	[26]

Table 5.2: List of Ethe1p mutations described in literature. The missense mutations are written in bold font.

Template	RMSD	TM-score	LG-score	Overall
1qh5	8.72	0.61	0.27	GOOD
2gcu	7.25	0.68	0.28	GOOD

Table 5.3: AIDE comparison between two models of Ethe1p.

The missense mutations (highlighted on tables 5.1 and 5.2 in bold font) can be analyzed at molecular level studying the three-dimensional structure of the Ethe1p, and integrating the analysis with the available biochemical data.
Given that the experimental 3D structure of Ethe1p is not known, it is worth to obtain an accurate 3D model structure of it using computational methods.

5.2.3 3D model of Ethe1p

The three-dimensional model of Ethe1p was build by homology modelling using the crystal structure of the Ethe1p of *A. thaliana* (PDB ID 2gcu) as template. The sequence alignment was produced using ClustalW, the sequence identity between the two proteins is 54%. The model was build with modeller and refined by means of molecular dynamics and using the built in fast refinement procedure of modeller.
It table 5.3 this model was compared using AIDE [15] with a previously build model based on the human Glyo-II crystal structure [25]. The two models are similar in all structures but in the C-terminal region where the 2gcu based shows a better agreement with the predicted secondary structure of Ethe1p (data not shown).
The model based on 2gcu was then used for the structural analysis (figure 5.2).

The three-dimensional model of Ethe1 is consistent with the β-lactamase fold of the N-terminal domain and the presence of three α-helices in the C-terminal domain. The charged residues in Ethe1 are located prevalently on the protein surface and are well exposed to the solvent, as deduced from a surface accessibility analysis (data not shown). The only charged residues which are buried in the protein interior are negative residues involved directly or indirectly in the metal ion binding or in stabilizing the interactions between the β-harpin and the second domain. Based on multiple alignment results (figure 5.1) and on previous work [25] the six putative metal ion ligands are predicted to be H79, H81 and H135 (first metal ion, M1), and D83, H84 and H195 (second metal ion, M2) (figure 5.3). These residues are conserved in Ethe1 family and corresponded to identical residues in the Glyo-II family. As previously mentioned, unlike the

Glyo-II and other metal β-lactamases, only one metal ion is bound in the metal binding site of Ethe1p [13].

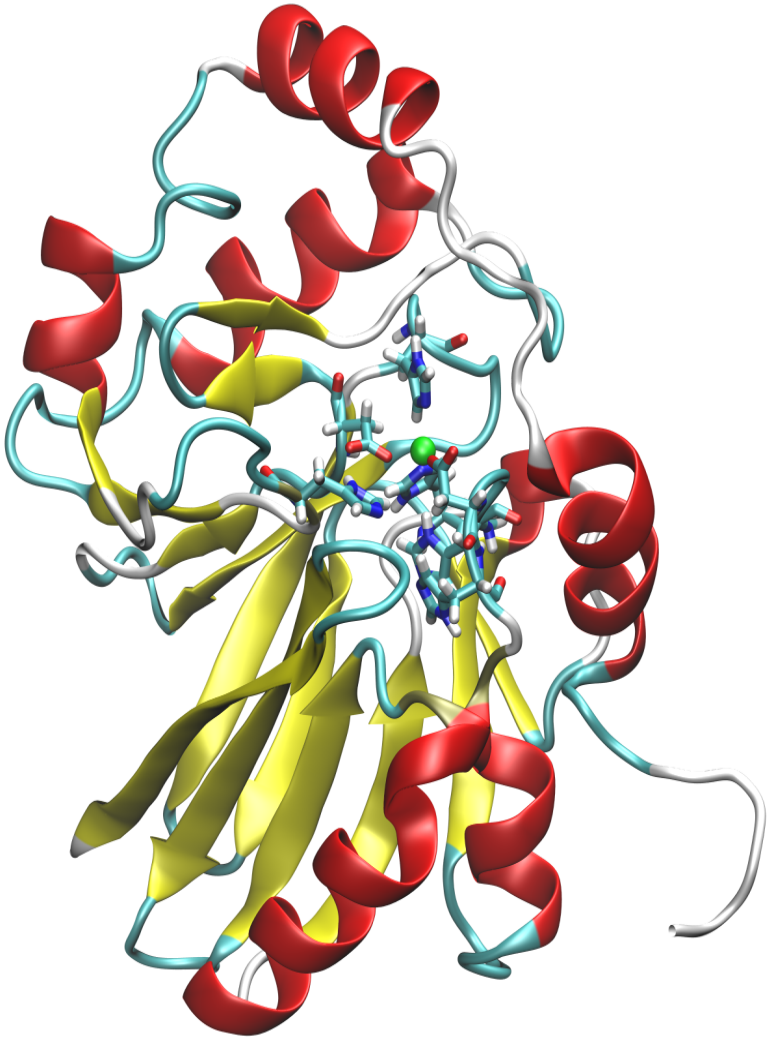


Figure 5.2: Three-dimensional model of Ethe1p build by homology modelling using the crystal structure the Ethe1p of *A. thaliana* (PDB ID 2gcu) as template.

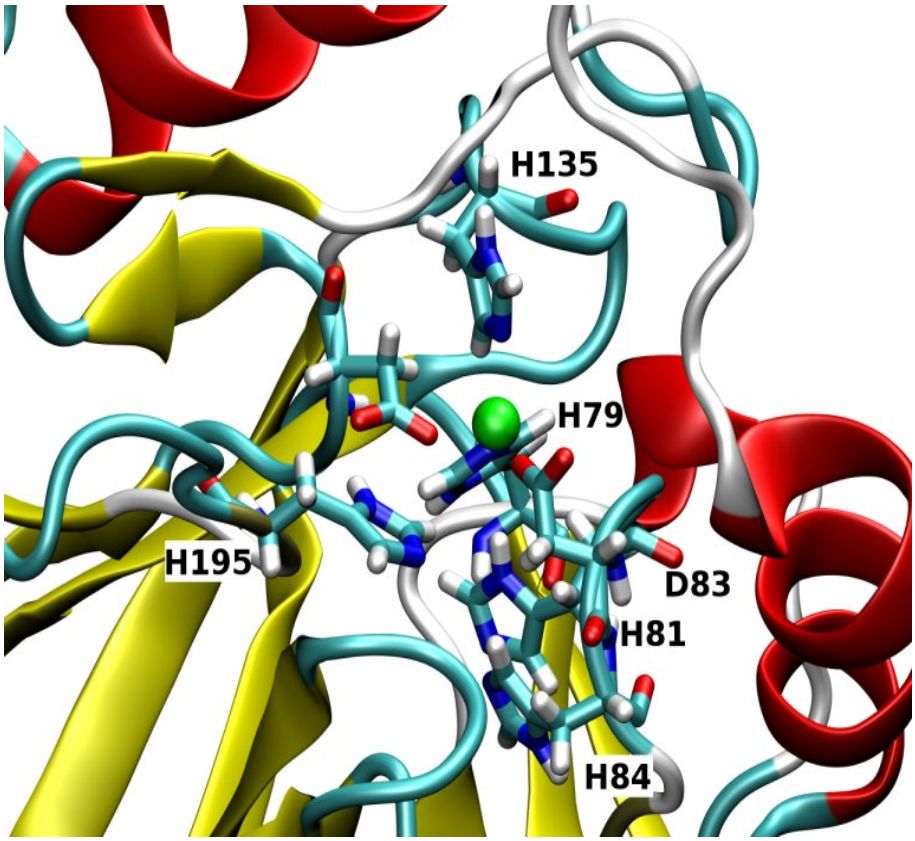


Figure 5.3: Ethe1p model catalytic site. The position of the Fe(II) metal ion (shown as a green sphere), is based on the position on the crystal structure of the *Arabidopsis* Ethe1p. The residues coordinating the metal ion and involved into the active site are also shown.

5.2.4 Biochemical and structural analysis

As shown in figure 5.1 , the missense mutations L55P, T152I, T164K, D196N, T136A, Y38C, L185R, C161Y, R163W and R163Q all affect highly conserved amino acid residues. The mutations in the Ethe1p protein can be classified as structural or catalytic depending on the presence/absence of the protein verified by western-blot analysis [25].

The T136A, Y38C, and L185R changes are structural mutations associated with the absence of Ethe1p protein specific cross reacting material (CRM). By contrast, the R163W, R163Q, and C161Y changes are catalytic mutations associated with the maintenance of normal or slightly reduced amounts of Ethe1p protein specific CRM [25]. The position of each missense mutation were analyzed in the three-dimensional (3D) model to make predictions of their possible functional consequences. T164 resides immediately adjacent to R163, an amino acid that is frequently mutated in patients with EE [11]. However, mutations affecting the R163 residue are associated with normal levels of Ethe1p protein, whereas the novel T164K mutation is associated with very low protein levels as shown by western-blot analysis [25]. Based on a 3D model and on the crystal structure of the Ethe1p protein in *A. thaliana* [13], this result can be explained by the localisation of this amino acid residue in a highly hydrophobic domain. The substitution of T by K can produce a distortion of the region due to steric hindrance or electric charge because K is bigger and more basic than T (figure 5.4). As a consequence, both the activity of the catalytic site and stability of the protein may be partially impaired. In contrast, the L55P occurs in a hydrophilic region of the protein, making it unlikely that the mutation can severely perturb the surrounding environment. However, proline is characterised by a peculiar conformational rigidity because of its cyclic structure, which locks the backbone dihedral angle at approximately -75°. This may lead to a distortion of the loop in which it is located, therefore affecting the protein tertiary structure. Moreover, mutant P55 is in close proximity to and may distort the orientation of H84, a residue involved in the catalytic site (figure 5.4). Our results show that both T164 and L55 are critical for the stability of the protein.

Mutation D196N is likely to affect the (unknown) catalytic activity of Ethe1p, because it is associated with normal protein levels [25]. This amino acid residue is located in the internal part of a loop and its carboxyl forms a hydrogen bond by interacting with the N-terminus of either F200 or H198 (figure 5.4, panel B). It is probable that the D196N mutation alters the conformation of the loop, thus interfering in an indirect way in substrate recognition and catalysis. The patient carrying this mutation apparently showed a milder phenotype than that

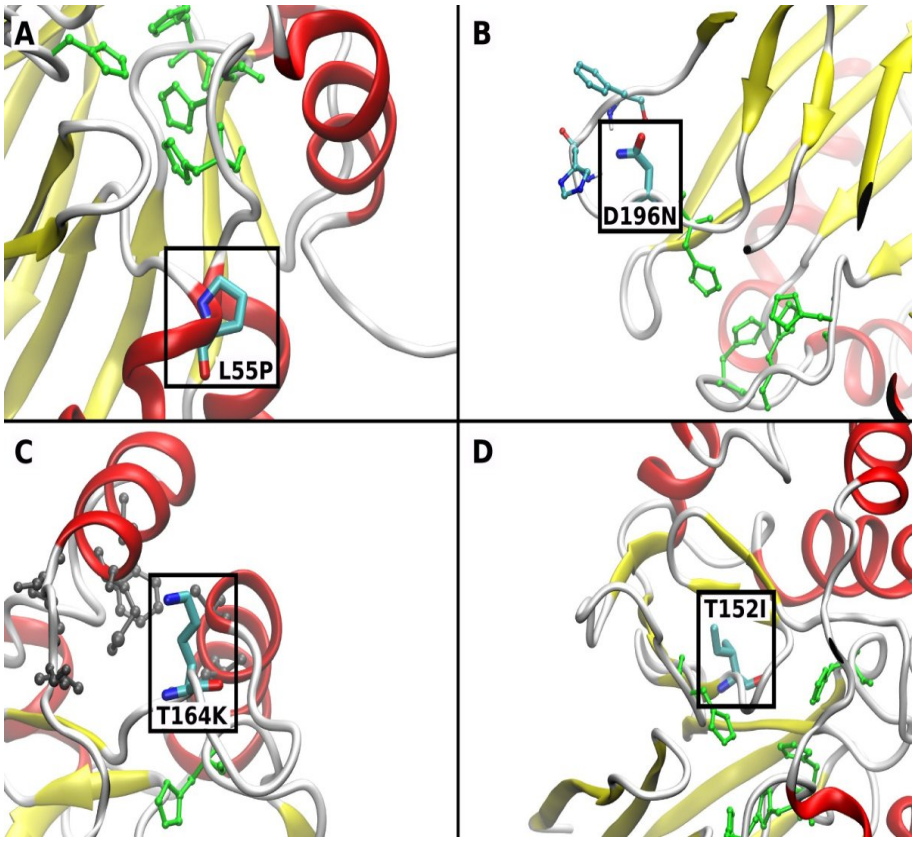


Figure 5.4: Localization of each mutant residue. A. The L55 residue is located in a hydrophilic region of the protein in relative proximity to the catalytic histidine cluste.r B. The D196 residue is likely to establish hydrogen bonds with residues F200 and H198. C. The hydrophobic amino acids surrounding the T164 residue are shown in grey. H135 (green) is part of the histidine catalytic cluster. D. The T152 residue is located within the catalytic pocket together with the histidine cluster.

seen in other patients presenting missense mutations associated with normal levels of the Ethe1p protein. Indeed, the patient with this mutation has a more benign course of the disease.
The T152 residue lies in a deeply buried position surrounded by hydrophobic tightly packed amino acids. Being isoleucine is bigger than threonine, the T152I substitution can result in a conformational rearrangement of the surrounding pocket. In particular, owing to the proximity of this mutation to the active site, the rearrangement may lead to a distortion in the orientation of the catalytic residues and in turn to loss of activity (figure 5.4).

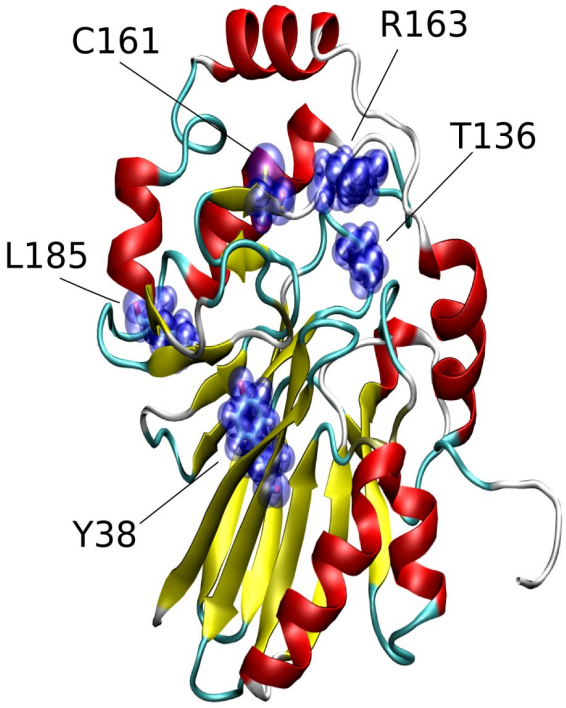


Figure 5.5: Localization of the following mutant residues on the 3D structure of the human Ethe1p model: T136A, Y38C, L185R, C161Y, R163W. The R163 is located near the active site, hence a mutation directly interferes with the enzyme catalysis. The other aminoacids are involved in the tertiary structural stabilization.

The T136 residue, which is strictly conserved in both Ethe1p and Glyo-II-like proteins (data not shown), belongs to the second coordination sphere of the metal ions , in close proximity to the active site (figure 5.5). The absence of Ethe1p protein specific CRM in T136A mutant fibroblasts indicates that the T136 residue is crucial in maintaining the structural integrity of the protein. The same is likely to occur for the Y38C and L185R mutations. In particular, the Y38 residue is located at the interface between the first and the second half of the first domain, where it is involved in the formation of a cluster of hydrophobic residues in the protein core. The L185 residue belongs to the first domain, in close proximity to two positively charged amino acids, K181 and R209. As a consequence, the effect of the L185R mutation, which changes an apolar leucine to a positively charged arginine, could be due to unfavourable electrostatic interactions between the three positively charged residues, leading to destabilisation and degradation of the protein.
The R163 residue in Ethe1p is located in a loop region, which is part of the putative catalytic site, near the binding site of Fe ions (figure 5.5). This observation suggests that R163Q and R163W mutations impair the catalytic activity, rather than the structural integrity, of Ethe1p.
The C161 residue, which does not interact with the metal binding site (fig 5.5), belongs to a GCG motif, which is conserved in both the Ethe1p-like and the Glyo-II-like protein families [17]. The GCG motif present in Ethe1p can play a key structural role in the stabilisation of the protein. This consideration can explain the low amount of C161Y Ethe1p protein specific CRM detected by western-blot analysis [25].
A computational and experimental approach, combining information from secondary structure predictions, multiple sequence alignments and comparative predictions allowed us to obtain a reliable structural model for Ethe1p. From the analysis of the model it was possible to define the Ethe1p functional sites, as well as clarify the effect of the mutations on the structure and the functionality of the protein.

5.3 Disease-associated mutation in COX6B1

Cytochrome c oxidase (COX, complex IV) is the terminal enzyme in the mitochondrial respiratory chain. It catalyzes the electron transfer from reduced cytochrome c to molecular oxygen.
This reaction is coupled to the extrusion of protons from the mitochondrial matrix to the intermembrane space, forming a proton-based membrane potential that allows ATP to be synthesized.
In mammals, the COX monomer is composed of 13 subunits, but the active form of the enzyme works as a dimer in vivo. Mitochondrial DNA (mtDNA) encodes the three larger, and most hydrophobic, subunits, including the two catalytic MT-CO1 and MT-CO2 subunits, as well as the core structural MT-CO3 subunit. The remaining ten smaller subunits, COX4, 5A, B, 6A, B, C, 7A, B, C, and 8, some of which have also tissue-specific isoforms, are encoded in the nucleus and imported into mitochondria [7]. Most of the nuclear encoded subunits of COX have hydrophobic domains spanning the inner mitochondrial membrane once [22]. However, subunits COX5A, COX5B, and COX6B are hydrophilic extramembrane proteins, the first two facing the matrix, whereas COX6B faces the intermembrane space. According to the bovine enzyme structure, COX subunit 6B, connects the two COX monomers into the physiological dimeric form [22, 20] and is also believed to interact with cytochrome c [24, 14].
COX deficiency (OMIM 220110) is one of the most common respiratory-chain defects in humans, being associated with different clinical phenotypes and caused by different genetic abnormalities. The study of COX defects is complicated as the biosynthesis and function of the enzyme depends on the contribution of both mitochondrial and nuclear encoded products.
Mutations in mtDNA-encoded COX genes (OMIM 516030, 516040, and 516050) are associated with a range of phenotypes including pure myopathy, MELAS (OMIM 540000), encephalomyopathy, and a motor neuron disease-like presentation [1]. In other cases, COX deficiency is associated with mutations in nuclear-encoded proteins that do not belong to, but participate in the biogenesis of, complex IV [16].
Mutations in nuclear-encoded COX structural subunits were searched for but never found [18, 10, 23]. Studying two siblings belonging to a consanguineous Arab family, affected by a combination of early-onset leukodystrophic encephalopathy, myopathy, and growth retardation, a COX deficiency of unknown cause was identified.

5.3.1 Genetic analysis

Sequencing the COX genes locus revealed the presence of a mutation in the COX6B1 gene.
In particular the homozygous substitution 221G→A in exon 2 of the COX6B1 gene was identified, leading to the missense mutation R19H (the numeration of the amino acid residues corresponds to the mature COX6B1 as the first M is cleaved in the import process of the precursor into mitochondria) [23].
A multiple sequence alignment was build searching omologues sequences with BLAST [6] and aligning them with ClustalW [5]. The R19H mutation falls into the strictly conserved motif *RFP*, which is conserved into all the aligned sequences (see figure 5.6).

5.3.2 Biochemical and structural analysis

The structural analysis has been done on the crystal structure of the bovine COX which subunity COX6B1 shares more than 90% of sequence identity with the human COX. Considering the crystal structure of bovine COX, the R19 residue is predicted to form a strong saline bond with the adjacent highly conserved D17 residue and a weaker saline bond with conserved D35 residue (see alignment on figure 5.6). These bonds help maintain the appropriate conformation of the COX6B1 N-terminal loop, which is predicted to interact with subunit 2 of COX (figure 5.7).
The substitution of an elongated, flexible R with a bulkier, shorter and rigid H residue may well prevent the formation of the salt bridge with D35, and weaken the salt bridge with D17, as well; this altered conformation could in turn compromise the stability of the COX6B1 subunit within the COX dimer, which is the physiologically active form of the enzyme. The stability of the mutated COX was tested using a western blot analysis with antibodies specific to several COX subunits [23]. In particular, in experiments based on denaturing, sodium dodecyl sulfate-poly-acrylamide gel electrophoresis (SDS-PAGE) on mutant versus control muscle homogenate samples, a reduced crossreacting material (CRM) was detected for all tested COX subunits, including COX6B1 [23].
To further investigate the structural composition of mutant COX, the holocomplex was analyzed from muscle homogenates extracted in native conditions and separated by blue-native gel electrophoresis (BNGE). The fully assembled COX was reduced to approximately 40% in mutant versus control muscle [23]. Taken together, these results indicate that the R19H change in COX6B1 compromises

the stability of the muscle COX holocomplex, thus reflecting the severely reduced specific activity measured in this tissue.

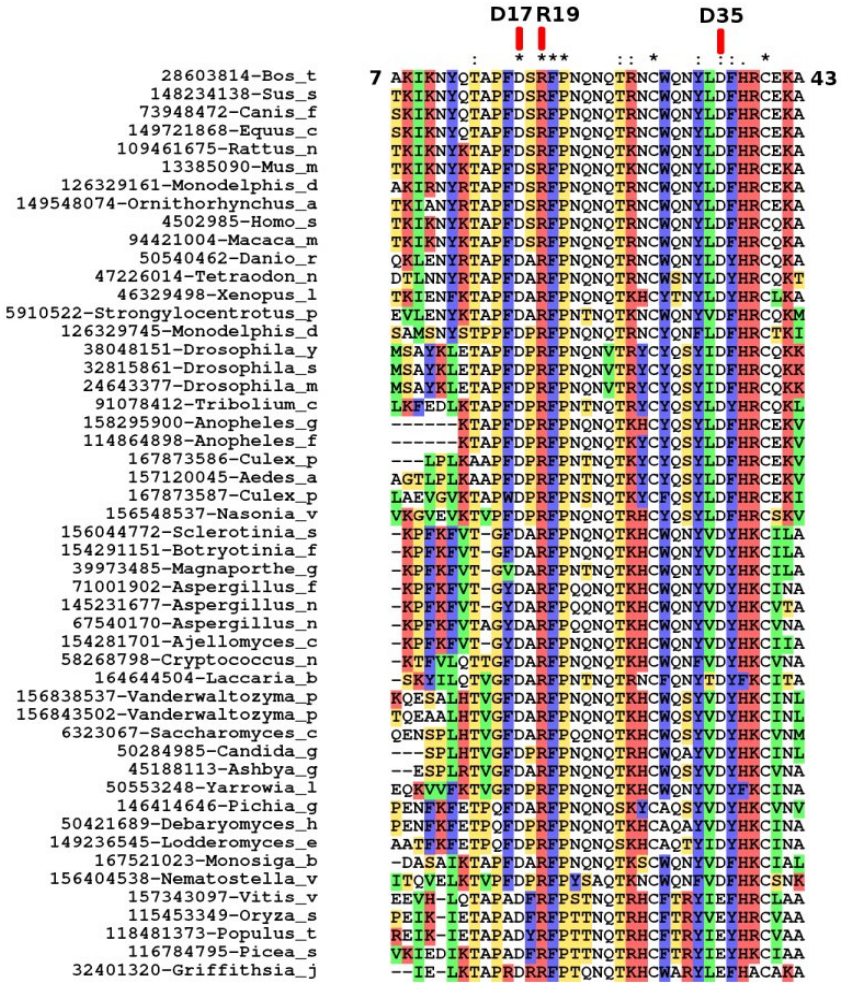


Figure 5.6: The alignment of the protein sequence in different species is shown. The numeration of the amino acid residues corresponds to the mature COX6B1 as the first M is cleaved in the import process of the precursor into mitochondria. The R19 and D35 residues are shown. It is clearly visible the high conservation of the R19 into all the aligned species.

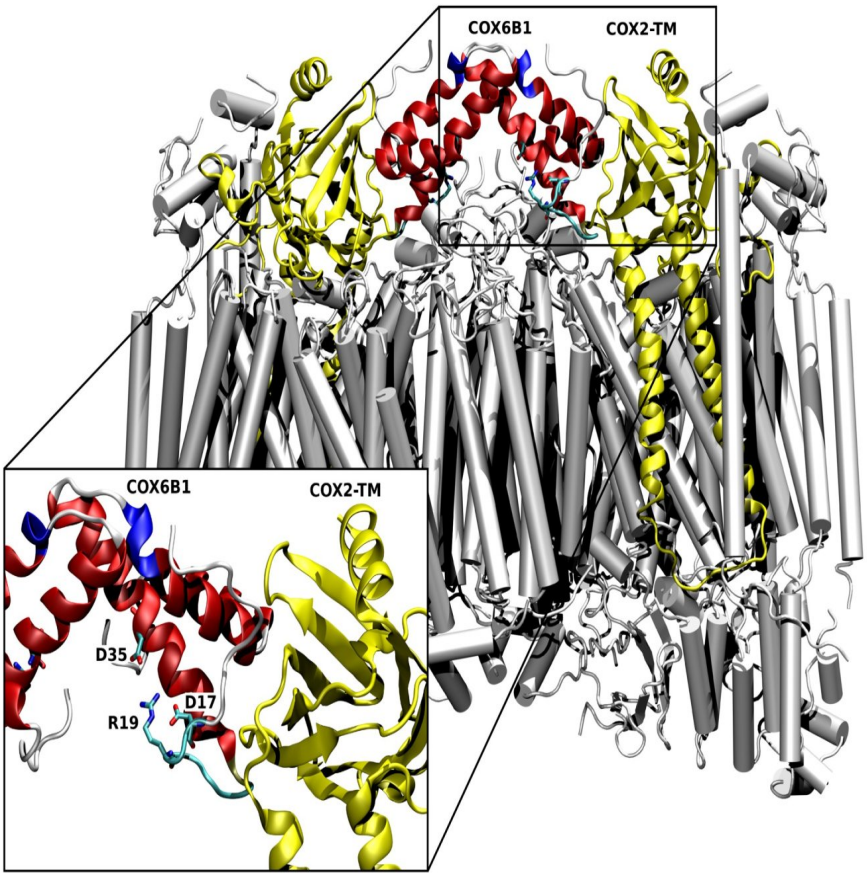


Figure 5.7: Crystal structure of bovine cytochrome C oxidase (PDB ID 2eij). The two chains of the subunit VIb isoform 1 are shown in red whereas the yellow chains representing the subunit 2 are shown in yellow. The interaction region between the subunit VIIb and the subunit 2 is magnified, moreover, the position of the residues R19, D35 and D17 VIIb are labelled for clarity.

So far, only mutations of mtDNA encoded COX subunits or nucleus encoded COX assembly factors have been associated with human COX deficiency raising the conjecture that mutations in nucleus encoded COX structural subunits were not viable extra utero. However, very recently, mutations that disrupt COX subunit 6A and knockdown of subunit COX5A were reported in Drosophila and zebrafish, respectively [23]. Both models do not result in embryonic lethality but rather in neurodegeneration and COX deficiency. In agreement with these

experimental models, our own data show that disease-causing abnormalities of nuclear COX subunits are also possible in humans.

5.4 Mutations of Mitochondrial elongation Factors EFG1 and EFTu

The mitochondrial respiratory chain (MRC) is a multiheteromeric enzymatic structure that performs oxidative phosphorylation (OXPHOS), a fundamental reaction of life that supplies about the 90% of the energy used by mammalian cells [11]. The MRC consists of five complexes, composed of 85 structural proteins, 13 of which are encoded by mtDNA (designated "mit" genes, according to the yeast mtDNA terminology), whereas the others are encoded by nuclear genes. Four MRC complexes contain the 13 mtDNA-encoded polypeptides as integral parts; seven are subunits of complex I (cI), one is a subunit of complex III (cIII), three are subunits of complex IV (cIV, which is the cytochrome c oxidase [COX]), and two are subunits of complex V (cV). Complex II (cII), also named succinate:ubiquinone oxidoreductase is the only MRC complex that lacks mtDNA-encoded subunits.

The mit genes are translated into proteins within the mithocondria by a protein-synthesis machinery, composed of both RNAs and proteins, which is largely independent from that responsible for translation of genes contained in the nuclear genome, which takes place in the cytosol. The RNA component of mitochondrial translation consists of 22 tRNAs and 2 ribosomal RNAs (rRNAs) encoded by mtDNA genes (designated "syn" genes, according to the yeast terminology), whereas the protein component is encoded by nuclear genes and consists of about 50 ribosomal proteins, the aminoacyl-tRNA synthetases, several tRNA maturation enzymes, the translation initiation, elongation, and termination factors, and a large number of unidentified factors, including ribosome-assembly factors [28]. Abnormalities in either gene set mitochondrial or nuclear can compromise mitochondrial translation. leading to multiple biochemical defects that may occur in the mtDNA-dependent MRC complexes, that, in turn, give to faulty OXPHOS and disease.

Over 100 disease-causing mutations are known in either tRNA- or rRNA-encoding mtDNA syn genes. In contrast, only a few mutations in mitochondrial translation protein factors have been reported. In particular, a missense mutation in pseudouridine synthase 1 (PUS1) was identified in Persian Jewish families affected by myopathy, lactic acidosis, and sideroblastic anemia (MLASA [OMIM 600462]) [11].

PUS1 converts uridine into pseudouridine in several positions of tRNAs synthesized in both nuclear and mitochondrial compartments. Lack of this posttranscriptional maturation of tRNAs leads to defective cytosolic and mitochondrial

translation [11]. A second observation concerned a homozygous stop mutation in MRPS16 (OMIM 609204), a protein of the mitochondrial small ribosomal subunit, which was found in one patient with severe lactic acidosis, developmental defects in the brain, and facial dysmorphisms [11]. A missense mutation in the mitochondrial elongation factor G1 (EFG1) was found in a singleton case affected by fatal neonatal liver failure and lactic acidosis associated with severe mitochondrial translation defect (combined oxidative phosphorylation deficiency 1 [OMIM 609060]) [11]. Finally, the same homozygous missense mutation in the mitochondrial translation elongation factor Ts (EFTs) has recently been reported in two unrelated babies, one affected by mitochondrial encephalomyopathy, the other by fatal hypertrophic cardiomyopathy [11]. The number of nuclear-encoded proteins that partecipate in the traslation of mithocondrial transcripts is about 200, thus it is plausible that defects in these proteins are either lethal or are underdiagnosed to a major extent.

5.4.1 Structural analysis of Mutant EFG1M496R and EFTuR339Q Proteins

Given that the three-dimensional structure of both the human EFG1 and EFTu is unknown, to carry out a structural analysis the 3D structure was computationally predicted.
Each protein model was built using the automatic homology modelling server SWISS-MODEL [21]. The model of EFTu was built using the coordinates of the crystal structure of EFTu from *Bos taurus* (PDB ID 1d2e), which shows 96% sequence identity with the human homolog. Whatcheck [27] (table 5.4) and AIDE [15] (table 5.5) analysis confirmed the quality of the model.

The EFG model was built using the protein crystal structure from *Thermus thermophilus* (PDB ID 2bm0) that shares 40% of residues identical to the human sequence. Despite the relatively low sequence identity, this protein contains four highly conserved domains that make it a suitable homology-modelling target. In fact, the Whatcheck and AIDE report shows that the overall model quality is only slightly lower in comparison to the EFTu model (see table 5.4 and 5.5).

To make predictions about the structures of human EFG1 and EFTu and about the effects of the human EFG1M496R and EFTuR339Q missense mutations, we took advantage of information on the crystal structure of these proteins available for a mammalian organism, *Bos taurus* (NCBI accession number NP_776632), and a micro-organism, *Thermus aquaticus* (NCBI accession number CAA46998). Like other translocases, EFG1 is a single polypeptide with a

Parameter	Z-score	
	EF-G1	EF-Tu
1st generation packing quality	-0.997	0.042
2nd generation packing quality	-2.565	0.063
Ramachandran plot appearance	-1.504	-0.210
chi-1/chi-2 rotamer normality	2.201	4.475
Backbone conformation	-0.656	0.515
Bond lengths	1.088	0.714
Bond angles	1.102	1.056
Omega angle restraints	1.078	0.678
Side chain planarity	2.700*	1.935
Improper dihedral distribution	1.313	1.277
Inside/Outside distribution	1.061	0.946

Table 5.4: Most relevant parameters of whatchek validation of the EF-G1 model. All parameter are expressed as Z-score values. All parameters give a score that is normal for well refined protein structures.

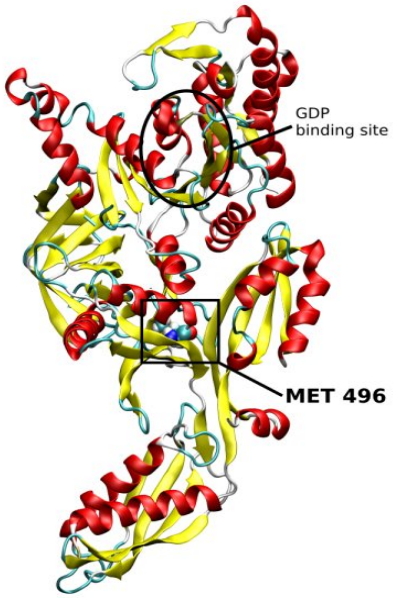
Template	RMSD	TM-score	LG-score	Overall
EF-G1	2.27	0.95	0.006	EXCELLENT
EF-Tu	1.13	0.81	0.124	EXCELLENT

Table 5.5: AIDE validation of EF-G1 and EF-Tu model.

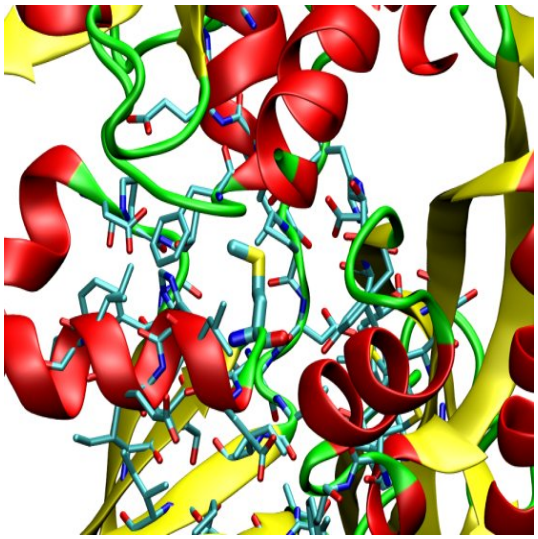
molecular weight of about 80 kDa. As shown in figure 5.8a, the EFG1M496R mutation is located at the end of a α-helix and before a β-sheet, in domain III of the mammalian protein.

Detailed analysis of this region, a pocket filled with both polar and apolar residues (figure 5.8b), suggests that a direct effect of the M496R amino acid substitution on the EFG1 GTPase hydrolytic activity is unlikely, because the region is far from the GTP-binding site (figure 5.8a). Rather, the replacement of the smaller, hydrophobic/negatively charged wild-type M with a bulkier, positively charged R residue is likely to produce a drastic structural rearrangement of the region that could, in turn, determine the destabilization of the entire protein or impede its correct interaction with the ribosome.

Human EFTu is also a GTPase consisting of a single polypeptide of about 45 kDa. The EFTuR339Q mutation is located on a solvent-exposed β-sheet on the outer surface of domain II of mammalian EFTu (figure 5.9a). This position makes it unlikely that the R339Q mutation can determine a drastic structural rearrangement of the protein, because the interaction with neighboring amino acid residues is minimal. However, domain II constitutes the tRNA-binding site of EFTu31 (figure 5.9b); therefore, the most probable effect of the R339Q substitution is to hamper the formation of the GTP:EFTu:aminoacyl-tRNA ternary complex. This hypothesis is supported by the demonstration that the amount and electrophoretic mobility of EFTuR339Q are both normal [11].

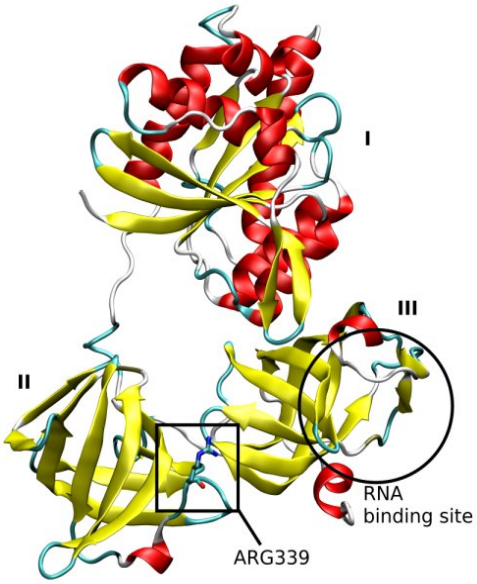


(a)



(b)

Figure 5.8: Modelling of mitochondrial EFG1. a. The EFG1 GDP-binding site and the wild-type Met 496 residue are indicated. b. Residues within 5Åof the M 496.



(a)

(b)

Figure 5.9: a. The three domains and the Arg 339 residue of EFTu are labeled. b. The model structure of human EFTu/tRNA complex is shown. The complex was obtained by superposing the bovine EFTu model on the crystal structure of the EFTu/tRNA complex of T. aquaticus (PDB ID 1ttt). The tRNA structure is in blue.

Bibliography

[1] Barrientos A, Barros M H, Valnot I, Rotig A, Rustin P, and Tzagoloff A. Cytochrome oxidase in health and disease. *Gene*, 286:53–63, 2002.

[2] Burlina A, Zacchello F, Dionisi-Vici C, Bertini E, Sabetta G, Bennet MJ, Hele DE, Schmidt-Sommerfeld E, and Rinaldo P. New clinical phenotype of branched-chain acyl-coa oxidation defect. *Lancet*, 14:1522–1523, 1991.

[3] Merinero B, Perez-Cerda C, Ruiz Sala P, Ferrer I, Garca M, Marinez Pardo M, Belanger-Quintana A, de la Mota J L, Martin-Hernandez E, Vianey-Saban C, Bischoff C, Gregersen N, and Ugarte M. Persistent increase of plasma butyryl/isobutyrylcarnitine concentrations as marker of scad defect and ethylmalonic encephalopathy. *J. Inherit. Metab. Dis.*, 29:685, 2006.

[4] Cameron C A D, Ridderstrom M, Olin B, and Mannervik B. Crystal structure of human glyoxalase ii and its complex with a glutathione thiolester substrate analogue. *Structure*, 7:1067–1078, 1999.

[5] Higgins D, Thompson J, Gibson T, Thompson J D, Higgins D G, and Gibson T J. Clustal w: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680.

[6] Altschul Stephen F, Gish Warren, Miller Webb, Myers Eugene W, and Lipmanl David J. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[7] Grossman L I and Lomax M I. Nuclear genes for cytochrome c oxidase. *BBA*, 1352:17419, 1997.

[8] Zafeiriou D I, Augoustides-Savvopoulou P, Haas D, Smet J, Triantafyllou P, Vargiami E, Tamiolaki M, Gombakis N, van Coster R, Sewell A C, Vianey-Saban C, and Gregersen N. Ethylmalonic encephalopathy: clinical and biochemical observations. *Neuropediatrics*, 38:78–82, 2007.

[9] Marasinghe G P K, Sander I M, Bennett B, Periyannan G, Yang K, Makaroff C A, and Crowder M W. Structural studies on a mitochondrial glyoxalase ii. *J. Biol. Chem.*, 280:40668–40675, 2005.

[10] Adams P L, Lightowlers R N, and Turnbull D M. Molecular analysis of cytochrome c oxidase deficiency in leigh's syndrome. *Ann. Neurol.*, 41:268–270, 1997.

[11] Valente L, Tiranti V, Marsano R M, Malfatti E, Fernandez-Vizarra E, Donnini C, Mereghetti P, De Gioia L, Burlina A, Castellan C, Comi GP, Savasta S, Ferrero I, and Zeviani M. Infantile encephalopathy and defective mitochondrial dna translation in patients with mutations of mitochondrial elongation factors efg1 and eftu. *Am. J. Hum. Gen.*, 80:44–58, 2007.

[12] Di Rocco M, Caruso U, Briem E, Rossi A, Allegri A E, Buzzi D, and Tiranti V. A case of ethylmalonic encephalopathy with atypical clinical and biochemical presentation. *Mol. Genet. Metab.*, 89:395–397, 2006.

[13] Holdorf M M, Bennett B, Crowder M W, and Makaroff C A. Spectroscopic studies on arabidopsis ethe1, a glyoxalase ii-like protein. *J. Inorganic Biochemistry*, 102:1825–1830, 2008.

[14] Huttemann M, Jaradat S, and Grossman L I. Cytochrome c oxidase of mammals contains a testes-specific isoform of subunit vibthe counterpart to testes-specific cytchrome c. *Mol. Reprod. Dev.*, 66:8–16, 2003.

[15] Mereghetti P, Ganadu M L, Papaleo E, Fantucci P, and De Gioia L. Validation of protein models by a neural network approach. *BMC Bioinformatics*, 9, 2008.

[16] Pecina P, Houstkova H, Hansikova H, Zeman J, and Houstek J. Genetic defects of cytochrome c oxidase assembly. *Physiol. Res.*, 53:S213–S223, 2004.

[17] Mineri R, Rimoldi M, Burlina A B, Koskull S, Perletti C, Heese B, von Dbel U, Mereghetti P, Di Meo I, Invernizzi F, Zeviani M, Uziel G, and Tiranti V. Identification of new mutations in the ethe1 gene in a cohort of 14 patients presenting with ethylmalonic encephalopathy. *J. Med. Genet.*, 45:473–478, 2008.

[18] DiMauro S and De Vivo D C. Genetic heterogeneity in leigh syndrome. *Ann. Neurol.*, 40:5–7, 1996.

[19] Melino S, Capo C, Dragani B, Aceto A, and Petruzzelli R. A zinc-binding motif conserved in glyoxalase ii, beta-lactamase and arylsulfatases. *Trends Biochem. Sci.*, 23:381–382, 1998.

[20] Yoshikawa S, Shinzawa-Itoh K, and Tsukihara T. Crystal structure of bovine heart cytochrome c oxidase at 2.8a resolution. *J. Bioenerg. Biomembr.*, 30:7–14, 1998.

[21] Schwede T, Kopp J, Guex N, and Peitsch M C. Swissmodel: an automated protein homology-modeling server. *Nucleic Acids Research*, 31:3381–3385, 2003.

[22] Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, and Yoshikawa S. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 a. *Science*, 272:1136–1144, 1996.

[23] Massa V, Fernandez-Vizarra E, Alshahwan S, Bakhsh E, Goffrini P, Ferrero I, **Mereghetti P**, D'Adamo P, GasspariniP, and Zeviani M. Severe infantile encephalomyopathy caused by a mutation in cox6b1, a nucleus encoded subunit of cytochrome c oxidase. *Am. J. Hum. Gen.*, 82(6):1281–1289, 2008.

[24] Sampson V and Alleyne T. Cytochrome c/cytochrome c oxidase interaction. direct structural evidence for conformational changes during enzyme turnover. *Eur. J. Biochem.*, 268:6534–6544, 2001.

[25] Tiranti V, Briem E, Lamantea E, Mineri R, Papaleo E, De Gioia L, Forlani F, Rinaldo P, Dickson P, Abu-Libdeh B, Cindro-Heberle L, Owaidha M, Jack R M, Christensen E, Burlina A, and Zeviani M. Ethe1 mutations are specific to ethylmalonic encephalopathy. *J. Med. Genet.*, 43:340–346, 2006.

[26] Tiranti V, D'Adamo P, Briem E, Ferrari G, Mineri R, Lamantea E, Mandel H, Balestri P, Garcia-Silva M T, Vollmer B, Rinaldo P, Hahn D H, Leonard J, Rahman S, Dionisi-Vici C, Garavaglia B, Gasparini P, and Zeviani M. Ethylmalonic encephalopathy is caused by mutations in ethe1, a gene encoding a mitochondrial matrix protein. *Am. J. Hum. Genet.*, 74:239–252, 2004.

[27] Hooft R W, Vriend G, Sander C, and Abola E E. Errors in protein structures. *Nature*, 381(6580):272, 1996.

[28] Bykhovskaya Y, Casas K, Mengesha E, Inbal A, and Fischel-Ghodsian N. Missense mutation in pseudouridine synthase 1 (pus1) causes mitochondrial myopathy and sideroblastic anemia (mlasa). *Am. J. Hum. Gen.*, 74(6):1303–1308, 2004.

Chapter 6

Outlook

L'essentiel est invisible pour les yeux.
Antoine de Saint-Exupéry (1900 - 1944)

The continuous increasing of the computing power and the huge quantity of biological information stored into the databases, leads to the computational methods to reach a development good enough to be considered a crucial support for the experimental methods. In this work we have studied several cases by means of different computational approaches for the analysis of the structure and function relationships.
In chapter 2 we describe a method, based on neural networks, developed for evaluate the accuracy of predicted three-dimensional protein structures. The artificial neural networks are mathematical models developed in analogy to the real neurons. Being able to describe complex relationship, they have been chosen to map the relation between the 3D structure and its accuracy. One of the great advantages of the artificial neural networks, is the ability to learn from examples. Hence, trained on a set of structure with known accuracy, the neural network is able, given a 3D structure, to predict its accuracy. This tool has been used in different studies described in this work, in which the prediction of the 3D structure of the protein under study, has been necessary.

In particular, in chapter 3, an interaction study between a new class of natural sweeteners (steviol glycosides) and the human sweet taste receptor (t1r2-t1r3), has been described. The relevance of these sweeteners is recently increased due to their non-caloric property and their high sweetening power. Moreover, hyploglycemic, diuretic and cardiotonic effects associated to these molecules

make them an important target not only for the food industry but also for the pharmaceutical one. The sweet taste receptor (t1r2-t1r3) is an heterodimeric transmembrane G-protein coupled receptor, whose structure has been predicted and evaluated using the method above mentioned. The interaction between the steviol glycosides and all the possible binding sites of the receptor, has been analyzed by means of an in-silico docking study, which allowed to identify the preferential binding site for the steviol glycosides. In particular, the transmembrane binding site seems to be the most suitable for this class of compounds.

In chapter 4 the relationship between the dynamical properties and the function of some psychrophilic enzyme has been studied. The psycrophilic enzymes are adapted to work at low temperature, compared to the mesophilic enzymes, they show an high catatlytic efficiency at low temperature. Supported by literature models and results, an accurate comparative study (psychrophile vs mesophile) of the thermodynamic properties of two different enzymes belonging to the elastases and the uracil-DNA-glycosylases families has been done. This study, carried out with molecular dynamics simulations, revealed, according to previous evidences, that the low temperature adaptation is related to the different flexibility of the psychrophilic compared to the mesophilic enzyme. This difference influences how the enzymes interact with their substrates.

In the last chapter, we report three cases in which a structural study has been used to support biochemical and genetical data for the analysis of the impact of point mutations on the protein structure and function and its effect on the associated disease. In particular, we have studied three different serious rare diseases which involve grave metabolic disorder associated to point mutations in mithocondrial proteins.

It is worth to underline the importance of the combined use both the approaches. The experimental methods require the processing and the analysis of the data, whereas the computational methods need the experimental data to be accurate.

Chapter 7

Appendix

7.1 Neural Networks

7.1.1 Back-propagation algorithm

Given a sigmoid neuron k belonging to the output layer we can compute the partial derivative of the error function for each input weight $w_{j,k}$ using the chain rules as,

$$\frac{\partial E}{\partial w_{j,k}} = \frac{\partial E}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_{j,k}} = \frac{\partial E}{\partial z_k} x_{j,k} \tag{7.1}$$

Where $z_k = \sum_j w_{j,k} x_{j,k}$ and $x_{j,k}$ is the input of the neuron k coming from the hidden neuron j

$\frac{\partial E}{\partial z_k}$ is the same regardless of which input weight of unit j we are trying to update and we denote this quantity as δ_k

$$\begin{aligned} \delta_k \quad = \frac{\partial E}{\partial w_{j,k}} \quad &= \frac{\partial}{\partial z_k} \frac{1}{2}(t_k - o_k)^2 \\ &= -(t_k - o_k) \frac{\partial o_k}{\partial z_k} \\ &= -(t_k - o_k) \frac{\partial f(z_k)}{\partial z_k} \\ &= -(t_k - o_k)(1 - f(z_k)) f(z_k) \\ &= -(t_k - o_k)(1 - o_k) o_k \end{aligned} \tag{7.2}$$

Where $f(z_k) = \frac{1}{1+e^{-x}}$ and $\frac{\partial f(z_k)}{\partial z_k} = (1 - f(z_k)) f(z_k)$. Thus

$$\Delta w_{j,k} = -\eta \frac{\partial E}{\partial w_{j,k}} = \eta \delta_k x_{j,k} \tag{7.3}$$

Now consider the neuron j belonging to a hidden layer. We make the following two important observations:

- For each unit k downstream from j, z_k is a function of z_j
- The contribution to error by all units $l \neq j$ in the same layer as j is independent of $w_{i,j}$

We want to calculate $\frac{\partial E}{\partial w_{i,j}}$ for each input weight $w_{i,j}$ for each hidden unit j. Note that $w_{i,j}$ influences just z_j which influences o_j which influences $z_k \forall k \in Downstream(j)$ each of which influence E. So, using the chain rules, we can write

$$\begin{aligned} \frac{\partial E}{\partial w_{i,j}} &= \sum_{k \in Downstream(j)} \frac{\partial E}{\partial z_k} \cdot \frac{\partial z_k}{\partial o_j} \frac{\partial o_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{i,j}} \\ &= \frac{\partial E}{z_k} \cdot \frac{\partial z_k}{o_j} \frac{\partial o_j}{\partial z_j} x_{i,j} \end{aligned} \tag{7.4}$$

As before we name this quantity δ_j. Also note that $\frac{\partial E}{\partial z_k} = \delta_k$, $\frac{\partial z_k}{\partial o_j} = w_{j,k}$ and $\frac{\partial o_j}{\partial z_j} = o_j(1 - o_j)$. Substituting,

$$\begin{aligned} \delta_j &= \sum_{k \in Downstream(j)} \frac{\partial E}{\partial z_k} \cdot \frac{\partial z_k}{\partial o_j} \frac{\partial o_j}{\partial z_j} \\ &= \delta_k w_{j,k} o_j (1 - o_j) \\ &= o_j(1 - o_j) \delta_k w_{j,k} \end{aligned} \tag{7.5}$$

Given a fee feed-forward network with n_i inputs, n_j hidden units, and n_k output units.
For each training example $\langle \vec{x}, \vec{t} \rangle$, Do

- Input the instance $\vec{x}$ and compute the output ou of every unit.
- For each output unit k, calculate

$$\delta_k = o_k(1 - o_k)(t_k - o_k) \tag{7.6}$$

- For each hidden unit j, calculate

$$\delta_j = o_j(1 - o_j) \sum_{j \in Downstream(j)} w_{j,k} \delta_k \tag{7.7}$$

- Update each network weight $w_{i,j}$ as follows:

$$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j} \tag{7.8}$$

$$\text{where} \quad \Delta w_{i,j} = \eta \delta_j x_{i,j} \tag{7.9}$$

7.2 Molecular Dynamics

7.2.1 Position Verlet algorithms

The position Verlet algorithm can be derived form the left and right Taylor expansion of the position

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t)\delta t + \frac{1}{2}\ddot{\mathbf{r}}_i(t)\delta t^2 + \frac{\delta t^3}{3!}\dddot{\mathbf{r}}_i + \mathcal{O}(\delta t^4) \tag{7.10}$$

$$\mathbf{r}_i(t-\delta t) = \mathbf{r}_i(t) - \dot{\mathbf{r}}_i(t)\delta t + \frac{1}{2}\ddot{\mathbf{r}}_i(t)\delta t^2 - \frac{\delta t^3}{3!}\dddot{\mathbf{r}}_i + \mathcal{O}(\delta t^4) \tag{7.11}$$

Summing these two equations:

$$\mathbf{r}_i(t+\delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\delta t) + \ddot{\mathbf{r}}(t)\delta t^2 + \mathcal{O}(\delta t^4) \tag{7.12}$$

$$\mathbf{r}_i(t+\delta t) \approx 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\delta t) + \ddot{\mathbf{r}}_i(t)\delta t^2 \tag{7.13}$$

The coordinate at time $t - \delta t$ can be computed using the initial velocities and accelerations as follow,

$$\mathbf{r}_i(0+\delta t) \approx 2\mathbf{r}_i(0) + \mathbf{v}_i(0)\delta_t + \frac{1}{2}\mathbf{a}_i(0)\delta t^2 \tag{7.14}$$

The velocity are not explicitly computed and can be computed using

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t+\delta_t) - \mathbf{r}_i(t-\delta_t)}{2\delta_t} + \mathcal{O}(\delta t^2) \tag{7.15}$$

A related, and more commonly used, algorithm is the Velocity Verlet algorithm, whose main advantage is to obtain more accurate velocities.

7.2.2 Velocity Verlet algorithms

The positions and the velocities are computed as,

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta_t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{7.16}$$

$$\mathbf{v}_i(t+\delta t) = \mathbf{v}_i(t) + \frac{1}{2}(\mathbf{a}_i(t) + \mathbf{a}_i(t+\delta_t))\delta t^2 \tag{7.17}$$

This scheme, however, requires the knowledge of the accelerations, $\mathbf{a}_i$, at timestep $t + \delta_t$. One may therefore decompose equation 7.17 into two steps. First calculate

$$\mathbf{v}_i(t+\frac{\delta_t}{2}) = \mathbf{v}_i(t) + \frac{1}{2}\mathbf{a}(t)\delta_t \tag{7.18}$$

then compute the actual forces on the particles at time $t+\delta_t$ from the potential energy derivatives and fnish the velocity calculation with

$$v(t+\delta_t) = v(t+\frac{\delta_t}{2}) + \frac{1}{2}a(t+\delta_t)\delta_t \tag{7.19}$$

Chapter 8

Publications:

Chapter 2: **Mereghetti P**, Ganadu M L, Papaleo E, Fantucci P and De Gioia L
Validation of protein models by neural network approach.
BMC Bioinformatic, 9:66, 2008

Chapter 3: **Mereghetti P**, De Gioa L, Temussi A, Fantucci P, Ganadu M L
Identification and characterization of stevioside binding site
in fase di scrittura

Chapter 4: **Mereghetti P**, Riccardi L, Ganadu M L, Brandsdal B O, Piercarlo F, De Gioia L, and Papaleo E
Conformational and free-energy landscape in cold- and warm-adapted serine-proteases: a molecular and essential dynamics approach.
in fase di scrittura

Papaleo E, **Mereghetti P**, Fantucci P, Grandori R, and De Gioia L
Free energy landscape, principal component analysis, and structural clustering to identify representative conformations from Molecular Dynamics simulations: the myoglobin case
Journal Molecular Graphics and Modelling, submitted

Chapter 5: Massa V, Fernandez-Vizarra E, Alshahwan S, Bakhsh E, Goffrini P, Ferrero I, **Mereghetti P**, D'Adamo P, GasspariniP, Zeviani M
Severe infantile encephalomyopathy caused by a mutation in COX6B1, a

nucleus encoded subunit of cytochrome c oxidase
American Journal of Human Genetic, 82(6):1281-1289, 2008

Mineri R, Rimoldi M, Burlina A, Koskull S, Perletti C, Bryce H, Von Dobel U, **Mereghetti P**, Di Meo I, Invernizzi F, Zeviani M, Uziel G, Tiranti V
Identification of new mutations in the ETHE1 gene in a cohort of 14 patients presenting with Ethylmalonic Encephalopathy
Journal of Medical Genetics, 45:473-478, 2008

Valente L, Tiranti V, Marsano RM, Malfatti E, Fernandez-Vizarra E, Donnini C, **Mereghetti P**, De Gioia L, Burlina A, Castellan C, Comi GP, Savasta S, Ferrero I, Zeviani M.
Infantile encephalopathy and defective mitochondrial DNA translation in patients with mutations of mitochondrial elongation factors EFG1 and EFTu.
American Journal of Human Genetics, 80(1):44-58, 2007

Other publications not included in this thesis:

Ami D, Neri T, Natalello A,**Mereghetti P**, Doglia S M, Zanoni M, Zuccotti M, Garagna S and Redi C A
Embryonic stem cell differentiation studied by FT-IR spectroscopy
BBA - Molecular Cell Research, 1783(1):98-106, 2007