UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI
Scienze Neurologiche, Neuropsicologiche, Morfologiche e Motorie

DOTTORATO DI RICERCA IN
IMAGING MULTIMODALE IN BIOMEDICINA

CICLO XXIII

TITOLO DELLA TESI DI DOTTORATO

HIV-1 negative factor binding to human thioesterase 8: insights from computational biology

S.S.D. BIO/16

Coordinatore: Prof. Andrea Sbarbati

Firma ____________________

Tutor: Prof. Andrea Sbarbati

Firma ____________________

Co-Tutor: Dr. Alejandro Giorgetti

Firma ____________________

Dottorando: Dott. Ing. Antonio Pozzo

Firma A. Tonio Pozzo

Summary

Abstract

HIV-1 Negative factor (Nef) is a protein essential for the metabolism of the virus.

Here we investigate the interactions of NEF with one of its targets on infected human cells, the human thioesterase 8 (hTE8) enzyme.

Homology modeling, virtual protein-protein docking and Molecular Dynamics Simulation experiments are carried out on the structural models of the enzyme and the complex respectively, with the aim of characterizing the putative interaction region.

A plausible, albeit approximate, binding region is identified. The latter help interpret existing site directed mutagenesis data. Our calculations suggest also that the system large-scale dynamics change upon complex formation.

Part A - Background and Theoretical Methods

1.0 HIV

1.1 HIV disease

Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family) that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to lifethreatening opportunistic infections. Infection with HIV occurs by the transfer of blood. HIV is present as both free virus particles and virus within infected immune cells. The four major routes of transmission are unsafe sex, contaminated needles, breast milk, and transmission from an infected mother to her baby at birth. Screening of blood products for HIV has largely eliminated transmission through blood transfusions or infected blood products in the developed world.

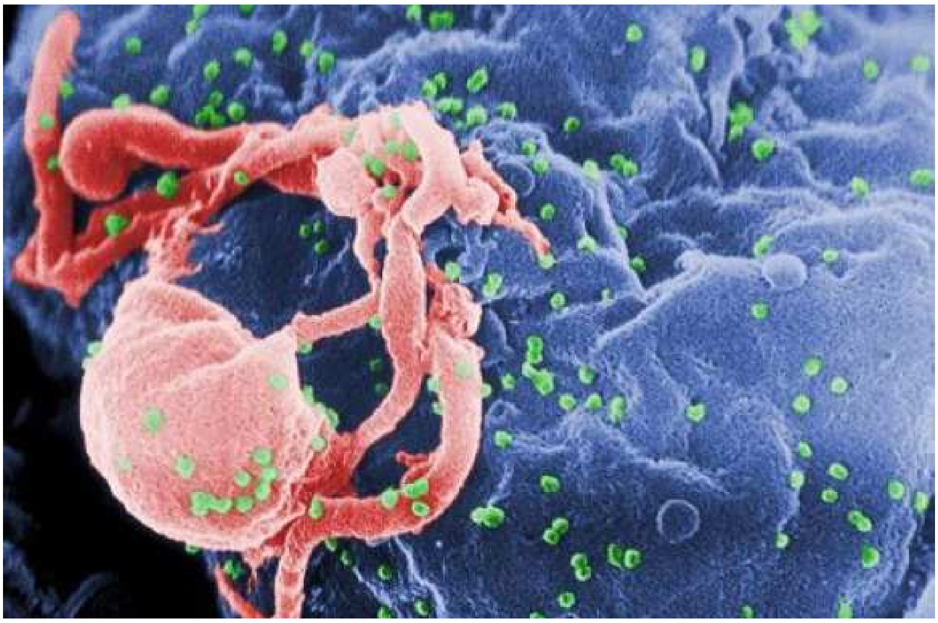


Figure 1. *Scanning electron micrograph of HIV-1 budding from cultured lymphocyte. (http://en.wikipedia.org/wiki/File:HIV-budding-Color.jpg)*

1.2 Negative Factor

The Negative Factor (Nef) protein from Human Immunodeficieny Virus type 1(HIV-1) is a 27-kDa-myristoylated protein required to produce a high viral load of the virus. In infected cells, this protein has been found to bind to a variety of proteins, including the src-family tyrosine kinases[1-2], the serin/threonine kinase]3-5], CD4[6,7], and thioesterase 8 (hTE8)[8]. Structural information of complexes between Nef and its identified cellular partners is currently of great interest to clarify Nef function in HIV-1 pathogenesis. So far, only the structure of the complex between Nef core domain and kinase SH3 domain has been solved using x-ray crystallography[9].

Although HIV-1 Nef was originally named "negative factor," it has been shown to have a positive role in viral replication and pathogenesis. Nef is a viral protein that interacts with host cell signal transduction proteins to provide for long term survival of infected T cells and for destruction of non-infected T cells (by inducing apoptosis). Nef also advances the endocyotsis and degradation of cell surface proteins, including CD4 and MHC proteins (CD4 is an integral membrane protein that functions in T-cell activation, and is the receptor for the HIV virus).

This action impairs cytoxic T cell function, thereby helping the virus to evade the host immune response. The multifunctional protein helps the virus maintain high viral loads and overcome host immune defenses, contributing to the progression of AIDS. Nef may be a valuable target for pharmaceutical intervention in AIDS progression.

Recently, point directed mutagenesis experiments were carried out on NEF with the aim of identifying the surface of interaction between the latter an one of its targets in human cells: the Acyl Thioesterase II. In the study, five residues that play a crucial role for the binding to hTE8 [10-12] were found, i.e. Asp108, Leu112, Phe121, Pro122 and Asp123.

1.3 Acyl Thioesterase

Acyl-CoA thioesterases are a group of enzymes that catalyse the hydrolysis of acyl-CoAs to the free fatty acid and coenzyme A (CoASH). They consequently have the potential to regulate intracellular levels of acyl-CoAs, free fatty acids and CoASH. They may also be involved in the metabolic regulation of peroxisome proliferation. Thioesters play a central role in cells as they participate in metabolism, membrane synthesis, signal transduction, and gene regulation.

Thioesterases catalyse the hydrolysis of thioesters to the thiol and carboxylic acid components. Many thioesterases have a hot dog fold, including YciA from Escherichia coli and its close sequence homologue HI0827 from Haemophilus influenzae(HiYciA). The E. coli thioesterase 8 reveals a new tertiary fold: a 'double hot dog'. It has an internal repeat with a basic unit that is structurally similar to the recently described beta-hydroxydecanoyl thiol ester dehydrase. The latter was shown to interact with the HIV-NEF protein in infected human cells [8]. Here we investigate the interactions of NEF with one of its targets on infected human cells, the human thioesterase 8 (hTE8) enzyme. Homology modeling, virtual protein-protein docking and Molecular Dynamics Simulation experiments are carried out on the structural models of the enzyme and the complex respectively, with the aim of characterizing the putative interaction region.

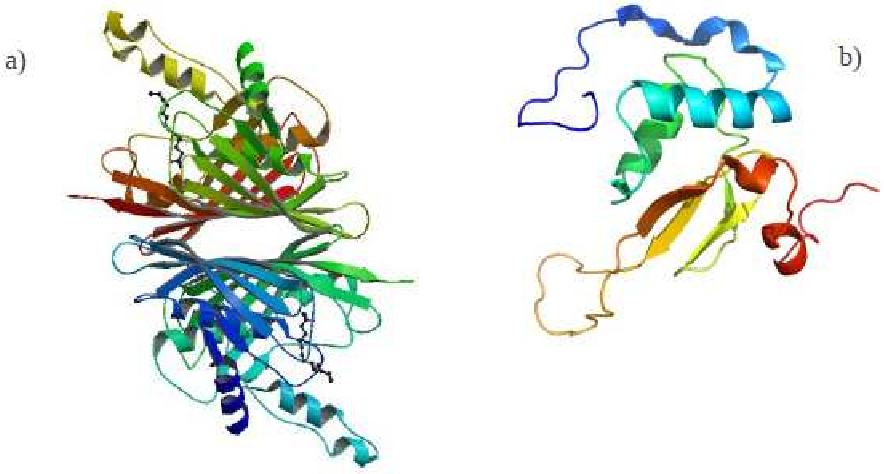


Figure 2. *Crystal structures from the a) Acyl-Coa thioesterase (PDB code 1C8U) and b) Nef (PDB code 2NEF)*

2.0 Comparative Modeling

2.1 The Levinthal paradox

One possible route to annotate a genome is to try and assign a structure to the protein products of the genes. In principle one could follow two routes: a physico-chemical approach whereby one tries to calculate the protein structure, or a heuristic approach where rules relating sequence to structure are derived from the analysis of known protein structures that have been experimentally determined.

The first route is clearly much more intellectually appealing. After all, given a protein sequence we know exactly its chemical composition, if we do not consider post-translational modifications, and all we need to know are the forces acting on each of the atoms so that we can compute their optimal relative position.

In order to follow this route we need to make sure that the functional protein structure is the conformation corresponding to the free energy minimum and, if this is the case, that we are able to calculate the energy of all possible protein conformation accurately enough to distinguish between the correct structure and all the others.

If one takes a folded protein, i.e. a protein in its functional conformation, places it in chemical conditions where all the forces are weakened and therefore where the protein unfolds, it is sufficient to remove the chemical agents used for denaturing the protein to recover the folded functional protein. This is the result of a very elegant experiment performed by Christian Anfinsen in 1973 [26]. The obvious interpretation of the experiment is that a protein sequence contains all the information needed to achieve its functional structure (the experiment is carried out in a test tube where there is nothing else but the protein) and that the functional or native structure is the one corresponding to the free energy minimum among those that the protein can explore (no matter how many times you repeat the experiment you always end up with the same final structure). Therefore we can assume that the native protein structure is the one corresponding to the free energy minimum (the limits of validity of this assumption are discussed later in this chapter).

All we need to do is to compute the energy of all possible conformations of a protein and select the one with minimum free energy. However there are at least two hurdles in this strategy, the first is that proteins are only marginally stable, i.e. the energy needed to unfold them is of the order of a few Kcal/mol and is brought about by a very large number of weak interactions, and therefore we would need to compute the energy of each interaction very accurately to distinguish between the native protein structure and all the others. The second is that the number of possible conformations of proteins is simply enormous. There are many interesting attempt to try and simulate the folding of a protein in a computer using various tricks, approximations and strategies, as it will be discussed later in this chapter, but in practice we do not have at the moment any method that can fold any protein only on the basis of the physico-chemical properties of its sequence and we have to recur to heuristic methods by exploiting the fact that we have access to several solved instances of our problem: all proteins of known sequence whose structure has been solved experimentally.

The enormous number of conformations available to a protein not only makes the task of computing them impossible, but implies that the protein itself cannot be randomly searching its conformational space.

The case against proteins searching conformational space for the global minimum of free energy was argued by Cyrus Levinthal in 1968[38]. The Levinthal paradox, as it is commonly known, can be demonstrated fairly easily. If we consider a protein chain of N residues, we can estimate the size of its conformational space as roughly 10N states. This assumes that the main chain conformation of a protein may be adequately represented by a suitable choice from just 10 different local conformations per residue. More technically, the assumption is that there are just 10 different common combinations of phi, psi and omega torsion angles for each residue type. This of course neglects the additional conformational space provided by the side chain torsion angles, but is a reasonable rough estimate, albeit an underestimate. The so-called paradox comes from estimating the time required for a protein chain to search its conformational space for the global energy minimum. Let’s think about a typical protein chain of length 100 residues and let’s assume that the atoms can move very fast - the speed of light even. Even at these physically impossible atom velocities, it would take the chain around 10^{82} seconds to search the entire conformational space, which compares rather unfavourably to the estimated age of the Universe (10^{17} seconds). Clearly proteins do not fold by searching their entire conformational space.

2.2 Comparative (homology) modeling

At some stage of the evolution of a species, some individuals might diverge sufficiently to give raise to a different species, i.e. become unable to interbreed in the wild producing fertile offspring with the other members of the originating species.
Proteins have limited stability brought about by a multitude of rather weak interactions among their atoms. This suggests that the delicate balance between destabilizing and stabilizing forces might be easily destroyed by a mutation and the mutated protein might not be able to fold. However, during evolution, function has to be preserved, therefore all the proteins that we observe can only contain non destabilizing mutations with respect to their immediate ancestor sequence. Can a small change destabilize the original protein structure and stabilize a completely different one, preserving stability, function, folding ability, etc.? This is rather unlikely, and indeed never observed. It follows that evolutionarily related proteins, that is proteins derived by a common ancestor via the accumulation of small changes, cannot but have similar structure, where mutations have been accommodated only causing small local rearrangements. If the number of changes, that is the evolutionary distance, is high these local rearrangements can cumulatively affect the protein structure and produce relevant distortions, but the general architecture, that is the fold, of the protein has to be conserved. On the other hand, if two proteins have evolved from a common ancestor it is likely that a sufficient proportion of their sequences has remained unchanged so that an evolutionary relationship can be deduced by their comparative analysis. Therefore if we can infer that two proteins are homologous, that is evolutionary related, the structure of one can be used as a first approximation of the structure of the other. This forms the basis of the technique known as comparative or homology modeling[45].

How well is a protein structure preserved during evolution? Chothia and Lesk [29] analyzed 32 pairs of homologous proteins of known structure and asked the question of how much the core of the structures diverged as a function of the sequence identity (a rough measure of the evolutionary distance). There are several definitions of the core of a protein structure. In their work, Chothia and Lesk used an almost tautological definition of core as the part of the protein structures that is more conserved between the two homologous proteins under study. Regardless the specific definition, we can intuitively understand what the core of a protein is: the part of the structure that is not peripheral to the folded nucleus of the protein, i.e. the protein without external “decorations” such as loops and small domains that are usually not very well conserved in evolution. In the same paper, Chothia and Lesk also analyzed the extent to which the core is conserved as a function of sequence identity. Their conclusion, supported by many subsequent analysis, is that there is a clear relationship between the divergence of the structures of homologous proteins and that it can be expressed as a function of their sequence identity.

2.3 Evolutionary history of the protein

The first step of a comparative modeling experiment is the detection of proteins evolutionarily related to it whose structure is known (templates). The next question we need to ask ourselves is: which amino acid of the target protein corresponds to which amino acid of the templates? In other words we need a sequence alignment between the target sequence and the sequence of the template protein. This is, without doubts, the most crucial aspect of a modeling procedure and one of the most difficult ones. There are several methods for aligning protein sequences, but here is the catch. All these methods try to reconstruct the evolutionary history of the protein. In other words, they tell us which amino acids are likely to be derived from the same amino acid of the ancestral protein that gave origin to the present sequences. However, this is not necessarily the alignment we need for homology modeling. Let us try and explain this with an example. Suppose that there is an insertion of one amino acid in a given position in our target sequence with respect to its template. Not only the inserted amino acid of the target does not have any equivalent amino acid in the template, but also the amino acids surrounding it are likely to have changed their position relative to the rest of the structure in order to accommodate the insertion and using their evolutionary counterparts as structural templates for their position is incorrect.

2.4 Best protein to be used as template

If more than one protein of known structure evolutionarily related to the target is available we have several possible choices. We can:

- use the one evolutionarily closer to the target, i.e. the one with the highest sequence similarity,
- "average" the coordinates of the templates and build a "theoretical template",
- take the structure of different regions from the different proteins selecting the regions where the local similarity is higher,
- build a model on the basis of each of the available templates and select the best one according to some criteria,
- derive constraints from the templates and subsequently build a structure that satisfies as many of them as possible.

Essentially, all these strategies are used in practice by different tools available to users[37] [42]. It is difficult to say which is the best in general, although it is becoming clearer that using multiple templates has to be preferred and probably the constraint based strategy is more effective in many cases.

2.5 Machine learning methods and template selection

Secondary structure and presence of disulfide bonds are among the features that can be successfully predicted. Generally speaking, prediction tools rely on the fact that, even if the overall structure is determined by the whole sequence, specific structure features can be strongly influenced by local features of the sequence. For example, alpha-helices and beta-sheets have different amino acid composition, and the same is true for the neighboring residues of disulfide-bonded and free cysteines. If we are able to understand these differences, we can use them to evaluate the probability for a residue to be in a secondary structure or for a cysteine to be disulfide bonded.

The basic idea is to analyze the set of proteins known at atomic resolution and adopt methods suited to extract correlations between structural features and local sequence features. Simplest methods are based on classical statistics and evaluate, for example, the propensity of alanine residues to be in a alpha-helical structure simply by computing the ratio between the alanine composition of alpha-helices and the overall alanine composition in proteins. Statistical methods can take into consideration more elaborate sequence features, but they often fail in extracting useful correlations when the complexity of the problem increases. For that reason, more versatile and flexible methods have been designed and implemented on the basis of the so called “machine-learning” theory. Among them, Neural Networks, Support Vector Machines and Hidden Markov Models are the most widely adopted. With different strategies, they are able to extract information from a set of known examples in an automatic way, on the basis of a rigorous mathematical framework. Owing to their architectures, they are able to deduce more complex rules of association between input (sequence) and output (structural feature) than classical statistic methods do. These rules are encoded in a set of numerical parameters whose values are fixed during the training phase and then used for predicting new sequences.

Versatility of machine learning methods allows different input encodings, more informative than the sole sequence, to be considered. In particular a general improvement of the performance can be obtained using sequence profiles upon multiple sequence alignments. In practice, given a sequence, similar sequences are searched in the data base and then aligned so as to obtain a representation of a whole family instead of a simple sequence. This representation highlights, for example, the conserved and mutated residues and this supplements the predictor input with evolutionary information.

The classical application of predictive methods to protein structure is the determination of secondary structure starting from sequence. Best methods for this task are based on Neural Network and Support Vector Machines and take as input the sequence profile of a 15/25-residue long window, centered around the residue to be predicted. When validated on proteins with known structure not used during the training phase, these methods predict the correct secondary structure for about 78% of residues[35] [46]. Better results can be obtained implementing a consensus of different methods[30] [46].

Another important structural feature that can be predicted is the presence of disulfide bonds, that is the bond between the sulfur atoms of two cysteine side chains. This is the only covalent bond that non adjacent residue can form in the native state and a correct prediction of the topology of disulfide connection strongly constrains the prediction of the overall structure. This task can be easily split into two steps. First of all, since only about 1/3 of cysteine residues are involved in disulfide bonding, it is necessary to discriminate them. Then the topology of the connections can be predicted. Concerning the first step, very efficient methods have been implemented that are able to predict the correct bonding state for 88% of cysteine residues and to give an overall correct prediction for 84% of proteins[40]. They are currently based on systems that integrate a Neural Network and a Hidden Markov Model. The former analyzes the composition of the profile in windows centered around each cysteine residue while the latter correlates the outputs that the neural networks computes for all the cysteine residue in the sequence[39].

The prediction of the topology of the disulfide bridges, i.e. of which cysteine pairs with which, is more difficult due to the combinatorial number of possible connection patterns for a given number of bonded cysteine residues. Important achievements have been reached, although a reliable prediction of the disulfide connectivity pattern can be performed only when two or three disulfide bonds are present in the protein[31].

In conclusion, the prediction of structural features starting from the sequence is not able to completely reconstruct the protein conformation. Nevertheless these procedures can greatly help this task since the predict constraints limit the number of possible conformations. Moreover the output of this tool can supply information useful in the implementation of fold recognition methods.

2.6 Side chain modeling

For insertion and deletions, methods are usually based on either an energy driven search for the possible conformations of the region of interest or on a database search of regions of protein of known structure that can provide a local template [27] [28] [32][34][43] [44]. The latter are usually selected on the basis of either a good fit of the regions flanking the region between target and local template, or on local sequence similarity. Side chain modeling often takes advantage of the preference of side chains for specific conformations, as deduced by the analysis of known protein structures. These preferences, tabulated in so called rotamer libraries[28] [33] [34] [36], are usually used as a starting point for subsequent refinement of the overall structure.
Once we have built our initial model, we need to “refine” it. What this simply means is that we now need to model the effect of the specific sequence changes that have occurred in our protein with respect to its template.

2.7 The CASP (Critical Assessment of Methods for Protein Structure Prediction)

Every two years crystallographers and NMR spectroscopists who are about to solve a protein structure are asked to make the sequence of the protein available together with a tentative date for the release of the final coordinates[41]. Predictors produce and deposit models for these proteins before the structures are made available and, finally, a panel of assessors compares the models with the structures as soon as they are available and tries to evaluate the quality of the models and to draw some conclusions about the state of the art of the different methods. The results are discussed in a meeting where assessors and predictors convene and the conclusions are made available to the whole scientific community via the World Wide Web and the publication of a special issue of the journal Proteins: Structure, Function, and Genetics. The collected data, amounting to tens of thousands of models for hundreds of targets is an invaluable resource for assessing the quality of protein models.

Although embarrassing, we have to admit that, so far, no available method, is able to consistently produce the correct structure for regions where insertions and deletions are located or to improve the initial model and make it “better”, i.e. closer to the real structure, while the accuracy of side chain modeling methods seems to be only limited by the quality of the prediction of the rest of the structure.

Notwithstanding the limitations of comparative modeling, this method remains the method of choice whenever possible for at least two reasons. First of all, the relative quality of a comparative model depends on the evolutionary distance between two proteins. In fact, both the probability of inferring the correct alignment between two proteins and the structural divergence between their structures are correlated with their evolutionary distance which can be estimated a priori. This implies both that it is possible to estimate the expected quality of a comparative model and its possible range of application beforehand and hence decide whether it is reasonable to embark in the task and also, perhaps most importantly, that one can attach an approximate reliability to any of the conclusions derived from the analysis of the model. The second, equally important aspect, is that the methodology will be especially effective in modeling regions of a protein that are more conserved during evolution. This implies that functionally important regions will be more correctly modelled than other, often of lower interest, regions.

3.0 Molecular Dynamics

3.1 Molecular Dynamics (MD) simulations

Molecular Dynamics (MD) simulation is a technique founded upon the basic principles of classical mechanics that provide a dynamical picture of the individual particles of the system at a microscopic level. Using this technique successive configuration of the molecular system (in the phase space of coordinates and momenta) is generated by integrating Newton's law of motion.

The result is a trajectory, which contains the microscopic time evolution of the system in the phase space. From the trajectory generated, one can compute the dynamical properties such as absorption spectra, rate constants and transport properties. Further, on combining MD with statistical mechanics as a mean of sampling, one can compute equilibrium properties such as

average thermodynamics quantities, structure, and free energies along the reaction path seen as a union of all possible states of the system. For instance, the statistical ensemble average of an observable **A** can be obtained as:

$$\langle A \rangle = \sum_{t=1}^{\tau \to \infty} A(t) \qquad (3.1)$$

The assumption made here is called the ergodic hypothesis (details in 3.3), i.e given an infinite amount of time, ensemble average of observable **A**, is equivalent to its time average. The main aspect in atomistic MD simulations are:

- An algorithm that samples the phase space
- The choice of the interaction potential, V(**r**), between the atoms of the system.

Several simulations approaches were developed in the last decades that differs in the method to sample the phase space. The most fundamental form used to describe equation of motion is the Lagrangian form:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\right) - \left(\frac{\partial L}{\partial q_j}\right) = 0 \qquad (3.2)$$

where

$$L(q,\dot{q})$$

is Lagrangian defined as the difference between the kinetic and potential energies L=K-V, where qj are generalized coordinates and q˙ j are the associated time derivative. The momentum conjugate to coordinate qj is given as:

$$p_j = \frac{\partial L}{\partial \dot{q}_j} \qquad (3.3)$$

On substitution with the usual definition of kinetic and potential terms with cartesian coordinates r_i, equation (3.2) becomes:

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i \text{ with } \mathbf{F}_i = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i} \qquad (3.4)$$

where V(**r**), the potential, is a function of the atoms positions and Fi represents the total force on atom i. In this equation one assumes that the nuclear motion of constituent particles obeys the laws of classical mechanics. This is an excellent approximation if the distance in the energetic (translational, rotational and vibrational) levels of the involved degrees of freedom is << kT, where k is the Boltzmann constant and T the temperature. In the Hamiltonian form the equation of motion for the cartesian coordinates is given by:

$$\dot{r}_i = \frac{p_i}{m_i} \text{ and } \qquad \dot{p}_i = \frac{\partial V(r)}{\partial r_i} \qquad (3.5)$$

3.1.1 *Integration of Newton equations of Motion*

Under the influence of a potential, the motions of atoms are strongly coupled to each other giving rise to many-body problems that cannot be solved analytically. Therefore, in MD calculation an iterative numerical procedure is employed to obtain an approximate solution for the equations of motion. The two important properties of the equations of motion to be noted are:

- They must be time reversibe (t = -t).
- Conservation of total Energy (Hamiltonian) of the system.

For the first point, as the Newton equations are time reversible also the algorithm used is supposed to satisfy the same time reversal symmetry. The algorithms that are not time reversible do not normally preserve the phase space volume, i.e. they do not satisfy the Liouville theorem. For the second point, conservation of Hamiltonian is equivalent to conservation of total energy of the system and provides an important link between MD and statistical mechanics. The energy conservation condition H(p,r) = E, defines a hypersurface in the phase space called the constant energy, imposing a restriction on system to remain on this surface. A good way to check the accuracy of the algorithm is to follow the temporal evolution of an observable A that should be conserved (e.g. the total energy). In general a good algorithm must be such that:

$$\frac{|A(t_n) - A(t_0)|}{\langle A(t)\rangle} \ll 1, \qquad \text{for } (t_n - t_0) \gg \Delta t \tag{3.6}$$

there is no drift in the total energy.

The MD integration of the Newton's equation which have a continuous form, are based on assumption that position, velocities and other dynamical properties can be discretized using the

aylor series expansion:

$$r(t+\delta t) = r(t) + \Delta t v(t) + \frac{1}{2}\Delta t^2 a(t) + \frac{1}{6}\Delta t^3 b(t) + ... \tag{3.7}$$

$$v(t+\delta t) = v(t) + \Delta t a(t) + \frac{1}{2}\Delta t^2 b(t) + \frac{1}{6}\Delta t^3 c(t) + ... \tag{3.8}$$

The choice of the integration method depends on the degree of accuracy of problem at hand. One of the most useful form used is the velocity verlet algorithm, a variant of verlet algorithm. The advantage is using velocity verlet method is that positions, velocities and acceleration are well synchronized that allow to calculate the kinetic energy contribution to the total energy at same time, from which potential energy is determined.

The equations are:

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(\Delta t)\Delta t + \mathbf{a}_i \frac{1}{2}\Delta t^2 + \mathbf{O}(\Delta t^3) \qquad (3.9)$$

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + [a_i(t) + a_i(t+\Delta t)]\frac{1}{2}\Delta t + \mathbf{O}(\Delta t^3) \qquad (3.10)$$

where $\mathbf{a}_i$,$\mathbf{r}_i$,$\mathbf{v}_i$ are respectively the acceleration on the atom *i*, the atom position and the atom velocity. The algorithm has an accuracy of O(Δt3) for the variables and it is reversible in time.

Together with conservation of energy and time-reversibility another important feature of an integrating algorithm is to permit long time steps Δt. It is expected that the numerical Newtonian trajectory will diverge from the "true" Newtonian trajectory. However, it is important that the integrating algorithm maintains a well defined energy tolerance ΔE throughout the simulation time.

The error (ΔE) is known to decrease on decreasing the time step Δt. The aim here is to find a balance between using the largest possible time step and maintaining an acceptable ΔE all along the simulation. A large time step would lead to faster exploration, but energy would fluctuate widely with the possibility of the simulation being catastrophically unstable, on other hand too short time step would lead to computation being needlessly slow. The choice of an integration step is determined by the nature of forces acting on the system. The golden rule is to choose time step ($\Delta t \sim 10^{-15}$ s) such that the fastest motion of the system can be integrated accurately. This requirement is a severe restriction, particularly as high frequency motions are relatively of less interest and have minimal effect on the overall behavior of the system. One suggested approach is to freeze out such vibrations by constraining the appropriate bonds to their equilibrium values. Details of this approach is discussed in subsection 3.1.4, or to use multiple time step approach which is discussed in subsection 3.1.2.

3.1.2 Multiple Time Step Integrator

One of the approaches to accelerate the integration of equations of motion, is to use "multi-time" step algorithm such as reverse reference system propagation algorithm (r-RESPA)[51]. In the algorithm (r-RESPA), the molecular system is classified into number of groups according to how rapidly the forces varies over time. The starting point is the Liouville operator formulation, which can cast the equations for the Hamiltonian system (see equation 3.5) in a general form:

$$\dot{\mathbf{x}} = iL\mathbf{x} \tag{3.11}$$

where x is the phase vector and iL is the Liouville operator. Consider a molecular system containing N atoms (or 3N degrees of freedom) with x = $\{r_i, p_i\}$ representing a point in the phase space. The Liouville operator in cartesian coordinates is defined as:

$$iL = \{..., H\} \equiv \sum_{i=1}^{3N} \left[\frac{\partial H}{\partial p_i} \cdot \frac{\partial}{\partial r_i} - \frac{\partial H}{\partial r_i} \cdot \frac{\partial}{\partial p_i} \right] \tag{3.12}$$

On subsituting equation 3.5 into equation 3.12, we get:

$$iL = \{..., H\} \equiv \sum_{i=1}^{3N} \left[\frac{p_i}{m_i} \cdot \frac{\partial}{\partial r_i} + \mathbf{F}_i \cdot \frac{\partial}{\partial p_i} \right] \tag{3.13}$$

where F_i is the force on i^{th} degree of freedom, and {...,...} is the

poisson bracket. The classical time propagator U(t) is unitary and defined as e^{iLT} , and the evolution of system Eq. 3.11 is expressed as:

$$\mathbf{x}(t) = e^{iLt}\mathbf{x}(0) \tag{3.14}$$

The action of operator U(t) on x(0) cannot be determined analytically, however the operator can be decomposed using Trotter theorem, such that the action of U(t) on x(0)for each part can be evaluated analytically. Applying the Trotter theorem we get:

$$\begin{aligned} e^{i(L_1+L_2)t} &= \left[e^{i(L_1+L_2)t/P}\right]^P = \left[e^{i(L_1+L_2)\Delta t}\right]^P \\ &= \left[e^{iL_1(\frac{\Delta t}{2})}e^{iL_2\Delta t}e^{iL_1(\frac{\Delta t}{2})}\right]^P + O(t^3/P^2) \end{aligned} \tag{3.15}$$

where Δt=t/P. For finite P, the numerical iteration procedure is accurate to the second order in the time step at long times.

From equation 3.15, for the three exponential terms, we define
the discrete time propagator (U_1,U_2) as:

$$\begin{aligned} \mathbf{G}(\Delta t) &= U_1(\frac{\Delta t}{2}) + U_2(\Delta t) + U_1(\frac{\Delta t}{2}) + O(t^3/P^2) \\ &= e^{iL_1(\frac{\Delta t}{2})} e^{iL_2 \Delta t} e^{iL_1(\frac{\Delta t}{2})} + O(t\Delta t^2) \end{aligned} \qquad (3.16)$$

Since the three exponential terms in G _t are separately unitary, G (Δt) is also unitary i.e $G^{-1}(t) = G(-t)$. Lets us now consider the propagator generated by subdivison as:

$$\begin{aligned} iL_1 &= \sum_{i=1}^{N} \frac{p_i}{m_i} \cdot \frac{\partial}{\partial r_i} \\ iL_2 &= \sum_{i=1}^{N} F_i \cdot \frac{\partial}{\partial p_i} \end{aligned} \qquad (3.17)$$

The operator $U_1(\Delta t/2)$ becomes a translation operator on the positions: $r_i \rightarrow r_i + \Delta t(p_i / m_i)$, and operator $U_2(\Delta t)$ becomes a translational operator of momenta: $p_i \rightarrow p_i + (\Delta t / 2), F_i(r)$. On combining these two facts to action of operators in equation 3.16 on complete set of positions and momenta, yields the approximate evolution:

$$\begin{aligned} \mathbf{r}_i(\Delta t) &= r_i(0) + \Delta t v_i(0) + \frac{\Delta t^2}{2m_i} F_i(0) \\ \mathbf{v}_i(\Delta t) &= v_i(0) + \frac{\Delta t}{2m_i}[F_i(0) + F_i(\Delta t)] \end{aligned} \qquad (3.18)$$

which is the famous velocity verlet [49] integrator derived using the operator formulation. The power of the operator based approach is its symplectic property which ensures no drift in the total energy, resistance to increase in time steps and allowsgenerating stable long trajectories.

r-RESPA algorithms have been successfully employed to incorporate motions on more than two time scales. Let us consider a system with three characteristics time scales, a reference force F_i^{ref} , and two corrections F_i^{del} and F_i^{Del}, such that $F_i = F_i^{ref} + F_i^{del} + F_i^{Del}$. We define their Liouville operators as iL^{ref}, $iL^{(del)}$ and $iL^{(Del)}$ and the corresponding timescales δt, Δt and $\Delta\mathcal{T}$ respectively. The three time step propogator can then be written as:

$$\begin{aligned}\exp(iL\Delta\mathcal{T}) = \exp\left(iL^{(Del)}\frac{\Delta\mathcal{T}}{2}\right)\left\{\exp\left(iL^{(del)}\frac{\Delta t}{2}\right)\left[\exp(iL_2^{(ref)})\right.\right.\\ \left.\left.\times\exp(iL_1^{(ref)}\delta t)\exp\left(iL_2^{(ref)}\frac{\delta t}{2}\right)\right]^n\exp\left(iL^{(del)}\frac{\Delta t}{2}\right)\right\}^m\\ \times\exp\left(iL^{Del}\frac{\Delta\mathcal{T}}{2}\right)\end{aligned} \quad (3.19)$$

Thus, the correction due to slowest time scale is applied every m·n timesteps, and the intermediate time scale correction is applied every n steps. Such numerical procedure lead to considerable saving in the CPU time to perform a MD simulation.

3.1.3 *The interaction potential*

The potential function V(**r**) from which the forces used in MD are derived depends on the atomic coordinates ri.
V(**r**) used in this thesis has the following expression:

$$\begin{aligned}V(r_1, r_2, \ldots, r_N) &= \sum_{\text{bonds}} \frac{1}{2}K_d(d-d_0)^2\\ &+ \sum_{\text{angles}} \frac{1}{2}K_\theta(\theta-\theta_0)^2\\ &+ \sum_{\text{improper dihedrals}} \frac{1}{2}K_\xi(\xi-\xi_0)^2\\ &+ \sum_{\text{dihedrals}} K_\phi\,[1+\cos(n\phi-\delta)]\\ &+ \sum_{ij\,LJ}\left[\left(\frac{C_{ij}^{12}}{r_{ij}^{12}}-\frac{C_{ij}^{6}}{r_{ij}^{6}}\right)\right]\\ &+ \sum_{ij\,\text{coulomb}}\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}\end{aligned} \quad (3.20)$$

The first two terms (two and three body interactions respectively) represent the bonds and angles potentials, that are approximated by harmonic functions. The third and fourth term describe four body interactions. Improper dihedral terms are typically described by an harmonic function. Instead proper dihedrals are described by periodic functions (i.e. cosine functions) of a given periodicity n. The last two terms are a Lennard-Jones (LJ) potential and a coulomb potential between pair (ij) of atoms.

The LJ terms reproduce the Van der Walls interactions, while the coulomb potential terms reproduce the electrostatic properties of a protein. These interaction are modelled using the two-body approximation which doesnot explicitly account for the polarization effects, but on a average. The parameters used in this kind of potentials are typically obtained from quantum chemical calculations and experimental data (e.g. crystallographic data, spectroscopic data, etc). Among the popular sets of parameters (force fields) for MD simulations of proteins we can cite for example AMBER, GROMOS, CHARMM and OPLS. They all use the potential function expression given above for all the atoms of the simulated system except for the GROMOS(and CHARMM19 force field) force field in which a united atom description is used for non-polar hydrogens.

In MD simulations the description of the solvent (water for most of the biologically interesting systems) can be explicit or implicit. In the first case solvent molecules with a full atomistic force field description are added in the simulation box at the experimental density. In the implicit solvent description the solvent is treated as a dielectric medium in which the system is embedded. This is clearly a more approximated description but it is also computationally much more efficient since in many practical cases the solvent constitutes the majority of the atoms.

3.1.4 *Constraints for Hydrogen*

Constraints are used in MD to fix bonds to their equilibrium value. This allows increasing the simulation time step _t. Constraining the bond lenght does not alter significantly the statistics as these are quantum degrees of freedom being mostly in their ground state at the normal simulation temperature. Using the bonds constraints it is possible to use $\Delta t \sim 2$fs (2-4 times larger than the one that can be used without constraints). A common method to introduce constraints is the algorithm SHAKE , in which after each time step the atoms positions iteratively are modified in order to satisfy the constraint.

SHAKE may have convergence problems when applied to large planar groups and its implementation could hinder the efficiency of computing. To improve these aspects the LINCS algorithm was recently introduced. For water molecules it is also possible to use an analytic solution of SHAKE called SETTLE.

3.1.5 *Boundary conditions*

To simulate a finite size system, boundary conditions are needed to avoid artifacts near the border of the simulation box. Typically periodic boundary conditions (PBC) are used. In this scheme short range non bonded interactions are calculated using the minimal image convention (only the nearest replica is considered).

Typically a cut-off radius (Rc) is used for LJ interactions of the order of 10 Å. To avoid interactions between a particle and its periodic image each box side must be larger than 2Rc.

The coulomb energy is instead treated considering the full periodicity of the system. For a periodic lattice made by N particles it is given by:

$$E = \frac{1}{8\pi\epsilon_0} \sum_{|n|=0}^{\infty}{}^{*} \left[\sum_{i=0}^{N} \sum_{j=0}^{N} \frac{q_i q_j}{|r_{ij} + n|} \right] \quad (3.21)$$

where *n* indicates the periodic images, *i*,*j* the particles and the

symbol * indicates that the summation does not contain the term with *i* = *j* if *n* = 0.

The periodicity of the system speeds up the evaluation of the coulombic term. Although convenient, effective, and simple to apply, certain subtle problems arise for long range forces (electrostatics), whose spatial range may extend beyond the boundaries of the container into surrounding images, present a challenge.

Long range forces can only be correctly calculated by summing over all the periodic replicas of the original system. However, the associated computational effort is considerable. Fortunately, methods have been developed to treat this problem. Specifically, the Ewald summation technique, developed originally to treat Coulomb interactions and later extended to treat general interactions of the form $1/r_n$

for n≤3 has proved enormously successful.

The basic idea behind the technique is to divide the relevant part of the potential into a short range and a long range contribution.

For the Coulomb potential, 1/r, for example, this can be achieved via the identity

$$\frac{1}{r} = \frac{\mathrm{erf}(\alpha r)}{r} + \frac{\mathrm{erfc}(\alpha r)}{r} \tag{3.22}$$

where erf(x) and erfc(x) are the error function and complementary error function, respectively (erf(x) + erfc(x)= 1). The variable, R, is a convergence parameter, which can be optimized for each system studied. The short range term, erfc(αr)/r, is treated as an ordinary short range interaction, i.e., using a spherical cutoff to truncate the interaction at large spatial distances where the potential is small. The long range term, erf(αr)/r, is Fourier transformed into reciprocal space, where it takes the short-ranged form, $\exp(-g^2/4\alpha^2)$, and can be evaluated accurately by summing over only a small number of reciprocal space vectors of the simulation cell. Such reciprocal space sums can be evaluated with high a degree of efficiency (N log N) using particle-mesh methods(PME)[54]. An extension of PME is the smooth PME. With respect to PME, this method uses a fixed cuttoff in the direct sum and uses the B-spline interpolation of the reciprocal space structures onto a rectangular grid, permitting the use of fast Fourier transforms to efficiently calculate the reciprocal sum.
In this thesis we use SPME method to evaluate the electrostatic energies.

3.1.6 *Statistical Ensembles*

Molecular dynamics can be performed in different statistical ensembles. The traditionally used ensemble to perform MD is the micro-canonical ensemble (NVE), where the number of particles (N), the volume (V), and the total-energy (E) of the system are fixed to a constant value.
The simple extension of NVE ensemble is the canonical one (NVT), where the number of particles, the volume and the temperature are fixed to a constant value. The temperature T, in contrast to the number of particles N and volume V, is an intensive parameter. The temperature T is related to the time average of the kinetic energy given as:

$$\mathbf{T} = \frac{2\,E_{kin}}{3\,Nk_B} = \frac{1}{3Nk_B}\sum_{i=1}^{N}\frac{p_i{}^2}{m} \tag{3.23}$$

where, E$_{kin}$ is the kinetic energy, k$_B$ is the Boltzmann constant. The simplest way to control the temperature, is to rescale the velocities at each step by the factor

$$\lambda = \sqrt{\frac{T_{req}}{T_{curr}}}$$

where T$_{curr}$ is the current temperature calculated from the kinetic energy and T$_{req}$ is the desired temperature (for instance 300 K). However, an alternative way to maintain is to couple the system to an external heat bath that is fixed at the desired temperature. The bath acts as a source of thermal energy, supplying or removing heat from the system as appropriate. This thermostat is named as the "Berendsen" thermostat. It is extremely efficient for relaxing a system to the target temperature, but once the system has reached equilibrium, it might be more important to probe a correct canonical ensemble.

Extended system methods, was originally introduced for performing constant MD simulation by Nosè in 1984, and subsequently developed by Hoover in 1985. The idea of the method was to reduce the effect of an external system, acting as a heat reservoir, to an additional degree of freedom s. This reservoir has a potential energy $(f+1)k_B T \ln s$, where f is the number of degrees of freedom in the physical system and T is the desired temperature.

The kinetic energy of the reservoir is given as $(Q/T)(ds/dt)^2$, where Q is considered as the fictitious mass of the extra degree of freedom. The magnitude of Q determines the coupling between the reservoir and the real system and so influences the temperature fluctuation. If Q is large then the energy flow is slow; in the limit of infinite Q, conventional molecular dynamics is regained.

However, if Q is small then the energy oscillates, resulting in equilibrium problems. It has been suggested that Q should be proportional to $f k_B T$.

Another ensemble we discuss here it the NPT ensemble, an extension of NVT ensemble, where together with temperature the pressure of the system is maintained to a constant value. As most experimental measurements are usually made under conditions, which include a fixed pressure P, temperature T, and number of atoms N (constant-NPT ensemble), and so simulations in the isothermal-isobaric ensemble are the most directly relevant to experimental data. A simulation in NPT ensemble maintains the constant pressure by changing the volume of the simulation cell.

The amount of volume fluctuation is related to the isothermal compressibility, κ

$$\kappa = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \quad (3.24)$$

An alternative to maintain constant pressure is to couple the system to a “pressure“ bath, analogous to the temperature bath.

The rate of change of pressure is given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}(P_{bath} - P(t)) \quad (3.25)$$

where τ_p is the coupling constant, P_{bath} is the pressure of the ’bath’, and P(t) is the actual pressure at time t. The volume of the simulation box is scaled by a factor λ, which is equivalent to scaling the atomic coordinates by a factor $\lambda^{1/3}$. Thus:

$$\lambda = 1 - \kappa\frac{\delta t}{\tau_p}(P - P_{bath}) \quad (3.26)$$

and the new position are given by:

$$r_i^{new} = \lambda^{1/3} r_i \quad (3.27)$$

In the extended pressure-coupling systems, an extra degree of freedom, corresponding to the volume of the box, is added to the system. The kinetic energy associated with this degree of freedom (which can be considered to be equivalent to piston

acting on the system), is $(1/2Q)(dV/dt)^2$, where Q is the ’mass’ of the piston. The piston also has a potential energy PV, where P is the desired pressure and V is the volume of the system. The volume varies in the simulation with the average volume being determined by the balance between the internal pressure of the system and the desired external pressure. In this thesis, we have performed MD simulation in both NVT and NPT ensembles.

3.2 Long Time Scale Simulations

Molecular Dynamics (MD) simulations allow investigating processes occurring on timescales of ~100ns. However, most interesting and relevant biological process happen on time scales that are orders of magnitude larger, and are therefore termed as rare events. For example, protein folding (µs-few seconds), protein protein interactions, transport of molecules across membrane channels (order ~ µs) and many others. Over the years, we have observed an astounding increase in computer power (Blue gene, DESRES), which promise to increase utility of MD simulations to investigate more and more complex systems on µs timescale
However, these supercomputing machines are not available to all the research groups. Therefore another approach to overcome the timescale problem is to renounce the all atom approach and to use coarse grained models. This would retain the essential characteristics, however you require a detailed knowledge of system, that is often not available.
For systems, where its important to maintain the atomistic description, one can exploit methodology aimed at accelerating rare events to timescales reachable in MD simulations. Notable success has been achieved is using the accelerating methodology in diverse fields of interest. From their scope and range of applicability, they are classified in four categories :

1. Methods aimed at improving sampling, in a subspace of few predefined collective variables (CVs), that allow reconstructing the probability distributions as a function of chosen CVs. Examples of these methods include thermodynamic integration , free energy perturbation, umbrella sampling , conformational flooding, weighted histogram, steered MD, Jarzynski's identity based methods and adaptive force bias. The power of these methods in highly dependent on judicious choice of CVs, and computational performance degrades as a function of the number of variables.

2. Methods aimed at exploring the transition mechanism. Examples in these catogeries are transition path sampling, finite temperature string method, transition interface sampling and forward flux methods. These

methods do not require in most cases, an explicit definition of a reaction co-ordinate, but require a priori knowledge of initial and final states of process under investigation.

3. Methods for exploring the potential energy surfaces and localizing the saddle points that corresponds to a transition state. Examples in these catogeries are dimer method, hyperdynamics , multiple time scale accelerated MD and event based relaxation . The power of these methods is limited to low dimensionality, and reliability degrades with the complexity of system.

4. Methods in which the phase space is explored simultaneously at different values of temperatures, are parallel tempering and replica exchange, or as a function of the potential energy, such as multicanonical MD and Wang-Landau.

3.3 Statistical Bases

Molecular dynamics simulations generate information at the microscopic level, including atomic positions and velocities. The conversion of this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc., requires statistical mechanics. Statistical mechanics is fundamental to the study of biological systems by molecular dynamics simulation

In a molecular dynamics simulation, one often wishes to explore the macroscopic properties of a system through microscopic simulations, for example, to calculate changes in the binding free energy of a particular drug candidate, or to examine the energetics and mechanisms of conformational change. The connection between microscopic simulations and macroscopic properties is made via statistical mechanics which provides the rigorous mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas. With molecular dynamics simulations, one can study both thermodynamic properties and/or time dependent (kinetic) phenomenon.

Statistical mechanics is the branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules making up the system. The system could range from a collection of solvent molecules to a solvated protein-DNA complex. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced.

The thermodynamic state of a system is usually defined by a small set of parameters, for example, the temperature, T, the pressure, P, and the number of particles, N. Other thermodynamic properties may be derived from the equations of state and other fundamental thermodynamic equations.

The mechanical or microscopic state of a system is defined by the atomic positions, q, and momenta, p; these can also be considered as coordinates in a multidimensional space called phase space. For a system of N particles, this space has 6N dimensions. A single point in phase space describes the state of the system. An ensemble is a collection of points in phase space satisfying the conditions of a particular thermodynamic state. A molecular dynamics simulations 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 and their respective momenta.

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages; one justification for this is that there has been good agreement with experiment. An ensemble average is average taken over a large number of replicas of the system considered simultaneously.

The dilemma appears to be that one can calculate time averages by molecular dynamics simulation, but the experimental observables are assumed to be ensemble averages. Resolving this leads us to one of the most fundamental axioms of statistical mechanics, the ergodic hypothesis, which states that the time average equals the ensemble average.

The Ergodic hypothesis states

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}$$

Ensemble average = Time average

The basic idea is that if one allows the system to evolve in time indefinitely, that system will eventually pass through all possible states. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this equality is satisfied. If this is the case, experimentally relevant information concerning structural, dynamic and thermodynamic properties may then be calculated using a feasible amount of computer resources. Because the simulations are of fixed duration, one must be certain to sample a sufficient amount of phase space.

4.0 Software

4.1 HHpred

HHpred [13] is a tool for structural modeling of amino acids chains, with the help of profiles generated using the Hidden Markov Models methodology and starting from a FASTA sequence from a multiple sequence alignment.

Once the profile is generated, the program use an alignment algorithm profile versus profile, to proceed in searching of structural templates in an internal database that contains Markov profiles for every PDB entry.

Search options include local or global alignment and scoring secondary structure similarity. HHpred can produce pairwise query-template sequence alignments, merged query-template multiple alignments (e.g. for transitive searches), as well as 3D structural models calculated by the MODELLER software from HHpred alignments.

HHpred gives very good and reliable results. To testify this merits, you can find in the official site (http://toolkit.tuebingen.mpg.de/hhpred) the CASP9 results

The most successful techniques for protein structure prediction rely on identifying homologous sequences with known structure to be used as template. This works so well because structures diverge much more slowly than sequences and homologous proteins may have very similar structures even when their sequences have diverged beyond recognitio. If the relationship is so remote that no common function can be assumed, one can generally still derive hypotheses about possible mechanisms, active site positions and residues, or the class of substrate bound. When a homologous protein with known structure can be identified, its structure can be used as a template to model the 3D structure for the protein of interest, since even remotely homologous proteins generally have quite similar 3D structure. The 3D model may then help to generate hypotheses to guide experiments. When searching for remote homologous, it is wise to make use of as much information about the query and database proteins as possible in order to better distinguish true from false positives and to produce optimal alignments. This is the reason why sequence-sequence comparison is inferior to profile-sequence comparison. Sequence profiles contain for each column of a multiple alignment the frequencies of the 20 amino acids.

They therefore contain detailed information about the conservation of each residue position, i.e. how important each position is for defining other members of the protein family, and about the preferred amino acids. Profile Hidden Markov Models (HMMs) are similar to simple sequence profiles, but in addition to the amino acid frequencies in the columns of a multiple sequence alignment they contain information about the frequency of inserts and deletions at each column. Using profile HMMs in place of simple sequence profiles should therefore further improve sensitivity. HHpred is the first server to employ HMM-HMM comparison, based on a novel statistical method that we have developed recently. Using HMMs both on the query and the database side greatly enhances the sensitivity and selectivity over sequence-profile based methods such as PSI-BLAST.

4.2 MODELLER

MODELLER is a computer program that models three dimensional structures of proteins and their assemblies by satisfaction of spatial restraints[14].

MODELLER is most frequently used for homology or comparative protein structure modeling: the user provides an alignment of a sequence to be modeled with known related structures and MODELLER will automatically calculate a model with all non-hydrogen atoms.

The inputs of MODELLER are:

a) the sequence alignment between the target (the protein to be modeled) and the template (homologous protein with known structure;

b) the crystal structure of the template.

MODELLER extracts the spatial restraints from the template to the target producing a 3D structure that satisfies these restraints as well as possible. Restraints can be derived from a number of different sources. These include related protein structures (comparative modeling), NMR experiments (NMR refinement), rules of secondary structure packing (combinatorial modeling), cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site directed mutagenesis, intuition, residue-residue and atom-atom potentials of mean force, etc. The restraints can operate on distances, angles, dihedral angles, pairs of dihedral angles and some other spatial features defined by atoms or pseudoatoms. The final 3D model is then obtained by optimization of a molecular probability density function (pdf). The molecular pdf for comparative modeling is optimized with the variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing. MODELLER can also perform multiple comparison of protein sequences and/or structures, clustering of proteins, and searching of sequence databases. The program is used with a scripting language and does not include any graphics. The method and its applications to biological problems are described in detail in references listed in Section 1.2. Briefly, the core modeling procedure begins with an alignment of the sequence to be modeled (hTE8) with related known 3D structures (1C8U). This alignment is usually the input to the program (Figure 3). The output is a 3D model for the target sequence containing all main-chain and side-chain non-hydrogen atoms. Given an alignment, the model is obtained without any user intervention. First, many distance and dihedral angle restraints on the target sequence are calculated from its alignment with template 3D structures (Figure 3). This analysis relied on a database of 105 family alignments that included 416 proteins with known 3D structure. The form of these restraints

was obtained from a statistical analysis of the relationships between many pairs of homologous structures. By scanning the database, tables quantifying various correlations were obtained, such as the correlations between two equivalent Cα−Cα distances, or between equivalent main-chain dihedral angles from two related proteins. These relationships were expressed as conditional probability density functions (pdf) and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of a residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins.

Another example is the pdf for a certain Cα−Cα distance given equivalent distances in two related protein structures (Figure 4). An important feature of the method is that the spatial restraints are obtained empirically, from a database of protein structure alignments. Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry are combined into an objective function.

Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method employing methods of conjugate gradients and molecular dynamics with simulated annealing (Figure 5). Several slightly different models can be calculated by varying the initial structure. The variability among these models can be used to estimate the errors in the corresponding regions of the fold.

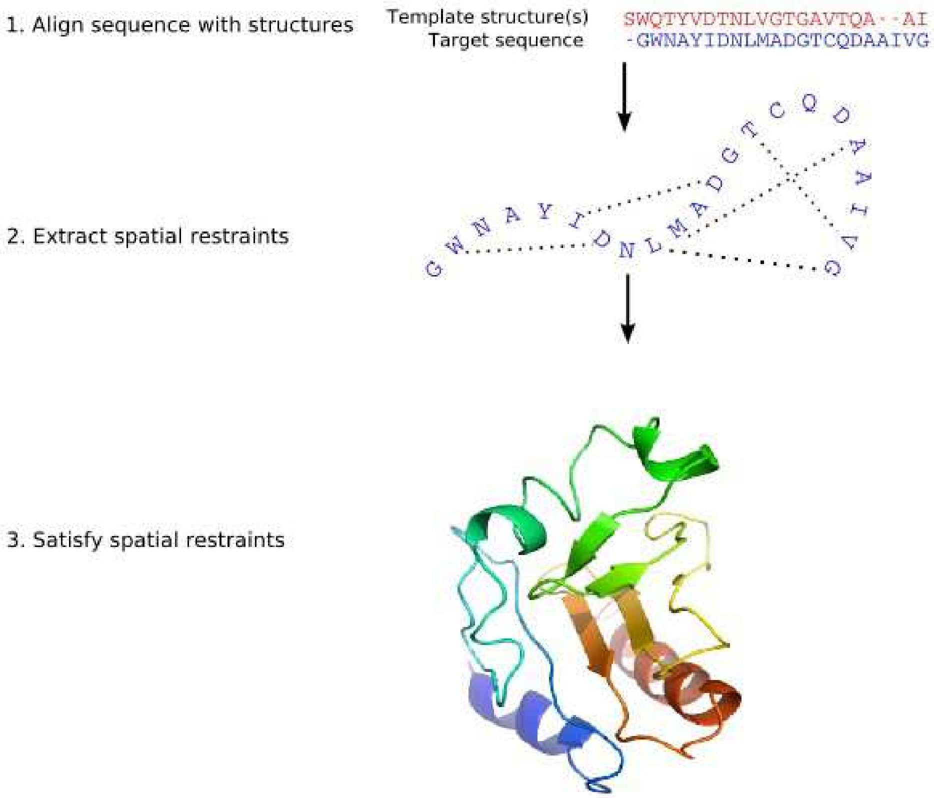

1. Align sequence with structures
Template structure(s)
Target sequence
SWQTYVDTNLVGTGAVTQA--AI
-GWNAYIDNLMADGTCQDAAIVG
2. Extract spatial restraints
G W N A Y I D N L M A D G T C Q D A A I V G
3. Satisfy spatial restraints

Figure 3. *First, the known, template 3D structures are aligned with the target sequence to be modeled. Second, spatial features, such as Cα−Cα distances, hydrogen bonds, and main-chain and side-chain dihedral angles, are transferred from the templates to the target. Thus, a number of spatial restraints on its structure are obtained. Third, the 3D model is obtained by satisfying all the restraints as well as possible.*

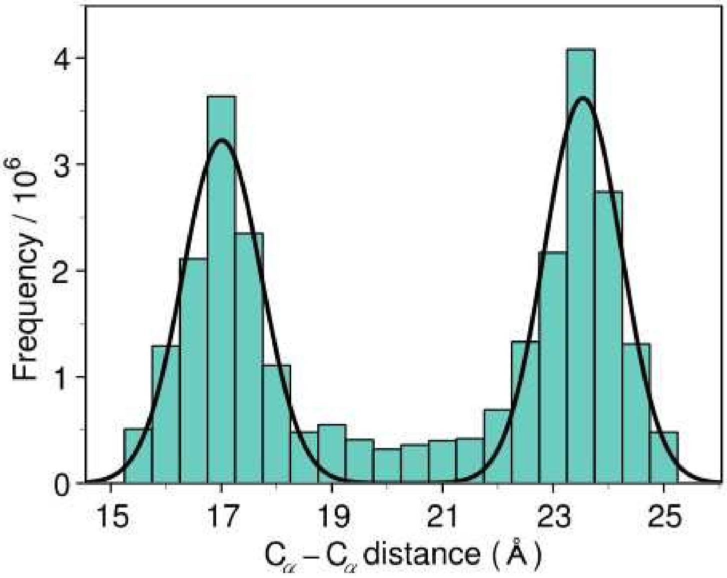


Figure 4. *The restraint (continuous line) is obtained by least-squares fitting a sum of two Gaussian functions to the histogram, which in turn is derived from the database of alignments of protein structures. In practice, more complicated restraints are used that depend on additional information, such as similarity between the proteins, solvent accessibility, and distance from a gap in the alignment [42].*

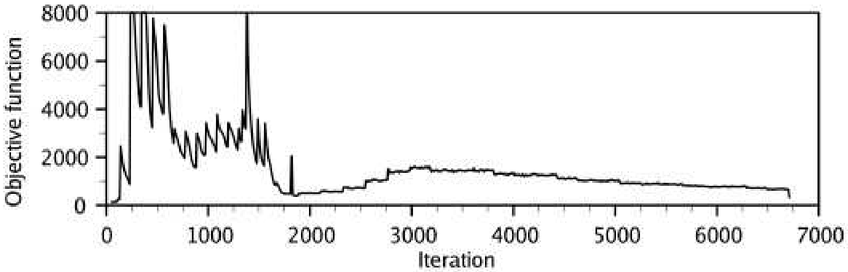


Figure 5. *Optimization of the objective function (curve) starts with a distorted average of template structures. In this run, the first 2000 iterations correspond to the variable target function method relying on the conjugate gradients technique. This approach first satisfies sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimized.*

4.3 Rosetta docking

The Rosetta software[15] focuses on the prediction and design of protein structures, protein folding mechanisms, and protein-protein interactions. Rosetta has been consistently successful in CASP and CAPRI competitions. Rosetta also addresses aspects of protein design, docking and structure.

RosettaDock predicts the structure of a protein-protein complex from the individual structures of the monomer components.

In the standard protein-protein docking protocol, it starts with two protein structures in space, firstly carry out a very fast but crude search to find a rough shape fit between these two proteins. During the first stage, the proteins are represented by only backbones (which defines the shape) and one pseudo atom for side-chains. Afterwards, side-chain atoms are added back and the docking protocol enters the full-atom refinement stage in which the relative orientation between the two proteins and the detailed side-chain interactions across the interface are optimized simultaneously. Each trajectory will end up with a model with certain docking orientation and also an energy function to rank them. Standard settings were used in the process.

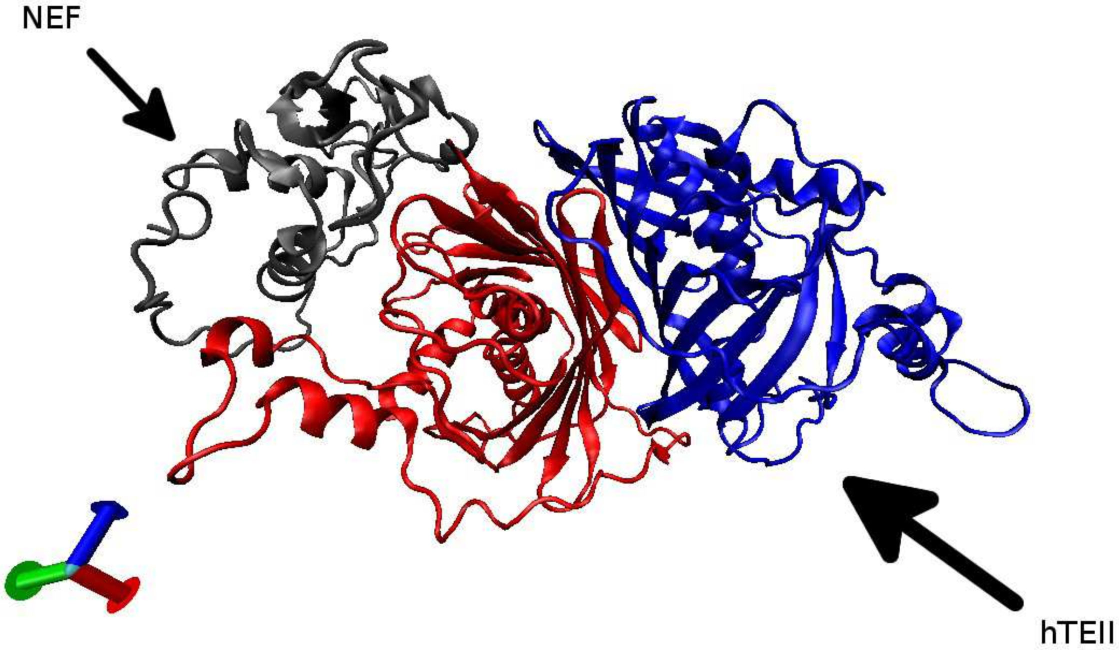


Figure 6. *Representative models of the most populated cluster as obtained from RosettaDock*

RosettaDock Protocol:

a) Nef position randomized

b) distance constraints between Nef residues: D108, F121, P122 and D123 and the hTE8 surface

c) 3000 decoys produced and hierarchically clustered with a cut-off RMSD value of 2.5 Å

4.4 LovoAlign

LovoAlign is a new protein structural alignment package. The methods used for structural alignment are based on Low Order Value Optimization (LOVO) theory.

The use of LOVO theory led to the development of fast convergent algorithms that provide very robust optimization of scoring functions[16]. The structural alignment is highly customizable and the package can be used for general structural alignments or particular chains of each protein may be selected.

The goal of the algorithm is to maximize a scoring function with a solid convergence properties. This is useful for the refinement of protein folding maps, and for the development of new scores designed to be correlated with functional similarity.

The maximization of scoring functions in protein alignment is interpreted as a Low Order Value Optimization (LOVO) problem. The resulting algorithms are convergent and increase the scoring functions at every iteration. The solutions obtained are critical points of the scoring functions. Two algorithms are introduced: One is based on the maximization of the scoring function with Dynamic Programming followed by the continuous maximization of the same score, with respect to the protein position, using a smooth Newtonian method. The second algorithm replaces the Dynamic Programming step by a fast procedure for computing the correspondence between Cα atoms. The algorithms are shown to be very effective for the maximization of the STRUCTAL score.

4.5 Clustering

The last step was to execute a home-made script written in Python, a high –level programming language[17] whose design philosophy emphasizes code readability.
This script, called Clustering.py, takes as input the list of all pdb's to clusterize and an align.log file taken from the previous step, using LovoAlign. The script builds as many clusters as the user wants and sort them ascending, and calculates the centroid of every cluster and all the models that are contained within. The utility of this step is to highlight the representative models (the one that contains more cluster for example) to make the research easier and give more sensibility to the problem.
Finally, all the visual analysis and the figures were produced with the program VMD (Visual Molecular Dynamics) [18].

4.6 VMD: Visual Molecular Dynamics

VMD is a molecular graphics program designed for modeling, visualization, and analysis of biological systems such as proteins, nucleic acids, lipid bi- layer assemblies, etc. It may be used to view more general molecules, as VMD can read standard Protein Data Bank (PDB)files and display the contained structure. VMD provides a wide variety of methods for rendering and coloring a molecule. VMD can be used to animate and analyze the trajectory of a molecular dynamics (MD) simulation. In particular, VMD can act as a graphical front end for an external MD program by displaying and animating a molecule undergoing simulation on a remote computer. All protein figures in this document were created using this computer program .

4.7 NAMD: Scalable Molecular Dynamics

NAMD[18] is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. Simulation of large molecules, however, require enormous computing power. One way to achieve such simulations is to utilize parallel computers. Distributed memory parallel computers have been offering cost-effective computational power. NAMD was designed to run efficiently on such parallel machines for simulating large molecules.

NAMD uses program VMD for simulation setup and trajectory analysis. NAMD has several important features which were used during the simulation:

- Force Field Compatibility: The force field used by NAMD includes local interaction terms consisting of bonded interactions between 2, 3, and 4 atoms and pairwise interactions including electrostatic and Van Der Waals forces.
- Efficient Full Electrostatics Algorithms: NAMD incorporates the Particle Mesh Ewald (PME) algorithm, which takes the full electrostatic interactions into account. This algorithm reduces the computational complexity of electrostatic force evaluation.
- Multiple Time Stepping: The velocity Verlet integration method is used to advance the positions and velocities of the atoms in time. To further reduce the cost of the evaluation of long-range electrostatic forces, a multiple time step scheme is employed. The local interactions (bonded, Van Der Waals and electrostatic interactions within a specified distance) are calculated at each time step. The longer range interactions (electrostatic interactions beyond the specified distance) are only computed less often. This amortizes the cost of computing the electrostatic forces over several time steps.
- Input and Output Compatibility: The input and output file formats include coordinate files in PDB format and structure files in PSF format. Output formats include PDB coordinate files and binary DCD trajectory files.
- Dynamics Simulation Options: MD simulations was carried out using several options, including

- Constant energy dynamics,
- Constant temperature dynamics via,
  + Velocity rescaling,
  + Velocity reassignment,
  + Langevin dynamics,
- Periodic boundary conditions,
- Constant pressure dynamics via,
  + Pressure coupling,
  + Langevin piston,
- Energy minimization,
- Fixed atoms,
- Rigid waters,
- Rigid bonds to hydrogen,
- Harmonic restraints,
- Spherical or cylindrical boundary restraints.

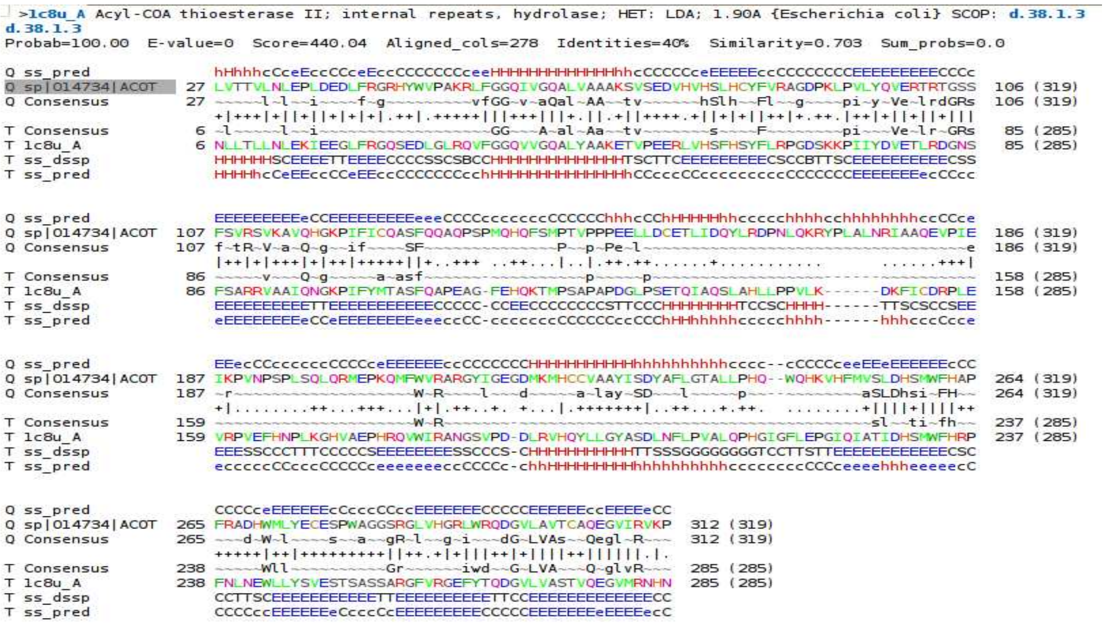
NAMD (CHARMM22 and TIP3P force fields) Protocol:

a) 9200 water molecules for solvatation and PBC
b) Time step : 2 fs
c) PME for electrostatic interactions
d) Constant temperature (300K) and pressure (1 atm)
e) 2000 steps of system minimization: using conjugate gradients
f) 30 ns in molecular dynamics simulation

Part B - Results

1.0 hTE8 model

All members of the hTE8 family were retrieved from the Uniprot [19] database using the program *ssearch*[20]. They were aligned with PROMALS [21]. This multiple sequence alignment was then used for the definition of the Hidden Markov profile (HMM) of hTEII. The latter was then funneled through the Hhsearch [22] program to identify the most plausible homologous structural templates. Such procedure is currently one of the best ones as evaluated from CASP9 experiment [23]. The multiple sequence alignment obtained in this way was used as the reference for the structural prediction of hTE by homology modeling (Figure 7). Homology models of the protein are here based on the crystal structure of the *E.coli* thioesterase (PDB code 1C8U). The sequence alignment between the target and the structural template was extracted from the multiple sequence alignment considering the entire family. We then constructed 50 different conformations of hTE (that were obtained with randomized initial structures and subsequent optimization by conjugate gradients and simulated annealing) based on each of the eight structural templates using Modeller9v3 [25]. All the three dimensional models of hTE obtained in this way do not deviate from currently available experimental geometries, that is, the secondary structures elements (12-stranded antiparallel β-sheets) and the typical tertiary fold (the double hot dog[13]) are also conserved in the model.



```
>1c8u_A Acyl-COA thioesterase II; internal repeats, hydrolase; HET: LDA; 1.90A {Escherichia coli} SCOP: d.38.1.3
d.38.1.3
Probab=100.00  E-value=0  Score=440.04  Aligned_cols=278  Identities=40%  Similarity=0.703  Sum_probs=0.0

Q ss_pred               hHhhhcCceEccCCceEccCCCCCCCceeHHHHHHHHHHHHHHhhcCCCCCceEEEEEccCCCCCCCCEEEEEEEEEECCCc
Q sp|O14734|ACOT    27 LVTTVLNLEPLDEDLFRGRHYWVPAKRLFGGQIVGQALVAAAKSVSEDVHVHSLHCYFVRAGDPKLPVLYQVERTRTGSS  106 (319)
Q Consensus        27 ~~~~~l~l~~i~~~~f~g~~~~~~~~~~vfGG~v~aQal~AA~~tv~~~~~~hSlh~~Fl~~g~~~~pi~y~Ve~lrdGRs  106 (319)
                        +|+++|+||+||+|+|+|.++|.+++++|||+++|||+.||.+||++++.+||+|+||++|+.++.|++|+||+||+|||
T Consensus         6 ~l~~~~~l~~i~~~~~~~~~~~~~~~~~~~~GG~~~A~al~Aa~~tv~~~~~~~s~~~~F~~~~~~~~~pi~~~Ve~lr~GRs   85 (285)
T 1c8u_A            6 NLLTLLNLEKIEEGLFRGQSEDLGLRQVFGGQVVGQALYAAKETVPEERLVHSFHSYFLRPGDSKKPIIYDVETLRDGNS   85 (285)
T ss_dssp             HHHHHHSCEEEEETTEEEEECCCCSSCSBCCHHHHHHHHHHHHHHHTSCTTCEEEEEEEEEECSCCBTTSCEEEEEEEEEEECSS
T ss_pred              HHHHhcCeEEccCCeEEccCCCCCCCcchHHHHHHHHHHHHHHHHhCCccccCCccccccccCCCCCCCEEEEEEEecCCcc

Q ss_pred               EEEEEEEEEEeCCEEEEEEEEEEEeeeCCCCcccccccCCCCCChhhcCChHHHHHhhcccccchhhhcchhhhhhhhccCCce
Q sp|O14734|ACOT   107 FSVRSVKAVQHGKPIFICQASFQQAQPSPMQHQFSMPTVPPPEELLDCETLIDQYLRDPNLQKRYPLALNRIAAQEVPIE  186 (319)
Q Consensus       107 f~tR~V~a~Q~g~~if~~~~SF~~~~~~~~~~~~~~~~~P~~p~Pe~l~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~e  186 (319)
                        |++|+|+++|+|++|+++++||+..+++ ..++...|..|.++.++......+...........       ......+++|
T Consensus        86 ~~~~~~v~~~Q~g~~~~~~a~asf~~~~~~~~~~~~~~~~~~~~p~~~~~p~~~~~~~~~~~~~~~~~~~-~~~~~~~~~~~~~~  158 (285)
T 1c8u_A           86 FSARRVAAIQNGKPIFYMTASFQAPEAG-FEHQKTMPSAPAPDGLPSETQIAQSLAHLLPPVLK------DKFICDRPLE  158 (285)
T ss_dssp             EEEEEEEEEEETTEEEEEEEEEEEEECCCCC-CCEECCCCCCCCSTTCCCHHHHHHHHHTCCSCHHHH------TTSCSCCSEE
T ss_pred              eEEEEEEEEEeCCeEEEEEEEEEEeeeccCC-cccccccCCCCCCccCCChHHhhhhhcccccchhhh------hhhcccCcce

Q ss_pred               EEecCCcccccccCCCCceEEEEEEEccCCCCCCCHHHHHHHHHHHHhhhhhhhhhhhcccc--cCCCCCceeEEeEEEEEEEcCC
Q sp|O14734|ACOT   187 IKPVNPSPLSQLQRMEPKQMFWVRARGYIGEGDMKMHCCVAAYISDYAFLGTALLPHQ--WQHKVHFMVSLDHSMWFHAP  264 (319)
Q Consensus       187 ~r~~~~~~~~~~~~~~~~~~~~~~W~R~~~~l~~~d~~~~~~a~lay~SD~~~l~~~~~~p~~~~~~~~~~~~~~aSLDhsi~FH~~  264 (319)
                        +|.........++...+++...|+|.++..+. +...|.+++++++|..++...+.++.  .........+||||+|||||++
T Consensus       159 ~~~~~~~~~~~~~~~~~~~~~~~~~~W~R~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~sl~~ti~fh~~  237 (285)
T 1c8u_A          159 VRPVEFHNPLKGHVAEPHRQVWIRANGSVPD-DLRVHQYLLGYASDLNFLPVALQPHGIGFLEPGIQIATIDHSMWFHRP  237 (285)
T ss_dssp             EEESSCCCTTTCCCCCSEEEEEEEEESSCCCS-CHHHHHHHHHHHHTTSSSGGGGGGGTCCTTSTTEEEEEEEEEEEEEEECSC
T ss_pred              eccccccCCcccCCCCCceeeeeeeccCCCCc-chhHHHHHHHHHHhhhhhhhhhhhccccccccCCCceeeehhheeeeecC

Q ss_pred               CCCCceEEEEEEEcCccccCCccEEEEEEECCCCCEEEEEEEccEEEEeCC
Q sp|O14734|ACOT   265 FRADHWMLYECESPWAGGSRGLVHGRLWRQDGVLAVTCAQEGVIRVKP  312 (319)
Q Consensus       265 ~~~d~W~l~~~~~s~~a~~gR~l~~g~i~~~dG~LVAs~~Qegl~R~~~  312 (319)
                        +++++|++|+++++++++||++.+|+|||++|+||||++||||||.|.
T Consensus       238 ~~~~~~Wll~~~~~~~~~~~~Gr~~~~~~~iwd~~G~LVA~~~Q~glvR~~~  285 (285)
T 1c8u_A          238 FNLNEWLLYSVESTSASSARGFVRGEFYTQDGVLVASTVQEGVMRNHN  285 (285)
T ss_dssp             CCTTSCEEEEEEEEEEEETTEEEEEEEEEEEETTCCEEEEEEEEEEEEEECC
T ss_pred              CCCCccEEEEEEeCccccCcEEEEEEEEEECCCCCEEEEEEEeEEEEecC
```

Figure 7. *Hhpred alignment between the sequences of thioesterase from E.coli and from H. sapiens.The secondary structure elements are detailed (E and H indicates β-strands and α-helices, respectively)*

2.0 Nef/hTE Complex

The complex between the the model and Nef was built by the use of the following protocol: 2500 hTE/NEF adduct structures were constructed using Rosetta-dock. A standard Lamarckian Genetic Algorithm, was used for conformational exploration with a rapid energy evaluation using grid-based molecular affinity potentials. The resulting structures were then clustered according to the three dimensional localization of Nef, regardless of the docking energies. Then LovoAlign program was used over all the models to obtain optimal structural superposition.

At the end a clustering Python home-written script was used for selecting the best decoys. The script gives as output the decoys representing the most populated clusters. In detail, the clusters were formed for decoys not deviating from the other members for more than 4Å of Root Mean Square Deviation (RMSD).

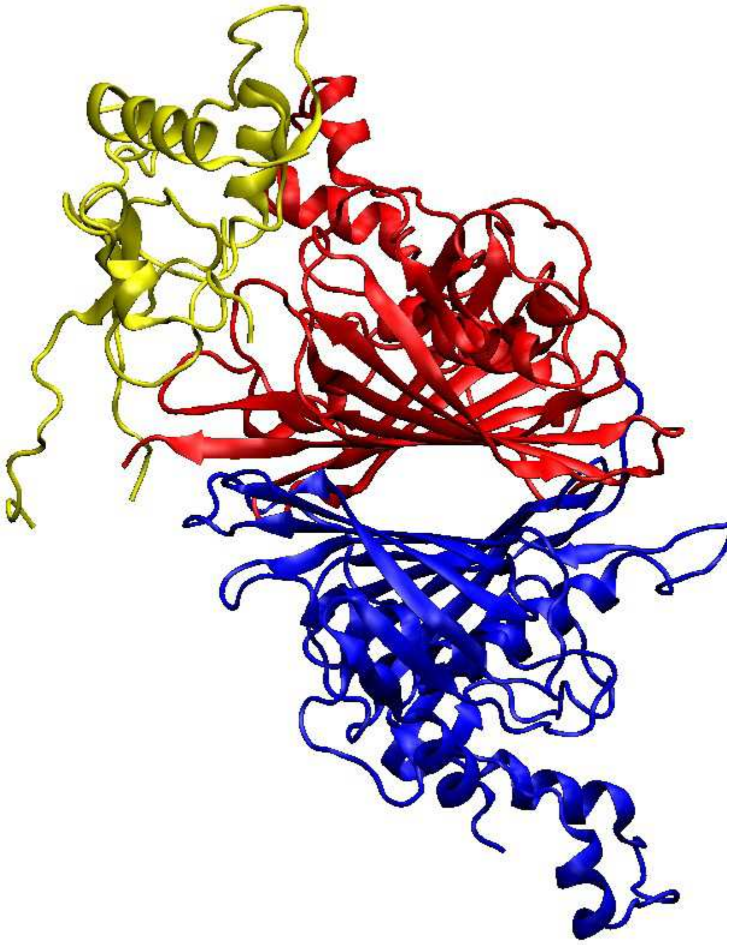
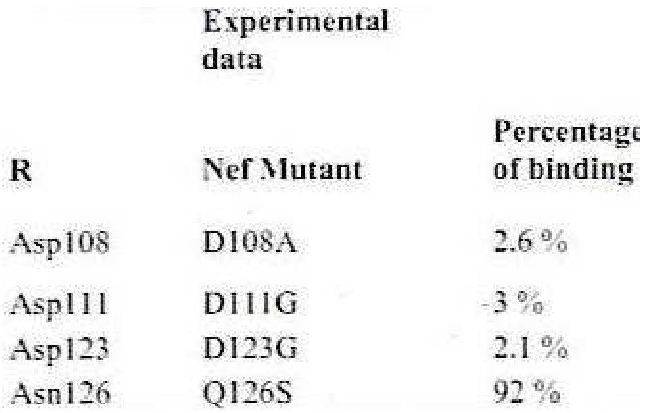


Figure 8. *Nef/hTE Complex*

The obtained representatives were selected among those that better satisfied the experimental restraints introduced in the docking procedure, that is, we have chosen the model that performs more contacts for the residues known to form part of the NEF interaction surface, i.e. Asp108, Asp123, Phe121 and Pro122 as in Figure 9 where the interacting details can be appreciated.



	Experimental data	
R	Nef Mutant	Percentage of binding
Asp108	D108A	2.6 %
Asp111	D111G	3 %
Asp123	D123G	2.1 %
Asn126	Q126S	92 %

Figure 9. *Mutagenesis data by Liu et al. The content of complexed protein relative to wild type (100%) is reported.*

In particular, several partners were identified showing electrostatic complementarity between both interacting surfaces. Furthermore, several hydrophobic contacts are formed between aromatic and aliphatic residues of Nef (Phe121, Pro 122) as well as aromatic(pyrrolic) and aliphatic(4methylated base) (Pro320, Lys322) residues of hTE, allowing a further stabilization of the interaction. In Figures 11, 12 the Nef/hTE electrostatic interactions can be visualized by a plot of the electrostatic potential calculated solving the Poison-Boltztmann equation. Figures 11, 12 shows that the contact surfaces of Nef and hTE subunit B are clearly complementary: while the surface of NEF is highly negative, the contact surface of the enzyme is highly positively charged. hTE's surface includes residues Lys361, Lys322, and no negatively charged residues. That of Nef is negative charged, containing residues Asp108, Asp123, and no positive residues.

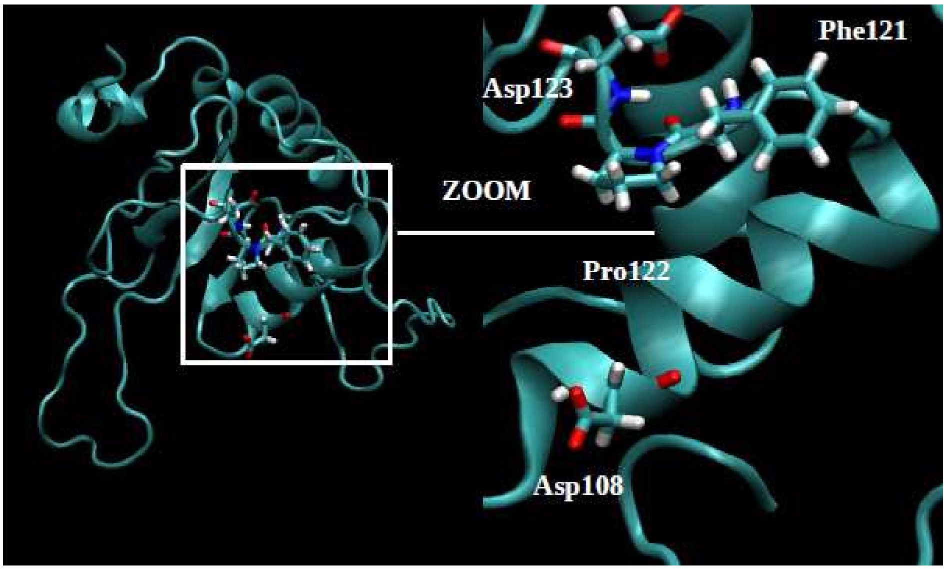


Figure 10. *Nef residues involved in the binding site*

List of contacts						
	Nef	**hTE (Subunit A)**	**Bond**	**%TIME <5Å**	**DIST(Å)**	**DEV.ST (Å)**
1	ARG105	ASP434(ASP143)	(SALTBRIDGE)	97%	2,74	1,04
2	ARG106	ASP434(ASP143)	(SALTBRIDGE)	50%	5,49	1,71
3	**ASP108**	**LYS322(LYS31)**	(SALTBRIDGE)	59%	4,56	1,33
4	ASP111	LYS322(LYS31)	(SALTBRIDGE)	74%	4,69	1,19
5	**PHE121**	**PRO320(PRO29)**	(AROMATIC+PYRROLIC)	43%	5,65	1,64
6	**PRO122**	**LYS322(LYS31)**	(ALIPHATIC+BASE)	58%	5,18	1,64
7	**ASP123**	**LYS361(LYS70)**	(SALTBRIDGE)	68%	3,89	1,33
8	GLU151	LYS581(LYS90)	(SALTBRIDGE)	79%	4,32	2,06
9	GLU201	ARG356(ARG65)	(SALTBRIDGE)	94%	3,02	1,50

Table I - *List of contacts involved in the binding site. Each contact was selected starting from the information given by the Nef mutagenesis and selecting all hTE's residues at a 5 Å distance along the entire 30 ns MD simulation (for details see below) of the Nef/hTE8 complex. Selected distances and residence times of residues at the interface between the viral factor and the subunit A of the enzyme are reported*

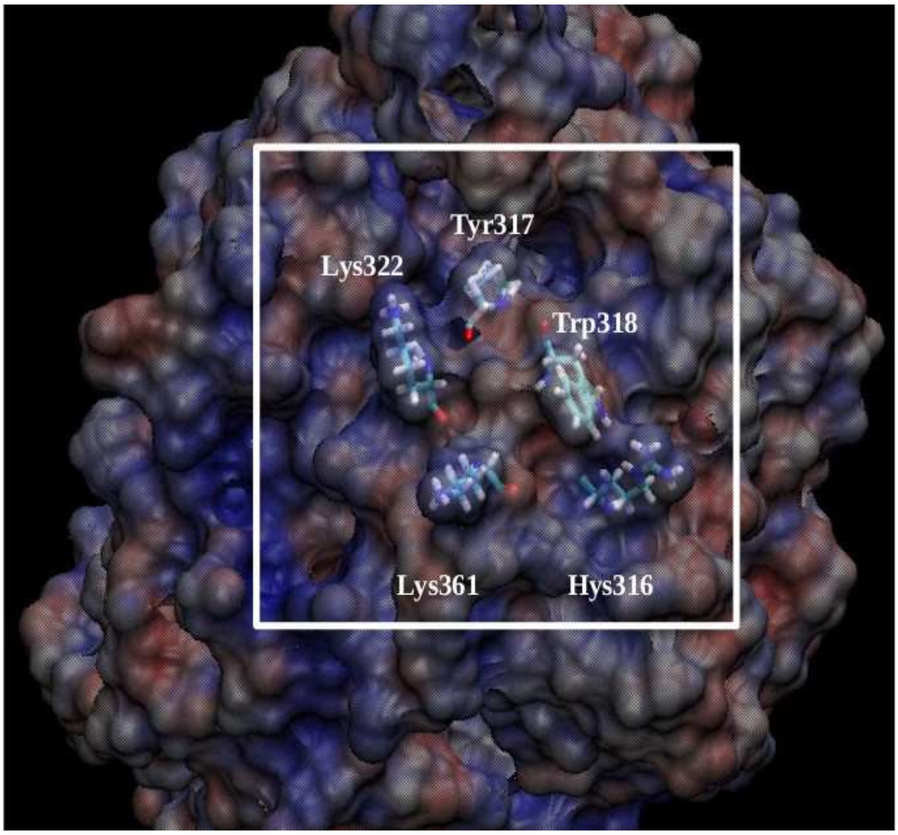


Figure 11. *Electrostatic potential of the hTE8 in the initial conformation as obtained from the docking procedure.*

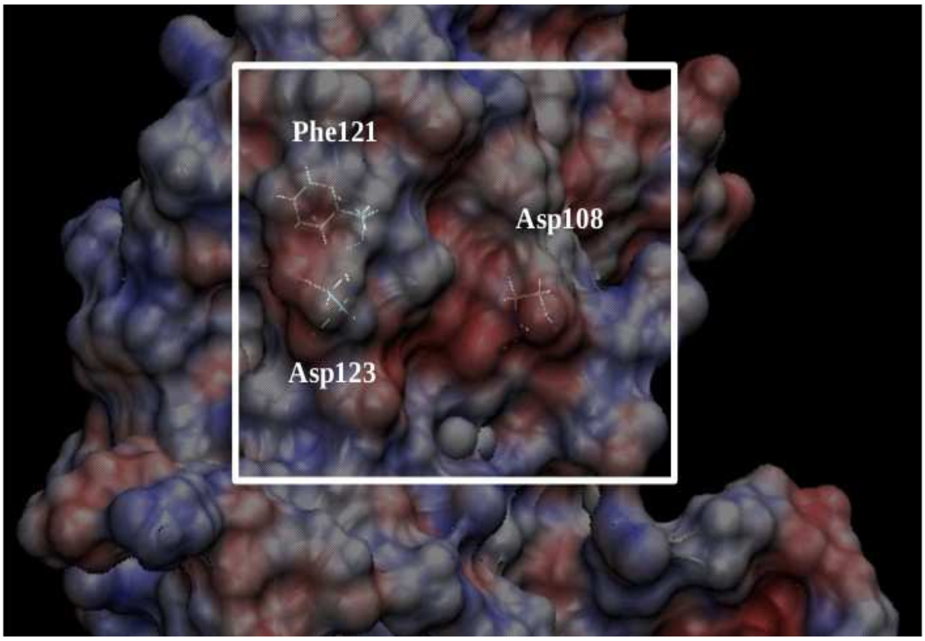


Figure 12. *Electrostatic potential of the Nef in the initial conformation as obtained from the docking procedure***.**

3.0 MD simulation

To test the stability of the docked complex and to gain insights into the dynamical properties of the hTE-Nef complex we have performed extensive molecular dynamics simulations of the solvated system. The latter simulations were carried out for 30ns. In the following paragraphs we will analyse the results obtained for the protein and the complex.

From the plot showing the evolution of the secondary structure elements along the entire MD simulation (Fig. 13) we can appreciate that the principal secondary structure elements are conserved through the entire simulation time (α-helix and β-strands in pink and yellow bands, respectively, in Fig 13).

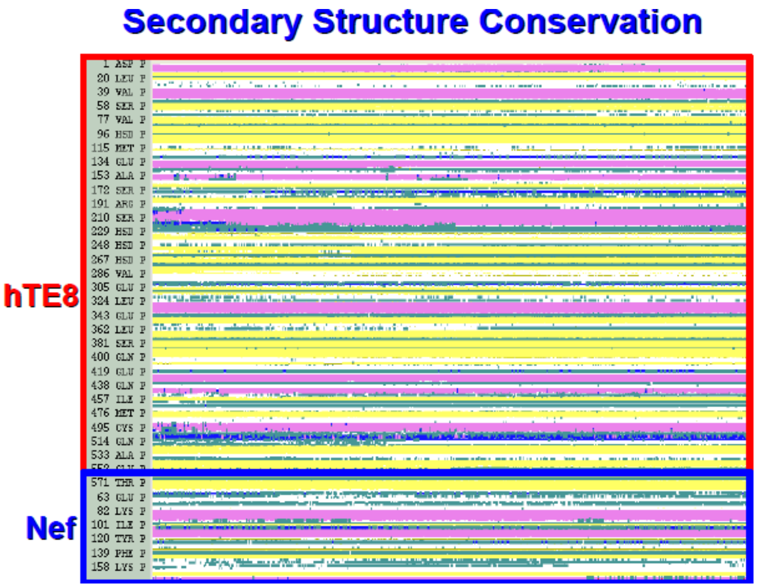


Figure 13. *Timeline continuity of the protein secondary structure elements*

The active site conformation is also rather preserved, the overall RMSD of residues Asp232, Ser234, Gln303 (the catalytic triad) between the initial and final minimized structures being as small as 0.4Å.

The interface between the two subunits(A and B) is totally buried in the protein. It involves the central fragments of the six central β-sheets from the two monomers to form several stabilizing interactions. The contact surface is structurally similar to that of the *E.Coli* enzyme (1980 and 2142 Å^2 per monomer for the enzyme from MD final structure and for that from *E.Coli*, respectively)

Our calculations suggest that also long-range subunit/subunit electrostatic dipole/dipole interactions stabilize the dimer , in part counterbalancing the charge/charge repulsion.

The largest scale motion of the protein in the multi ns timescale, here studied by diagonalization of the covariance matrix, involve essentially only the loop formed by residues 163-194 in both subunits (unit A: residues 454 to 485) ('active site loop'). As expected from these data, the analysis of the RMSF (Fig.14) calculated on each residue show that residues that experience the largest deviations are the ones forming the 'active site loop'.

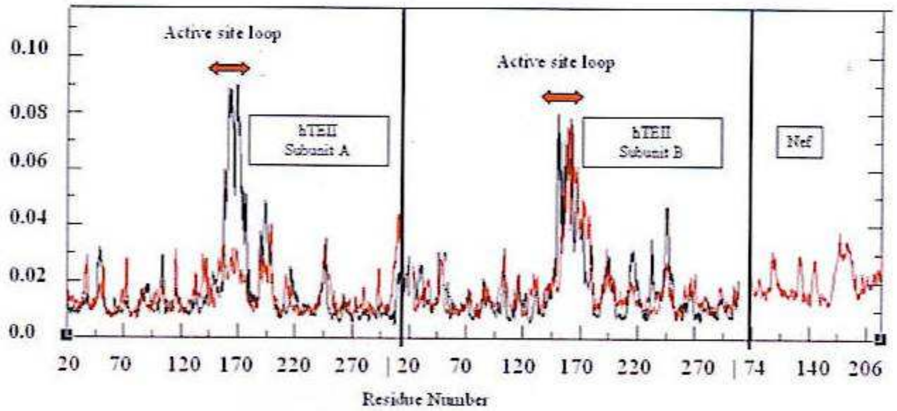


Figure 14. *RMSF value per residue***.** *Calculated for each atom backbone for hTE8 alone (black line), and for the Nef/hTE complex (red line*

The complex appears to be fairly equilibrated after ~15 ns, as shown by the plot of the RMSD deviation (Fig.15) from the energy-minimized structure as a function of time.

As expected, besides being stabilized by long range electrostatics (the net charges of Nef and hTE8 are -4 and +6, respectively), the complex is stabilized by electrostatic and hydrophobic interactions between residues located at the protein/protein interface, which were well-mantained during the dynamics.

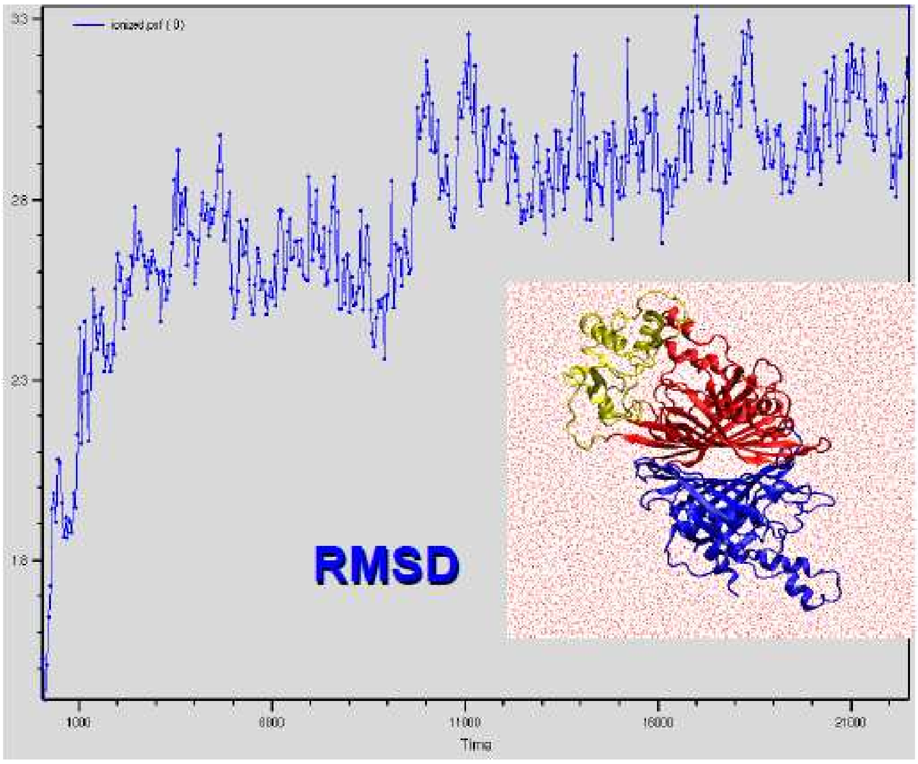


Figure 15. *Root Mean Square Deviation (RMSD) for the Nef-hTE8 complex along the entire 30ns MD simulation*

The hydrophobic contacts (Fig. 23-24) are formed by pyrrolic residues of hTE8 (Pro320) and 4-methylated base (Lys322) with aromatic (Phe121) and aliphatic (Pro122) residues of Nef respectively, and are conserved during the entire simulation.

Nef/hTE8 local electrostatic interactions can be vividly visualized by plot of the electrostatic potential calculated solving the Poisson-Boltzman equation (Fig.11-12).
The contact surface of the enzyme is highly positively charged This surface includes 4 positively charged (Lys322, Lys361, Lys581, Arg356) and only 1(Asp434) negatively charged residue.
That of Nef is negatively charged, containing five negative residues (Asp108, Asp123, Asp111,Glu151, Glu201) and only two positive residues (Arg105, Arg106).

In particular, the three Asp and two Glu groups on Nef surface provide a largely favorable contribution.

Asp108 interacts with Lys322, Asp123 with Lys361; and Asp111 with Lys322 while Glu151 and Glu 201 interact with Lys581 and Arg356 respectively. All the interactions are stable during the entire simulation as can be appreciated from Fig.18-19-20-21-22 (five salt bridges). Besides there are other two salt bridges between two positive Nef residues (Arg105, Arg106) and one negative hTE residue (Asp434) (Fig. 16-17).

We must point out that these values should be taken at a qualitative level as the model used contain implicit uncertainties due to the homology model structure. Here, we hypothesize that its binding to Nef modifies the dynamical properties of the enzyme. Although Nef does not bind to the active site, the 'active site loop' of subunit A (the one that binds to Nef) is much more rigid than that of subunits B, as evidently by an analysis of the large-scale motion of the complex.

These interactions, which are conserved during the entire simulation, might hinder the motion of subunit A. The RMSF fluctuations of the 'active site' loop from subunit A are considerably reduced in presence of Nef whereas those of the rest of the protein are essentially unaffected by the presence of the viral protein. It is important to take into account the conformation of this particular loop, which was modeled with gap of six residues in its middle region.

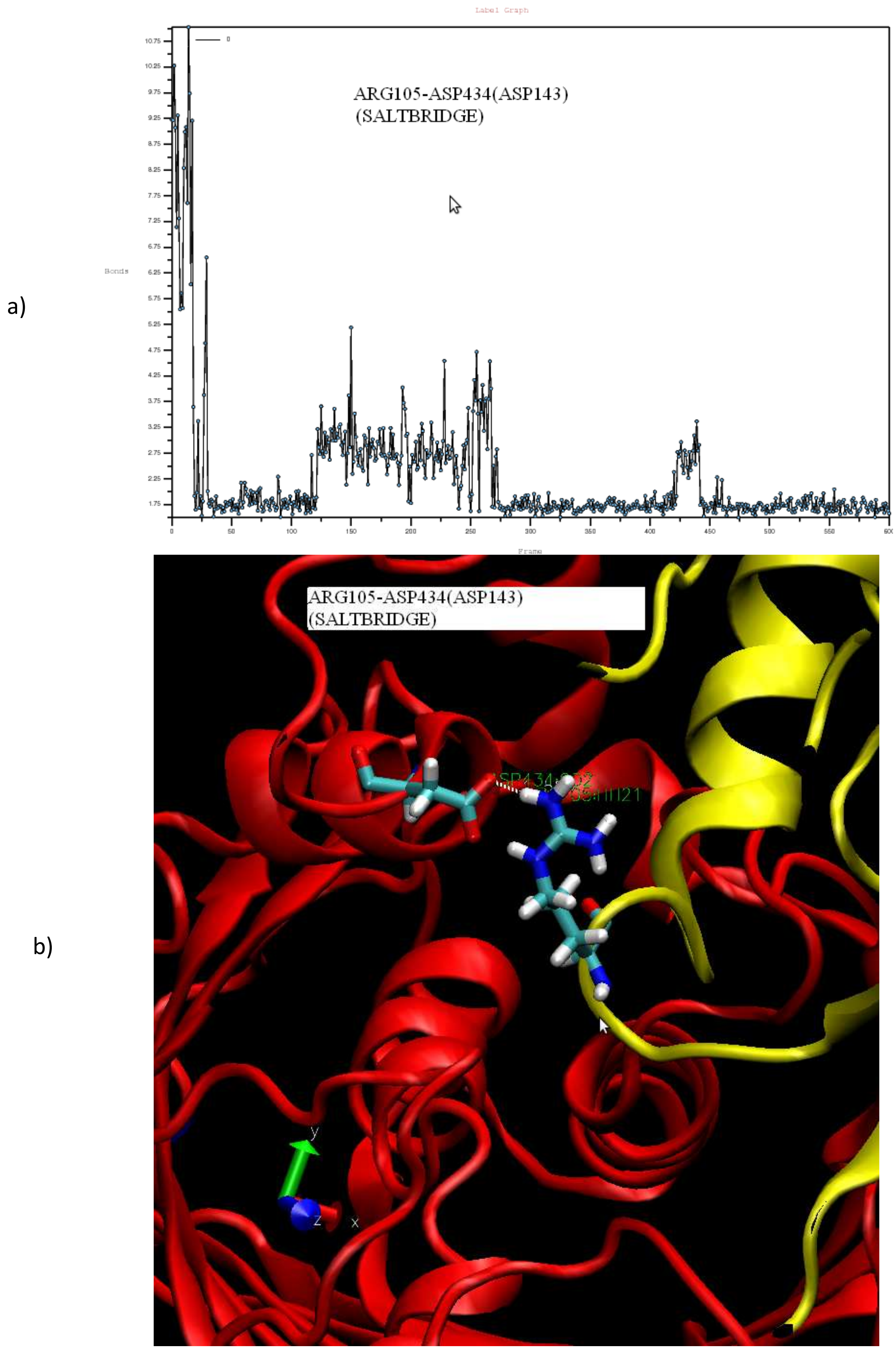


Figure 16. *Detail of the interaction between residues NEF-Arg105 and hTE-Asp434. a) distance along the MD simulation. b) molecular detail*

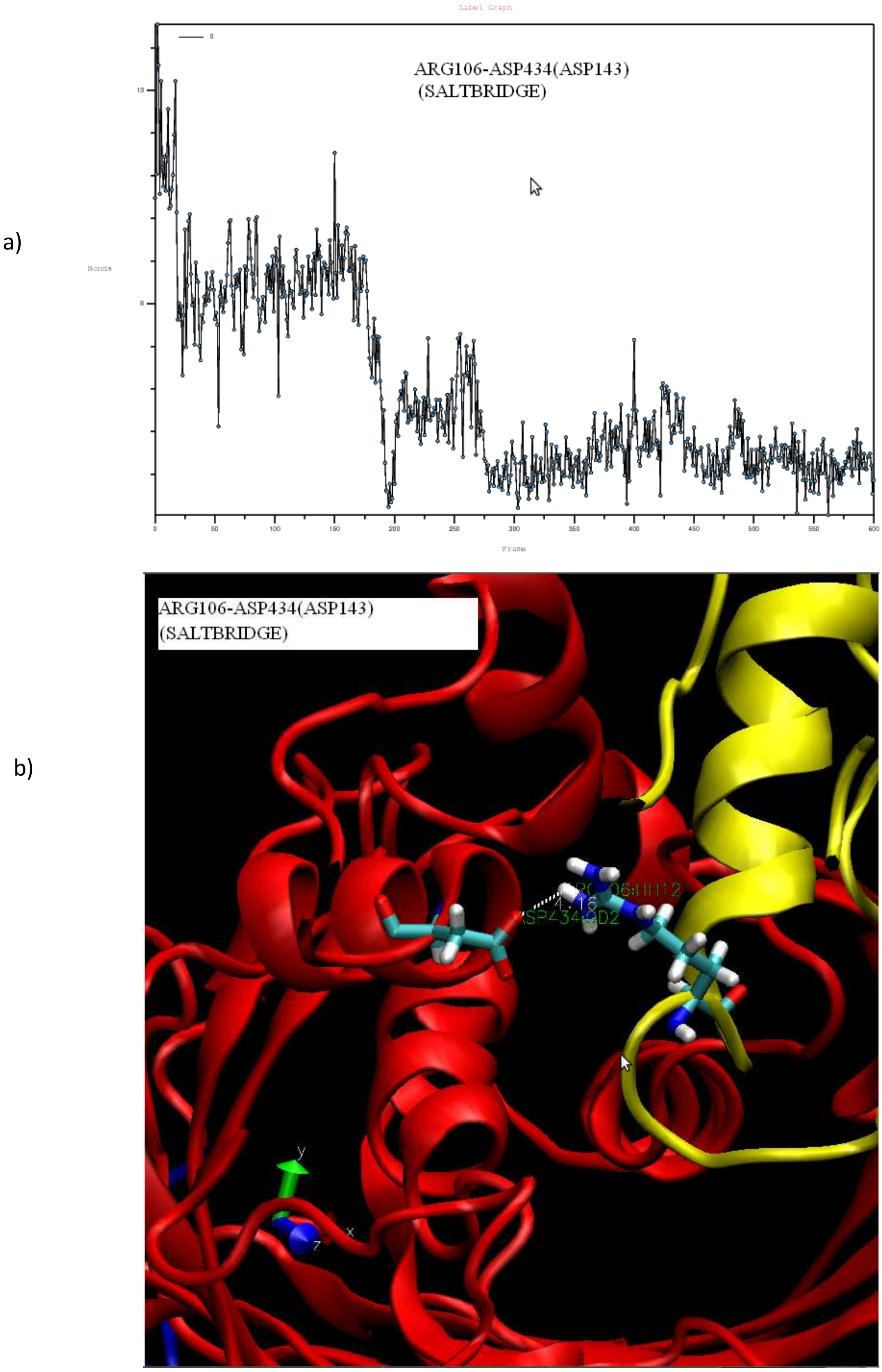


Figure 17. *Detail of the interaction between residues NEF-Arg106 and hTE-Asp434. a) distance along the MD simulation. b) molecular detail*

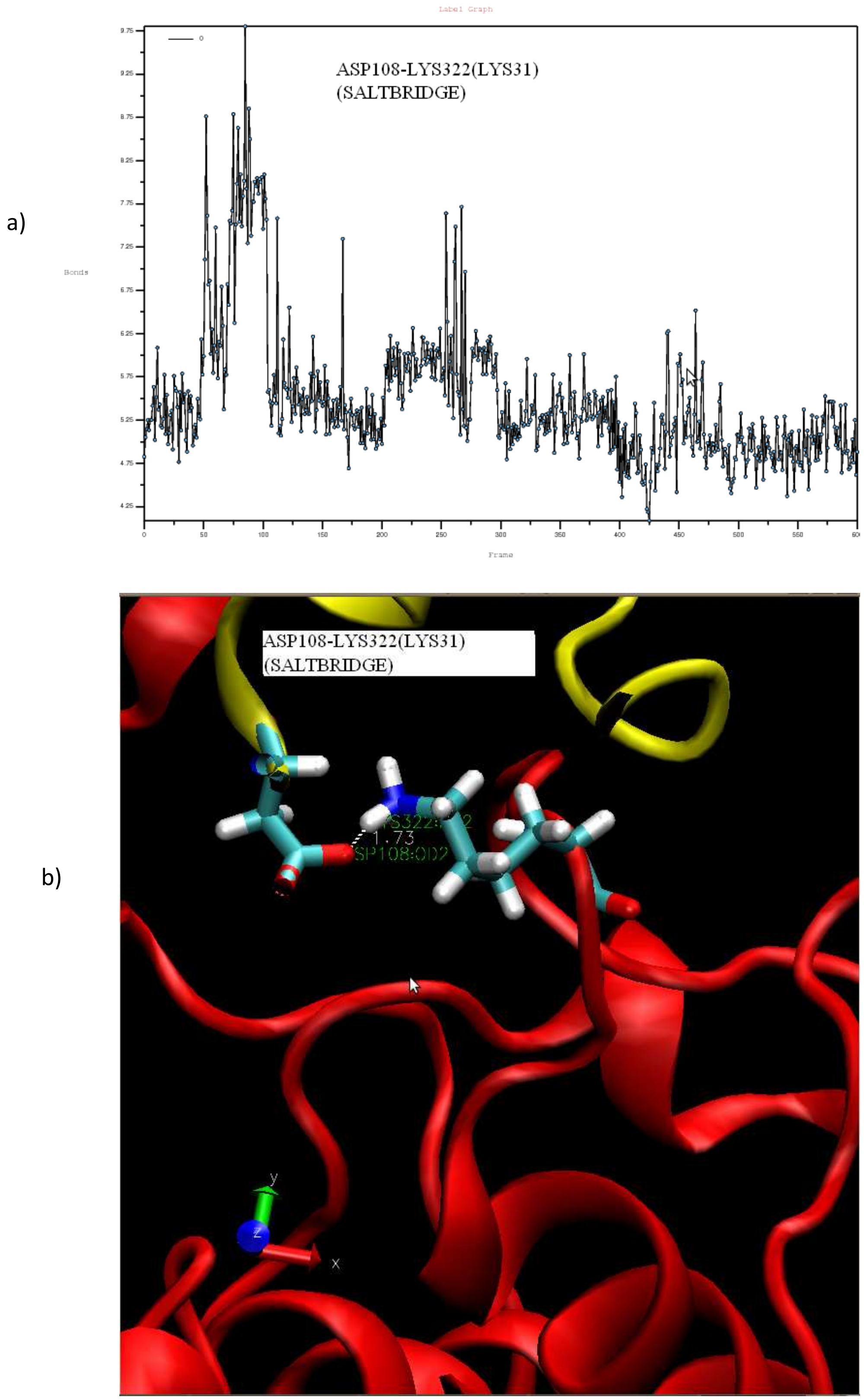


Figure 18. *Detail of the interaction between residues NEF-Asp108 and hTE-Lys322. a) distance along the MD simulation. b) molecular detail*

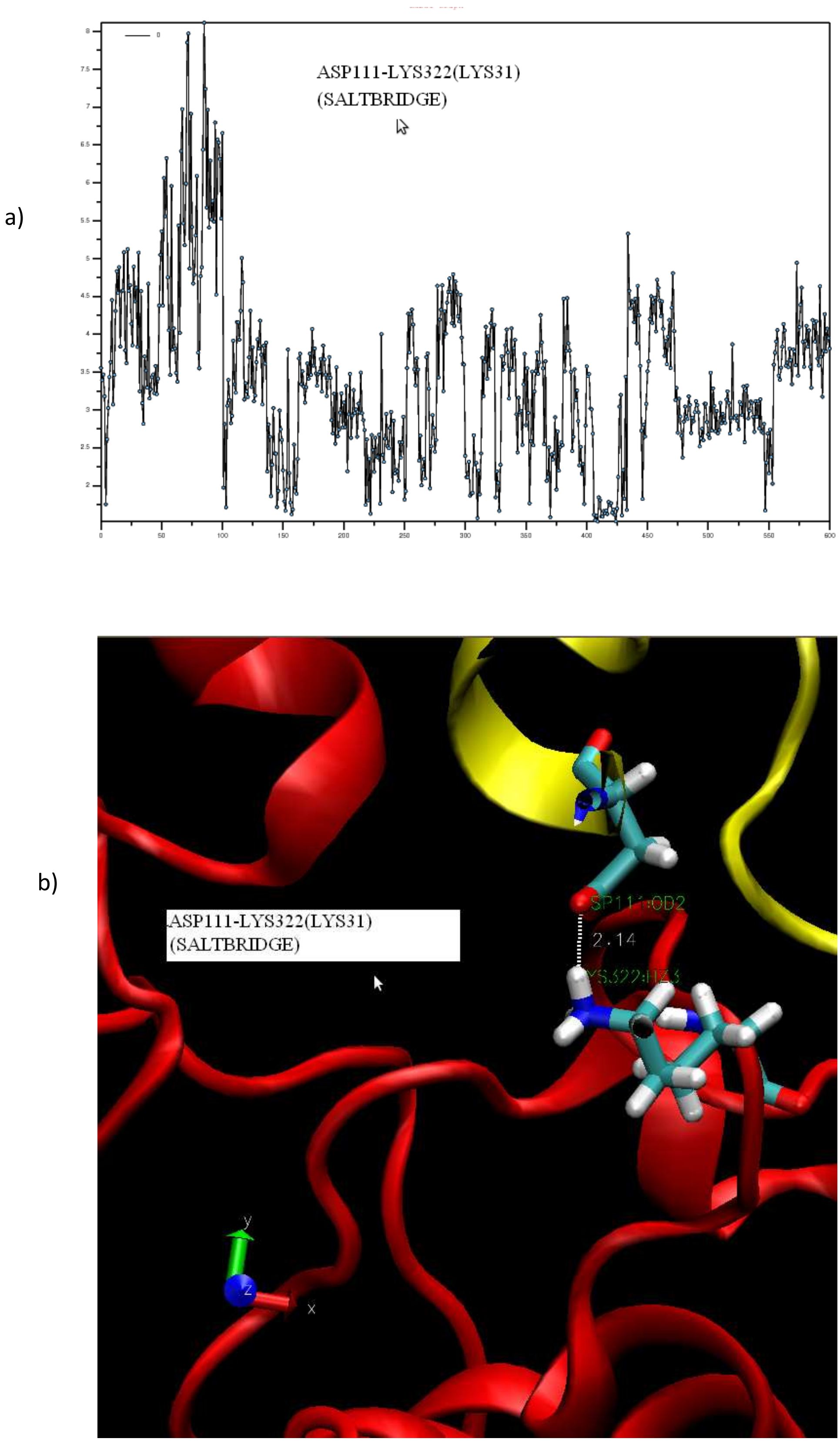


Figure 19. *Detail of the interaction between residues NEF-Asp111 and hTE-Lys322. a) distance along the MD simulation. b) molecular detail*

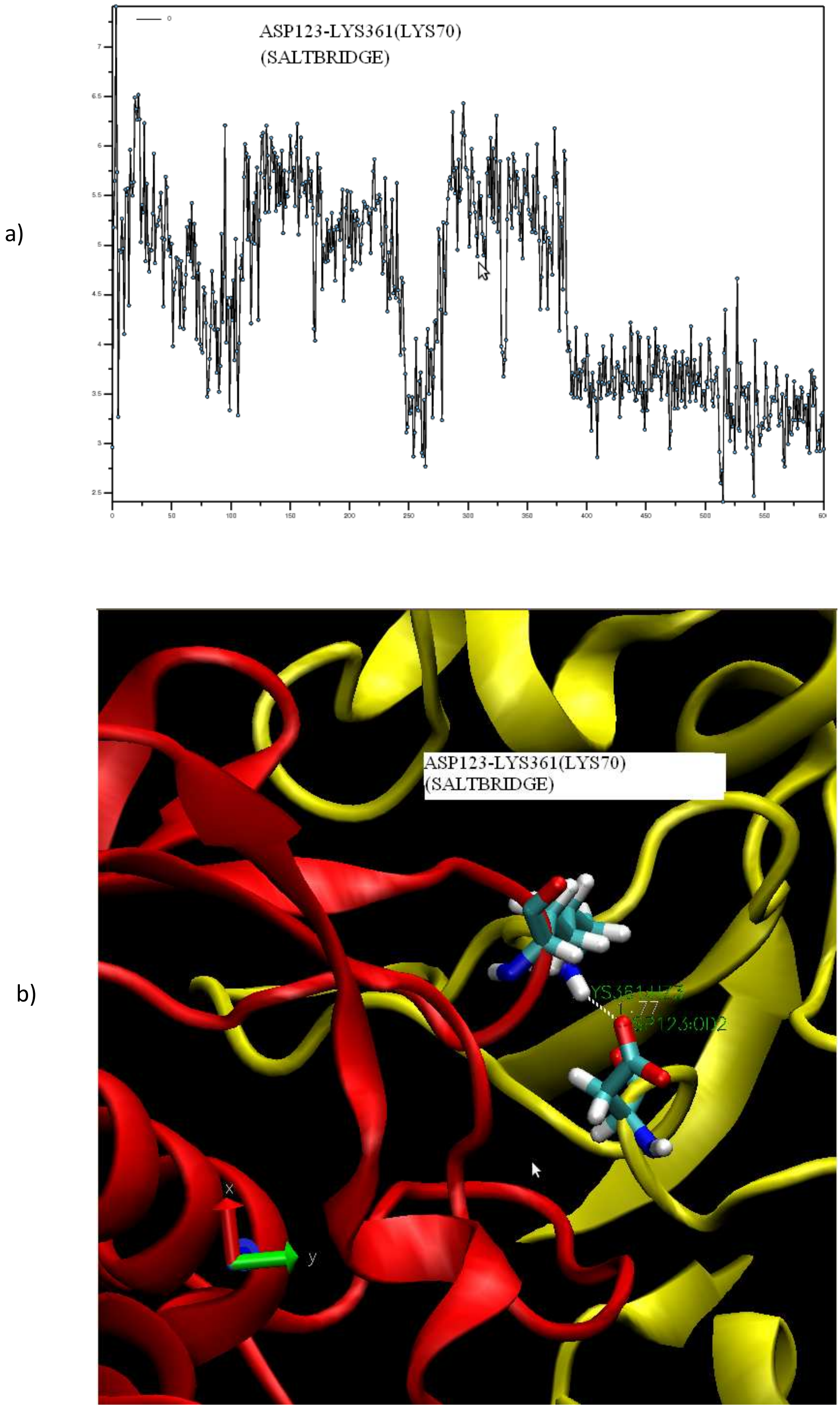


Figure 20. *Detail of the interaction between residues NEF-Asp123 and hTE-Lys361. a) distance along the MD simulation. b) molecular detail*

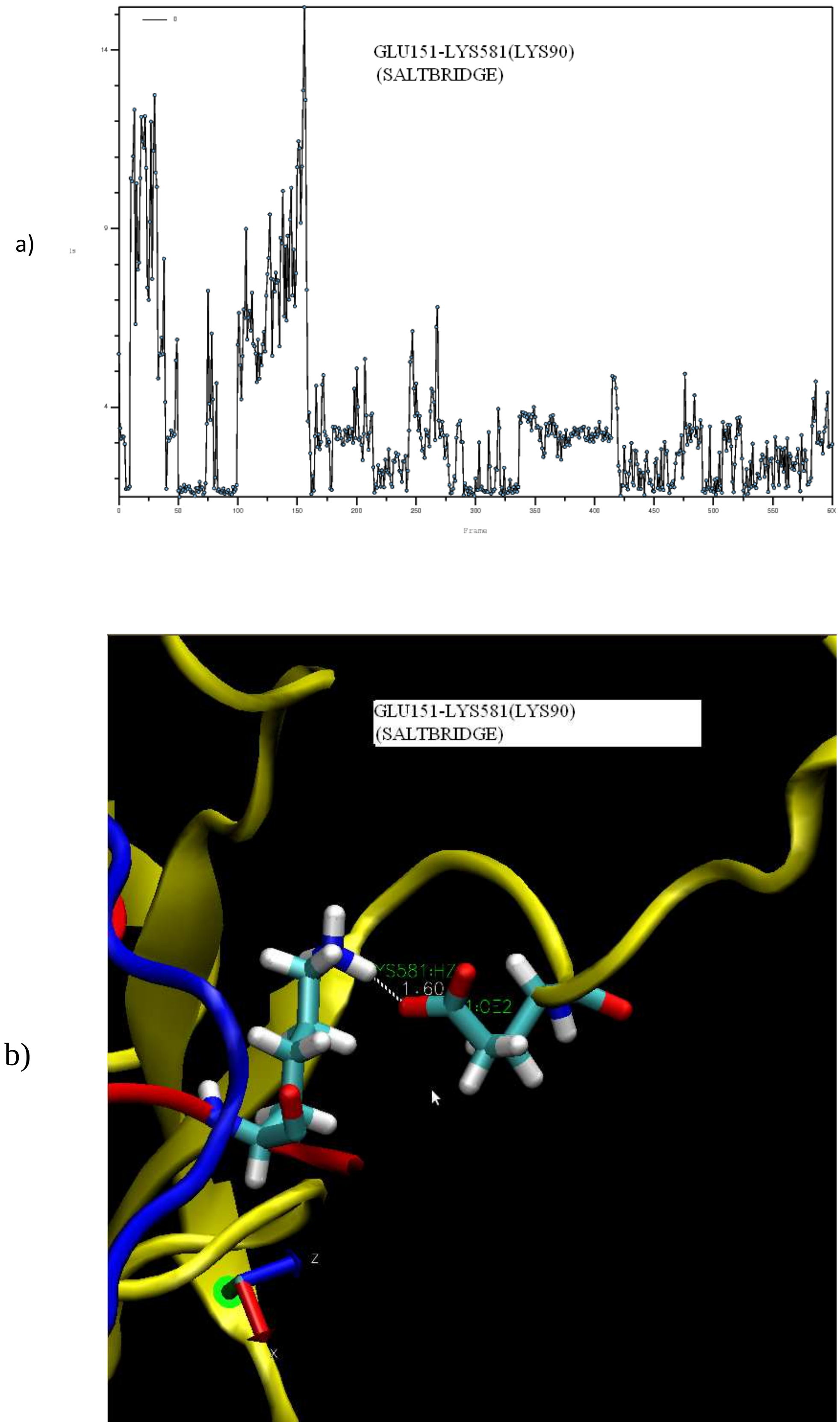


Figure 21. *Detail of the interaction between residues NEF-Glu151 and hTE-Lys581. a) distance along the MD simulation. b) molecular detail*

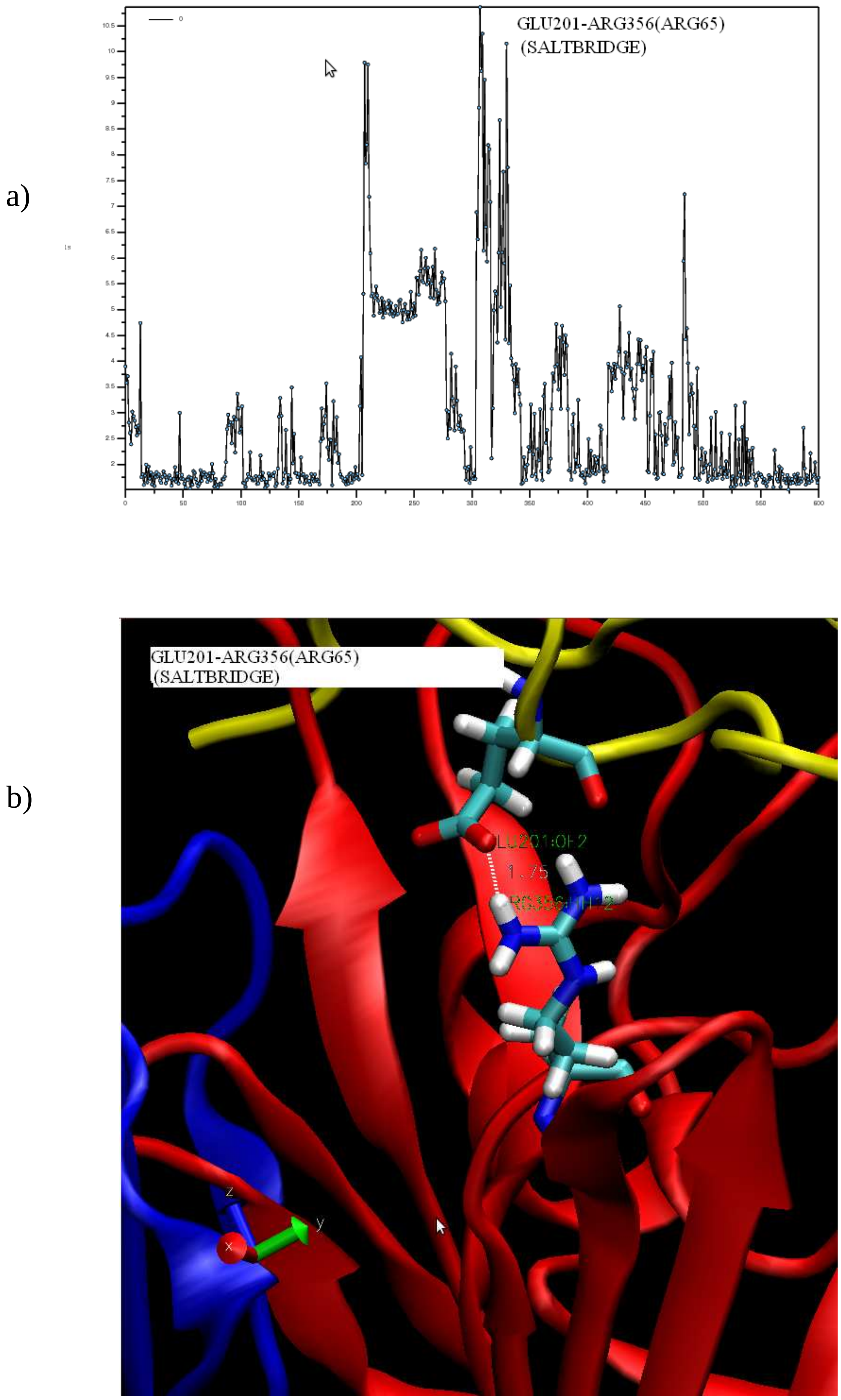


Figure 22. *Detail of the interaction between residues NEF-Glu201 and hTE-Arg356. a) distance along the MD simulation. b) molecular detail*

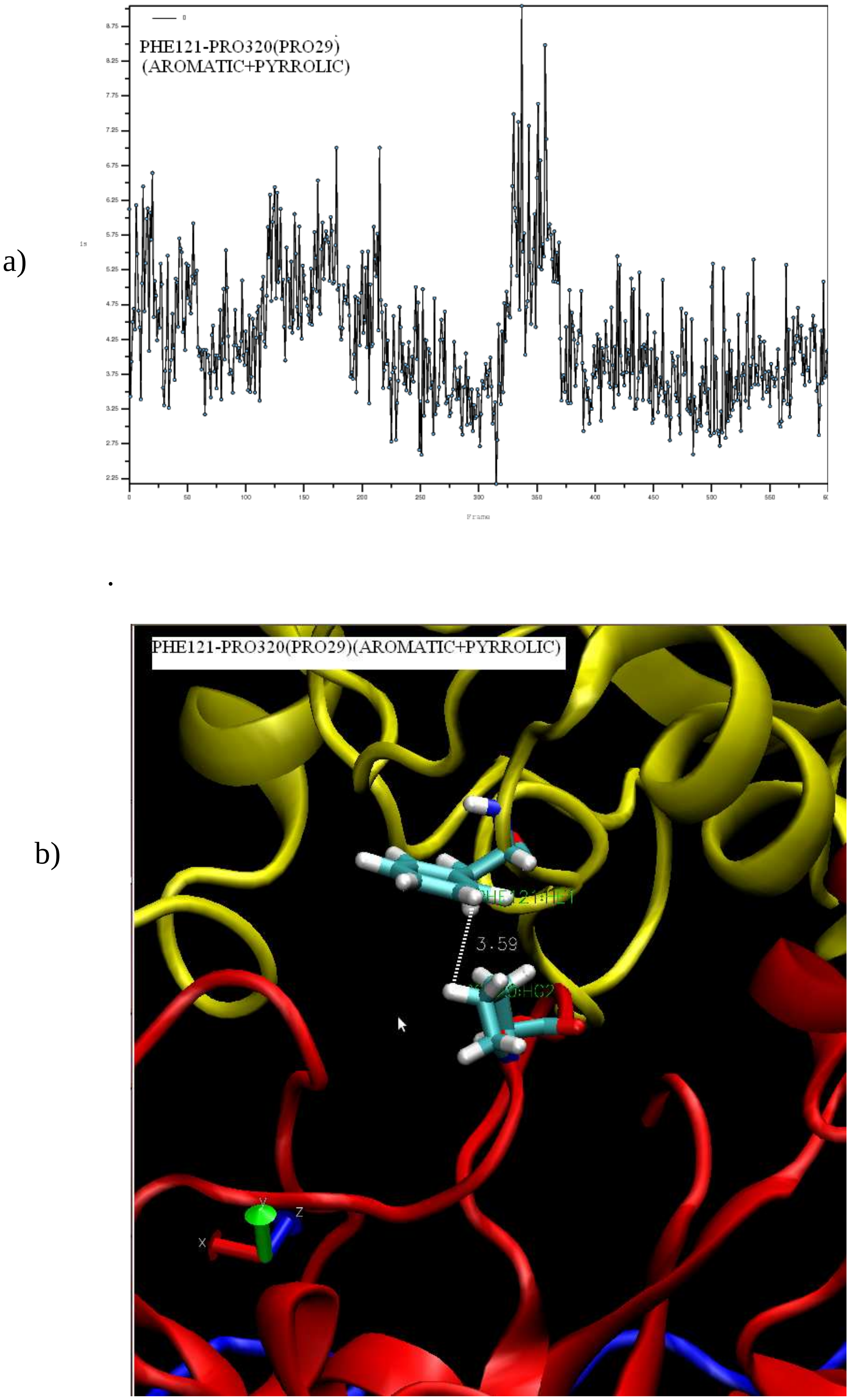


Figure 23. *Detail of the interaction between residues NEF-Phe121 and hTE-Pro320. a) distance along the MD simulation. b) molecular detail*

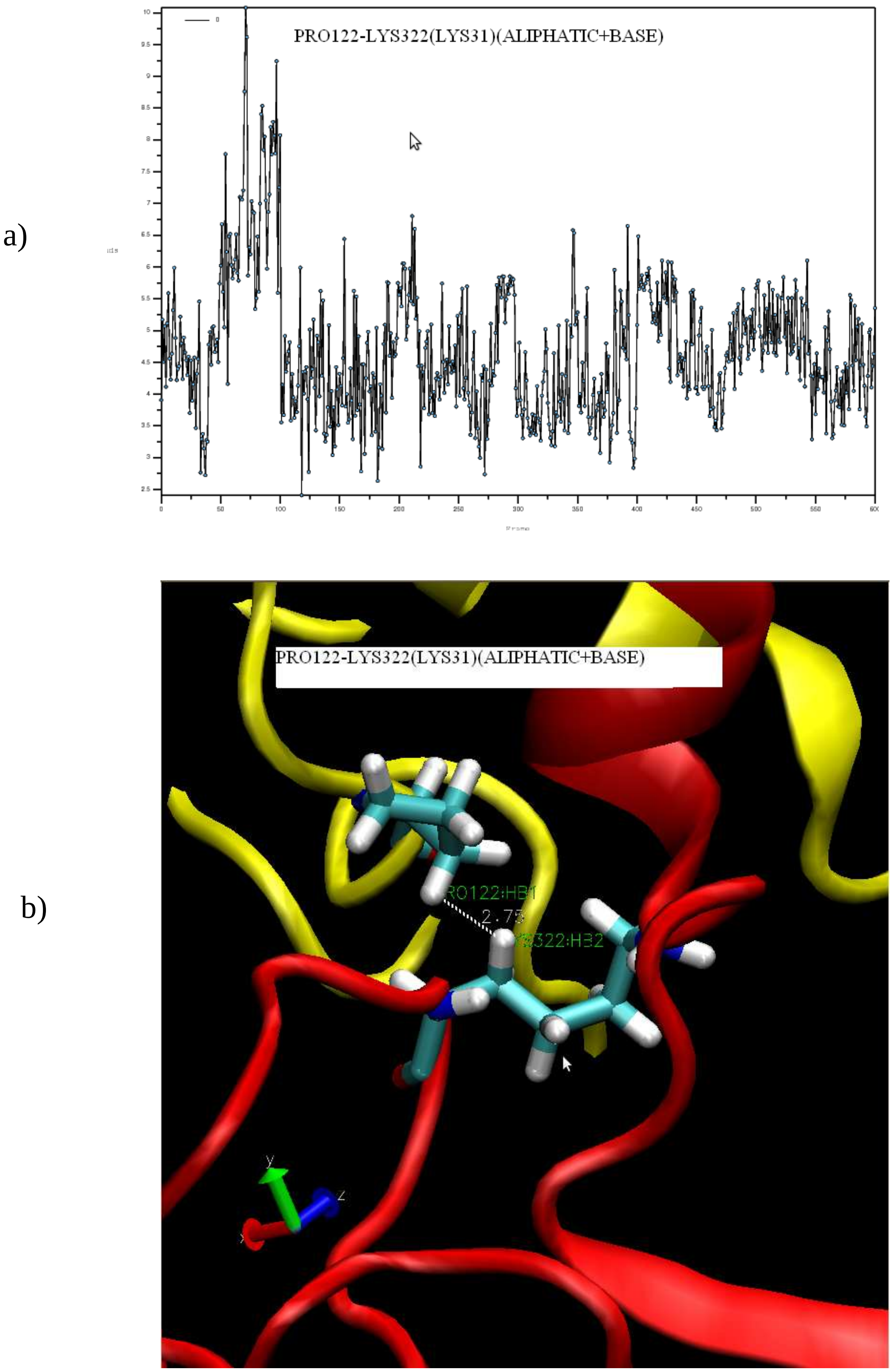


Figure 24. *Detail of the interaction between residues NEF-Pro122 and hTE-Lys322. a) distance along the MD simulation. b) molecular detail*

Discussion

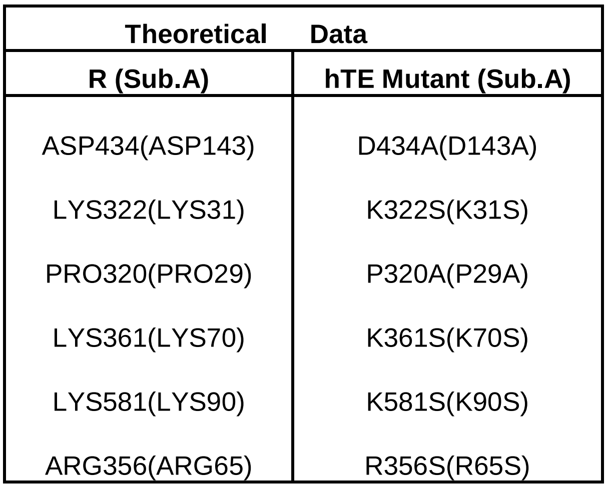
In this work, we have investigated the interactions between the HIV-1 Nef and its cellular partner hTE. We modeled the structure of Human thioesterase based on the isoenzyme from *E.coli*, which share a 42,3% of sequence identity. Long-range dipole/dipole interactions appear to counterbalance, at least in part, the repulsion between the charged subunits. An approximate model of the Nef/hTE complex was built using RosettaDock and using information derived from mutagenesis experiments, using the fact that residues Asp108, Phe121, Pro122 and Asp123 in Nef are critical for binding to hTE. Several charged and polar groups of Nef provide a negatively charged regions for the binding, i.e. Asp108, Asp111 and Asp123. Those aminoacids, experimentally proven to be located in the binding surface, play an essential role in the interaction, should be taken at the qualitative level due to uncertainties of the docking model. On the other hand, the residues found for the human thioesterase, are complementary to the ones in NEF. As expected, hydrophobic interactions may also play a role for the adduct stabilization, as can be observed from the electrostatic potential (Figures 11, 12).

Interestingly, in the uncomplexed hTE the two hydrophobic residues (Pro320 and Lys322) are in contact with the solvent. Instead, the three hydrophobic residues of Nef (Phe121, Pro122) are also important for the dimerization of the viral factor. In the case of uncomplexed Nef these residues form part of an open surface.

Our findings may help explain the experimentally low affinity of the *E. coli* isoenzyme for the viral factor. Indeed, in the isoenzyme from E. coli two negatively charged residues (an Asp and a Glu residues) replace two aromatic residues fundamental for the binding to Nef (Pro320 and Lys322) of hTE. These residues are expected to have low affinity for the hydrophobic pocket of Nef (which is constituted by Phe121, and Pro122) and may produce unstabilization as they will interact with Nef's negatively charged residues (Asp108, Asp123, Asp111).

Finally, our calculations suggest that the ‘active site loops’ (residues 454 to 485) of both subunit A and B are very mobile, as evident from an essential modes analysis on the multi ns timescale (Fig.15). Nef binding causes a significant change in the dynamics: in the complex, indeed, only ‘active site’ loop of subunit B is mobile whereas that of subunit A is relatively rigid. The motion of this region of the protein may be mechanically hindered by the presence of hydrophobic interactions between three residues on the ‘active site’ loop and other three belonging to Nef. At the speculative level, we suggest that Nef binding affect enzymatic activity, consistently with some experimental evidence. Also in this case, experimental data and activity calculations are needed to establish this proposal.

More experimental data as well as calculations of affinity and/or binding free energy are required to firmly establish these issues. In particular, we will propose to our experimental collaborators the mutations listed in Fig.25. In which the physico-chemical properties of the aminoacids putatively present on the interaction surface are modified to inhibit the hTE-Nef interaction.



Theoretical Data	
R (Sub.A)	**hTE Mutant (Sub.A)**
ASP434(ASP143)	D434A(D143A)
LYS322(LYS31)	K322S(K31S)
PRO320(PRO29)	P320A(P29A)
LYS361(LYS70)	K361S(K70S)
LYS581(LYS90)	K581S(K90S)
ARG356(ARG65)	R356S(R65S)

Figure 25. *Detail of the proposed Mutations*

Although the models generated in this thesis can be compared to low resolution crystal structures, the use of Homology modeling techniques and state-of-art bioinformatic tools makes room to the possibility of building and analysing thousands of complexes and models, thus allowing the identification, on the human thioesterase 8, of residues critical for the interaction with Nef.

The combined computational/experimental approaches allowed us to design a few new experiments aimed at a clear characterization of the residues involved, not only in Nef binding, but also in the enzymatic mechanism.

Further advancement in experimental structural biology , along with algorithms for free energy calculations, multiscale modeling, and protein-protein docking make us confident that the challenge of characterizing how the virus interacts with the different human targets can be undertaken in short time and that these approaches may provide a great improvement to our understanding of cell and molecular biology events upon virus infection.

Acknowledgements

I would like to thank Dr. Alejandro Giorgetti for all the time he spent with patience, for all the information he gathered, and the help he gave me to fix my thesis problems from the beginning to the end.
I would also like to thank Prof. Andrea Sbarbati for all the support he gave me for the last three years.

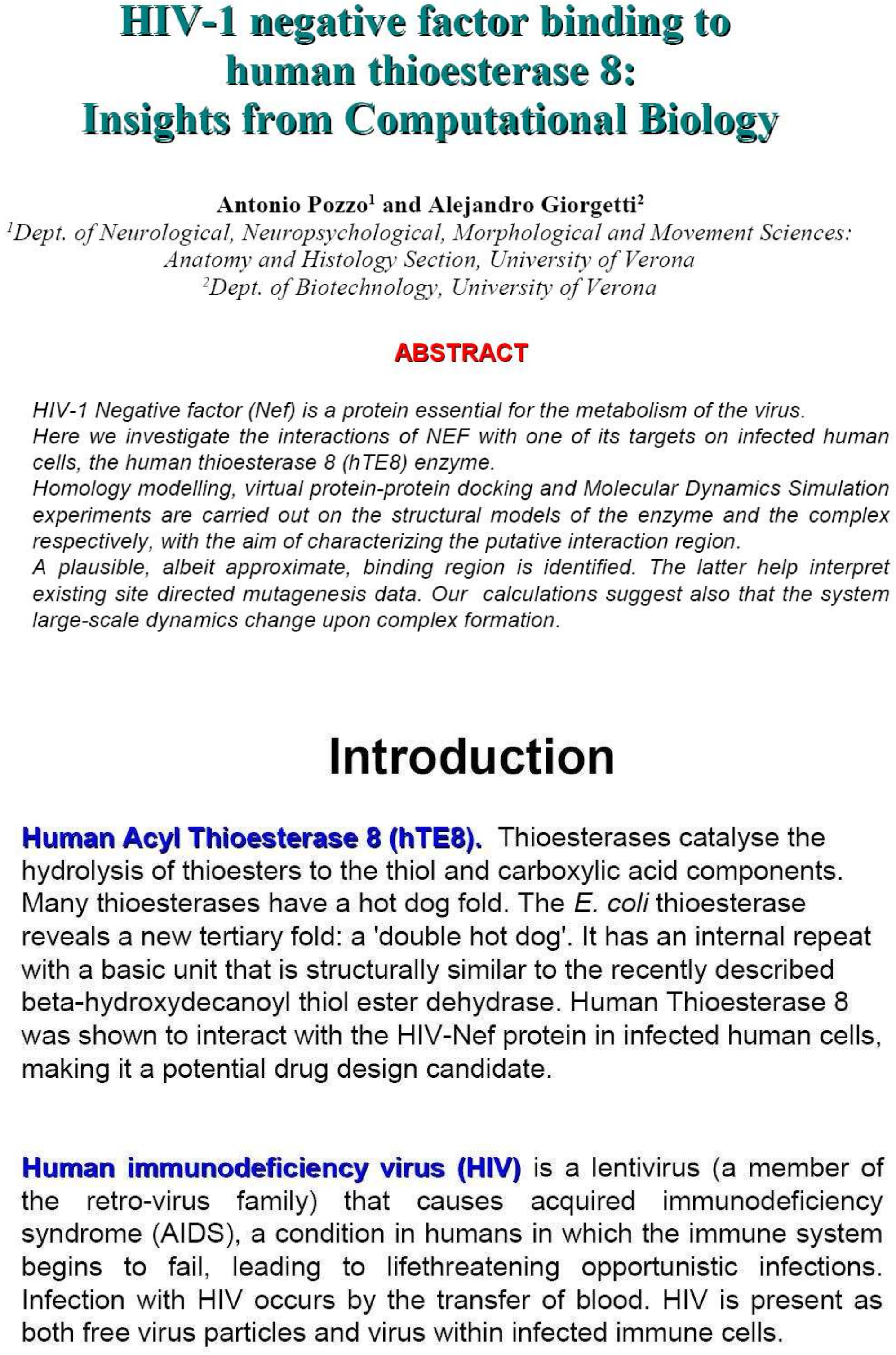
Reference list

1. Collette, Y. et al (1996) J Biol Chem 271, 6333-6341
2. Greenberg, M. E. et al (1998) EMBO J 17, 2777-2789
3. Barber, S. A. et al (1998) Virology 251, 165-175
4. Nunn, M. F. et al (1996) J Virol 70, 6157-6161
5. Sawai, E. T. et al (1994) Proc Natl Acad Sci U S A 91, 1539-1543
6. Harris, M. (1999) Curr Biol 9, R459-R461
7. Rossi, F. et al (1996) Virology 217, 397-403
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-10
9. Arold, S. et al (1997) Structure 5, 1361-1372
10. Watanabe, H. et al (1997) Biochem Biophys Res Commun 238,234-239
11. Liu, L. X. et al (2000) J Virol 74, 5310-5319
12. Cohen, G. B. et al (2000) J Biol Chem 275, 23097-23105
13. Johannes Soding et al (2005). Bioinformatics 21 : 951-960
14. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. John Wiley & Sons, et al (2006) Inc., Supplement 15, 5.6.1-5.6.30.
15. L. Martinez, R. Andreani, J. M. Martinez et al (2007) BMC Bioinformatics, 8:306.
16. Gray, J.J., Moughan S.E., Wang C., Schueler-Furman O., Kuhlman B., Rohl C.A., Baker D. et al (2003) J. Mol. Biol. 331(1), 281-299.
17. http://www.python.org/doc/faq/general/#what-is-python. Et al (2009) Python Software Foundation.
18. William Humphrey, Andrew Dalke, and Klaus Schulten. Et al (1996) Journal of Molecular Graphics, 14:33-38.
19. Wu CH, Apweiler R, Bairoch A, Natale, DA, Barker WC, et al. (2006) Nucleic Acids Res 34: D187–D191.
20. Ropelewski AJ, Nicholas HB Jr, Deerfield DW et al (2004) Bioinformatics. Chapter 3, Unit3
21. Pei J, Grishin NV et al (2007) Bioinformatics 23: 802–808.
22. Soding J et al (2005) Bioinformatics 21: 951–960.
23. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T et al (2007) Suppl 868–82.
24. Worth CL, Kleinau G, Krause G et al (2009) PLoS One 4: e7011.
25. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2006) Bioinformatics. Chapter 5, Unit.
26. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science:223-230
27. Bruccoleri RE, Karplus M (1990) Conformational sampling using high-temperature molecular dynamics. Biopolymers 29:1847-62

28. Chinea G, Padron G, Hooft R, Sander C, Vriend G (1995) The use of position-specific rotamers in model building by homology. Proteins 23:415-21
29. Chothia C, Lesk A (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823-826
30. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892-3
31. Fariselli P, Casadio R (2001) Prediction of disulfide connectivity in proteins. Bioinformatics 17:957-64
32. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Science 9:1753-73
33. Godzik A, Skolnick J (1992) Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proceedings of the National Academy of Sciences of the United States of America 89:12098-102
34. Holm L, Sander C (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins 14:213-23
35. Jacoboni I, Martelli PL, Fariselli P, Compiani M, Casadio R (2000) Predictions of protein segments with the same amino acid sequence and different secondary structure: a benchmark for predictive methods. Proteins 41:535-544
36. Keller D, Shibata M, Marcus E, Ornstein R, Rein R (1995) Finding the global minimum: a fuzzy end elimination implementation. Protein Eng 8:893-904
37. Kopp J, Schwede T (2006) The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res 34:D315-8
38. Levinthal C (1968) Are there pathways for protein folding? Journal de Chimie Physique et de Physico-Chimie Biologique 65:44-45
39. Martelli PL, Fariselli P, Casadio R (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. Proteomics4:1665-71
40. Martelli PL, Fariselli P, Malaguti L, Casadio R (2002) Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng 15:951-3
41. Moult J (1996) The current state of the art in protein structure prediction. Current Opinion in Biotechnology 7:422-427
42. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815
43. Tramontano A (1995) The architecture of loops in proteins. In: Villar HO (ed) Advances in Computational Biology. JAI Press, Greenwich, p 239-259
44. Tramontano A, Chothia C, Lesk AM (1989) Structural determinants of the conformations of medium-sized loops in proteins. Proteins 6:382-94
45. Ward JJ, L.J. M, Buxton BF, Jones DT (2003) Secondary structure prediction with support vector machines. Bioinformatics 19:1650-1655

Appendix A

Poster accepted for oral presentation to BIOTECHNO 2011-The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies-May 22-27, 2011 - Venice, Italy



HIV-1 negative factor binding to human thioesterase 8: Insights from Computational Biology

Antonio Pozzo[1] and Alejandro Giorgetti[2]
[1]*Dept. of Neurological, Neuropsychological, Morphological and Movement Sciences: Anatomy and Histology Section, University of Verona*
[2]*Dept. of Biotechnology, University of Verona*

ABSTRACT

HIV-1 Negative factor (Nef) is a protein essential for the metabolism of the virus. Here we investigate the interactions of NEF with one of its targets on infected human cells, the human thioesterase 8 (hTE8) enzyme. Homology modelling, virtual protein-protein docking and Molecular Dynamics Simulation experiments are carried out on the structural models of the enzyme and the complex respectively, with the aim of characterizing the putative interaction region. A plausible, albeit approximate, binding region is identified. The latter help interpret existing site directed mutagenesis data. Our calculations suggest also that the system large-scale dynamics change upon complex formation.

Introduction

Human Acyl Thioesterase 8 (hTE8). Thioesterases catalyse the hydrolysis of thioesters to the thiol and carboxylic acid components. Many thioesterases have a hot dog fold. The *E. coli* thioesterase reveals a new tertiary fold: a 'double hot dog'. It has an internal repeat with a basic unit that is structurally similar to the recently described beta-hydroxydecanoyl thiol ester dehydrase. Human Thioesterase 8 was shown to interact with the HIV-Nef protein in infected human cells, making it a potential drug design candidate.

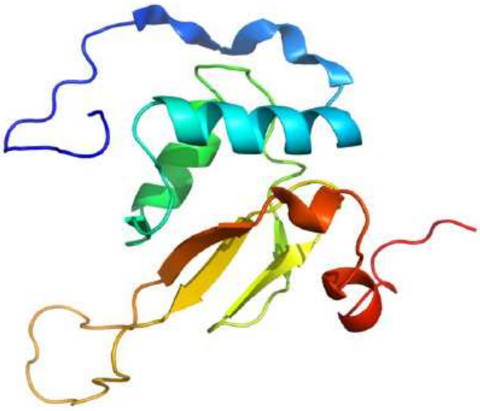
Human immunodeficiency virus (HIV) is a lentivirus (a member of the retro-virus family) that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to lifethreatening opportunistic infections. Infection with HIV occurs by the transfer of blood. HIV is present as both free virus particles and virus within infected immune cells.

Negative Factor (Nef)

The Negative Factor (**Nef**) protein from Human Immunodeficieny Virus type 1 (HIV-1) is a 27-kDa-myristoylated protein required to produce a high viral load of the virus.

Nef also advances the endocyotsis and degradation of cell surface proteins, including CD4, hTE8 and MHC proteins (CD4 is an integral membrane protein that functions in T-cell activation, and is the receptor for the HIV virus).

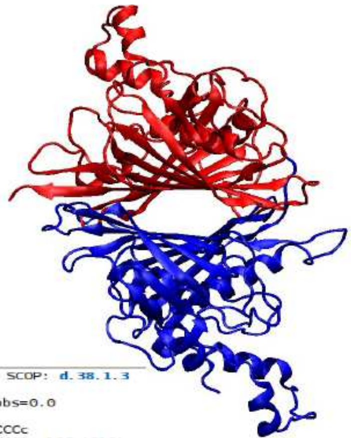
This action impairs cytoxic T cell function, thereby helping the virus to evade the host immune responsea

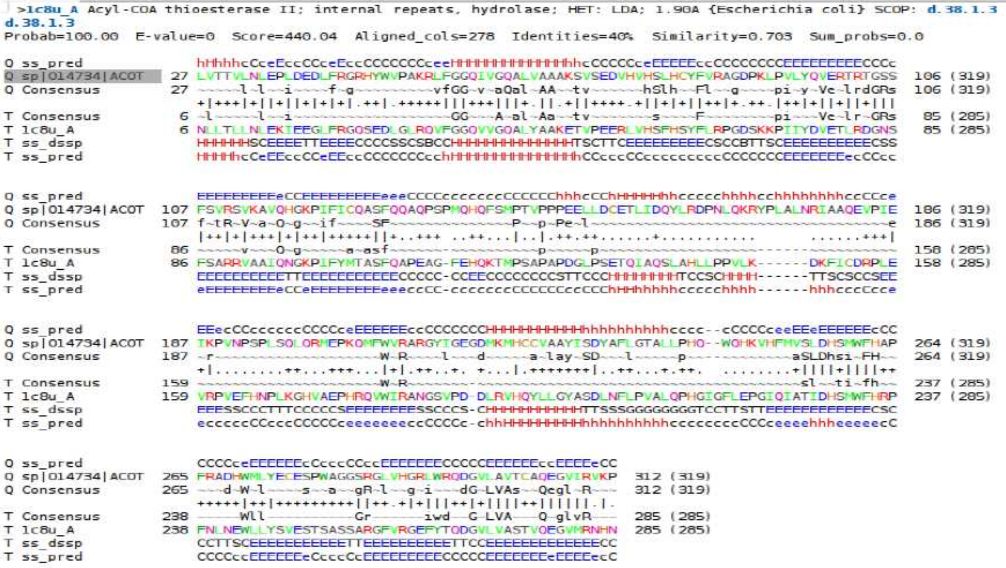


Recently, **point directed mutagenesis** experiments were carried out on **Nef** with the aim of identifying the surface of interaction between the latter an one of its targets in human cells: the **Acyl Thioesterase 8**. In the study, five residues that play a crucial role for the binding to hTE8 were found, i.e. **Asp108,Leu112, Phe121, Pro122 and Asp123.**

hTE8: Comparative Modeling

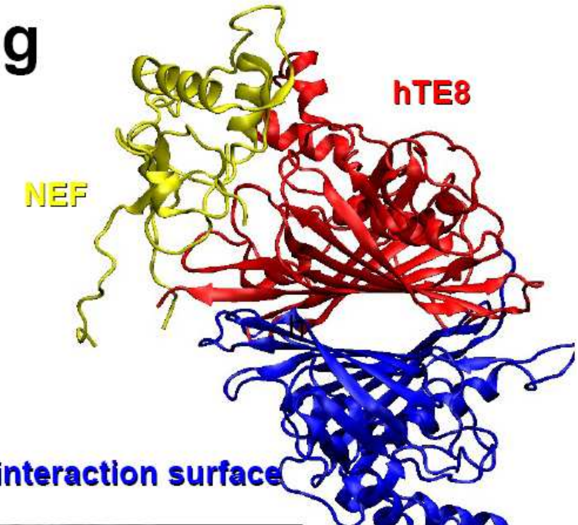
HHpred software was used for the template search and alignment. **Modeller9v4** was used for model construction. 100 models of the functional dimer were built and ranked using Modeller objective function and stereochemical analysis





```
>1c8u_A Acyl-COA thioesterase II; internal repeats, hydrolase; HET: LDA; 1.90A {Escherichia coli} SCOP: d.38.1.3
d.38.1.3
Probab=100.00  E-value=0  Score=440.04  Aligned_cols=278  Identities=40%  Similarity=0.703  Sum_probs=0.0

Q ss_pred              hHhhhcCceEccCCceEccCCCCCCCceeHHHHHHHHHHHHHhhcCCCCCceEEEEEccCCCCCCCCEEEEEEEEEECCCc
Q sp|014734|ACOT   27 LVTTVLNLEPLDEDLFRGRHYWVPAKRLFGGQIVGQALVAAAKSVSEDVHVHSLHCYFVRAGDPKLPVLYQVERTRTGSS  106 (319)
Q Consensus       27 ~~~~~l~l~~i~~~~~f~g~~~~~~~~~~vfGG~v~aQal~AA~~tv~~~~~~~hSlh~~Fl~~g~~~~pi~y~Ve~lrdGRs  106 (319)
                       +|+++|+||+||+|+|+|.++|.+++++|||+++|||+.||.+||++++.+||+|+||++|+.++.|++|+||+||+|||
T Consensus        6 ~l~~~~~~l~~i~~~~~~~~~~~~~~~~~~~GG~~~A~al~Aa~~tv~~~~~~~~s~~~~F~~~~~~~~~pi~~~Ve~lr~GRs   85 (285)
T 1c8u_A          6 NLLTLLNLEKIEEGLFRGQSEDLGLRQVFGGQVVGQALYAAKETVPEERLVHSFHSYFLRPGDSKKPITYDVETLRDGNS   85 (285)
T ss_dssp            HHHHHHSCEEEETTEEEECCCCSSCSBCCHHHHHHHHHHHHHHHTSCTTCEEEEEEEEEECSCCBTTSCEEEEEEEEEEECSS
T ss_pred              HHHHhcCeEEccCCeEEccCCCCCCCcchHHHHHHHHHHHHHHHHhCCcccCCcccccccccCCCCCCCEEEEEEEecCCcc

Q ss_pred              EEEEEEEEEEeCCEEEEEEEEEEeeeCCCCcccccccCCCCCChhhcCChHHHHHhhcccccchhhhcchhhhhhhhccCCce
Q sp|014734|ACOT  107 FSVRSVKAVQHGKPIFICQASFQQAQPSPMQHQFSMPTVPPPEELLDCETLIDQYLRDPNLQKRYPLALNRIAAQEVPIE  186 (319)
Q Consensus      107 f~tR~V~a~Q~g~~if~~~~SF~~~~~~~~~~~~~~~~~P~~p~Pe~l~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~e  186 (319)
                       |++|+|+++|+|+|++|+++++||+..+++ ..++...|..|.++.++......+..........      ......++|
T Consensus       86 ~~~~~v~~~~Q~g~~~~~~~a~asf~~~~~~~~~~~~~~~~~p~~~~~p~~~~~~~~~~~~~~~~~~~~~------~~~~~~~~~  158 (285)
T 1c8u_A         86 FSARRVAAIQNGKPIFYMTASFQAPEAG-FEHQKTMPSAPAPDGLPSETQIAQSLAHLLPPVLK------DKFICDRPLE  158 (285)
T ss_dssp            EEEEEEEEEEETTEEEEEEEEEEEEECCCCC-CCEECCCCCCCCSTTCCCHHHHHHHHTCCSCHHHH------TTSCSCCSEE
T ss_pred              eEEEEEEEEEeCCeEEEEEEEEEeeeccCC-cccccccCCCCCCcccCCChHHhhhhcccchhhh------hhhcccCcce

Q ss_pred              EEecCCcccccccCCCCceEEEEEEEccCCCCCCCHHHHHHHHHHHHhhhhhhhhhhcccc--cCCCCCceeEEeEEEEEEEcCC
Q sp|014734|ACOT  187 IKPVNPSPLSQLQRMEPKQMFWVRARGYIGEGDMKMHCCVAAYISDYAFLGTALLPHQ--WQHKVHFMVSLDHSMWFHAP  264 (319)
Q Consensus      187 ~r~~~~~~~~~~~~~~~~~~~~~W~R~~~~~l~~~~d~~~~~~a~lay~SD~~~l~~~~~~~p~~~~~~~~~~~~~aSLDhsi~FH~~  264 (319)
                       +|........++...+++...|+|.++..+. +...|.+++++++|..++...+.++.  ........+||||+||||++
T Consensus      159 ~~~~~~~~~~~~~~~~~~~~~~~~~~W~R~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~sl~~ti~fh~~  237 (285)
T 1c8u_A        159 VRPVEFHNPLKGHVAEPHRQVWIRANGSVPD-DLRVHQYLLGYASDLNFLPVALQPHGIGFLEPGIQIATIDHSMWFHRP  237 (285)
T ss_dssp            EEESSCCCTTTCCCCCSEEEEEEEEESSCCCS-CHHHHHHHHHHHHTTSSSGGGGGGGTCCTTSTTEEEEEEEEEEEEEECSC
T ss_pred              eccccCCccccCCCCCceeeeeeeeccCCCCc-chhHHHHHHHHHHhhhhhhhhhcccccccCCCceeeehhheeeeecC

Q ss_pred              CCCCceEEEEEEcCcccCCccEEEEEEEECCCCCEEEEEEccEEEEeCC
Q sp|014734|ACOT  265 FRADHWMLYECESPWAGGSRGLVHGRLWRQDGVLAVTCAQEGVIRVKP  312 (319)
Q Consensus      265 ~~~d~W~l~~~~~s~~a~~gR~l~~g~i~~~~dG~LVAs~~Qegl~R~~~  312 (319)
                       +++++|++|+++++++++++||++.+|+|||++|+||||++|||||.|.
T Consensus      238 ~~~~~~Wll~~~~~~~~~~~Gr~~~~~~iwd~~G~LVA~~~Q~glvR~~  285 (285)
T 1c8u_A        238 FNLNEWLLYSVESTSASSARGFVRGEFYTQDGVLVASTVQEGVMRNHN  285 (285)
T ss_dssp            CCTTSCEEEEEEEEEEEEETTEEEEEEEEEEETTCCEEEEEEEEEEEEEEEEECC
T ss_pred              CCCCccEEEEEEeCcccCcEEEEEEEEECCCCCEEEEEEEeEEEEecC
```

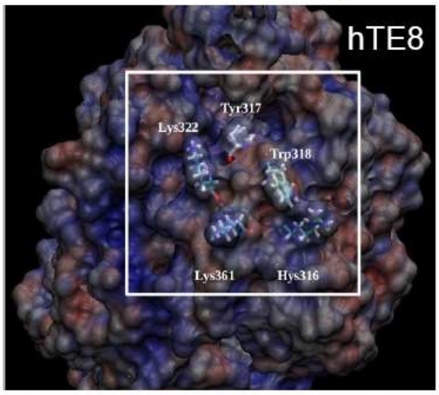
hTE8-Nef docking

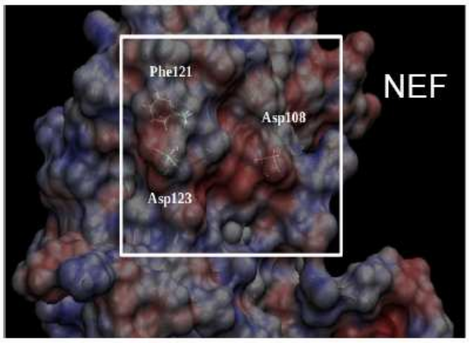


Software: RosettaDock

Protocol: **a)** Nef position randomized; **b)** distance constraints between Nef residues: **D108, F121, P122 and D123** and the **hTE8 surface**; **c)** 3000 decoys produced and hierarchically clustered with a cut-off RMSD value of 2.5 Å.

Electrostatic complementarity of the interaction surface



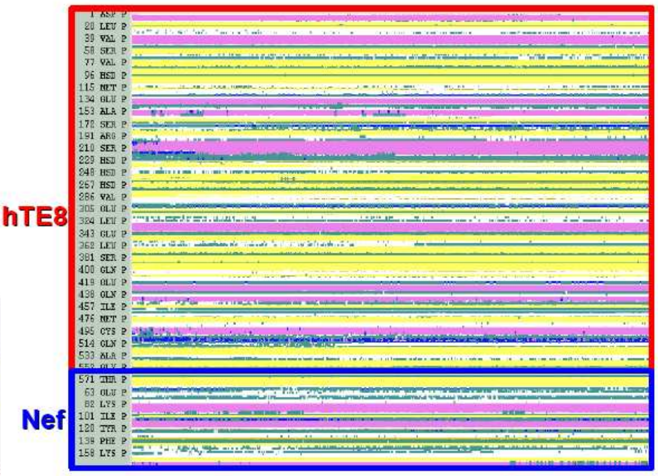


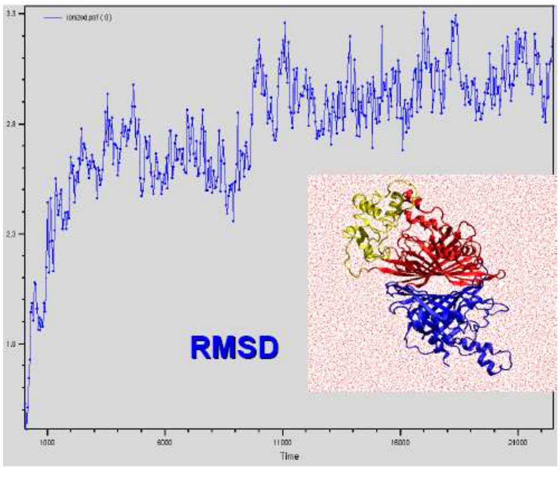
MD simulation

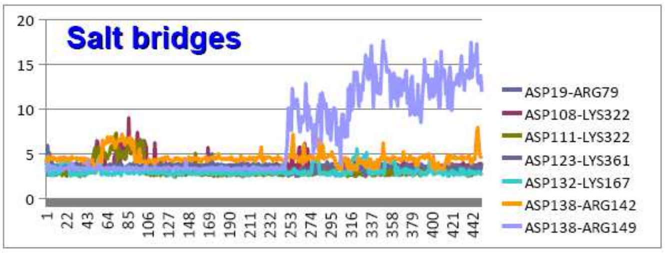
Program: NAMD (CHARMM22 and TIP3P force fields)

Protocol: **a)** 9200 water molecules for solvation and PBC; **b)** Time step: 2 fs; **c)** PME for electrostatic interactions; **d)** Constant temperature (300 K) and pressure (1 atm);**e)** 2000 steps of system minimization: using conjugate gradients; **f)** 22.5 ns molecular dynamics simulation.

Secondary Structure Conservation







Discussion

We have investigated the interactions between the HIV-1 Nef and its cellular partner hTE8: we modeled the structure of Human thioesterase 8 based on the isoenzyme from *E.coli.*

An approximate model of the Nef/hTEII complex was built using protein-protein docking: the docking was guided by using information derived from mutagenesis experiments.

The interacting surface is characterized by charge complementarity

The interactions are stable during the entire MD simulation: hydrophobic interactions may also play a role for the adduct stabilization, as can be observed from the electrostatic potential

Our findings can help explain the experimentally low affinity of the E. coli isoenzyme for the viral factor. Indeed, in the isoenzyme from E. coli two negatively charged residues replace two aromatic residues found in the binding region of hTE8.

More experimental data as well as calculations of affinity and/or binding free energy are required to firmly establish these issues.

Appendix B

Introduction to Biochemistry. In the following pages description of the aminoacids and their putative interactions are described (http://en.wikipedia.org/wiki/Amino_acid)

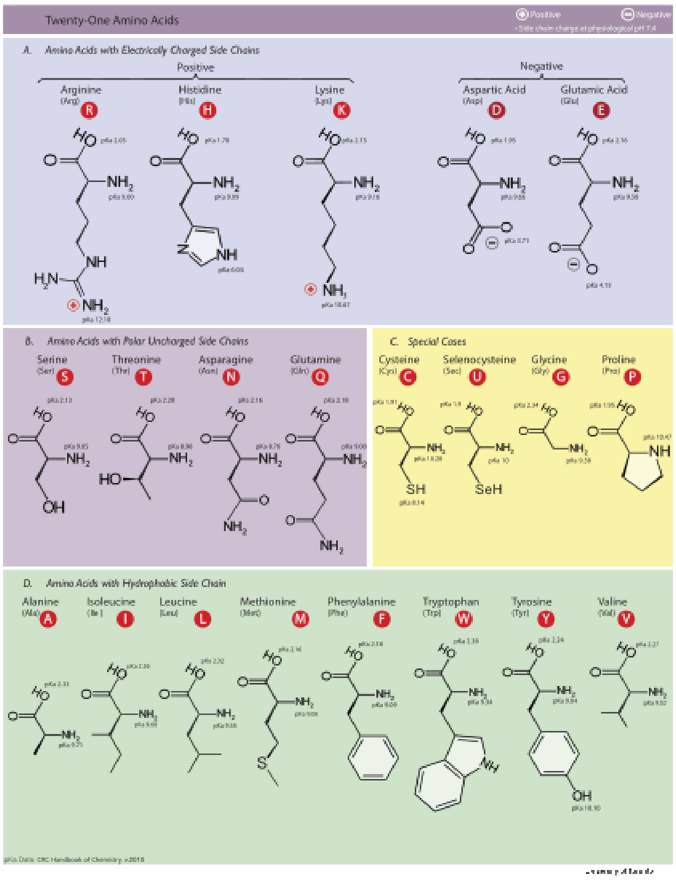


Figure B1. *The 21 amino acids found in eukaryotes, grouped according to their side-chains' pKas and charge at physiological pH 7.*

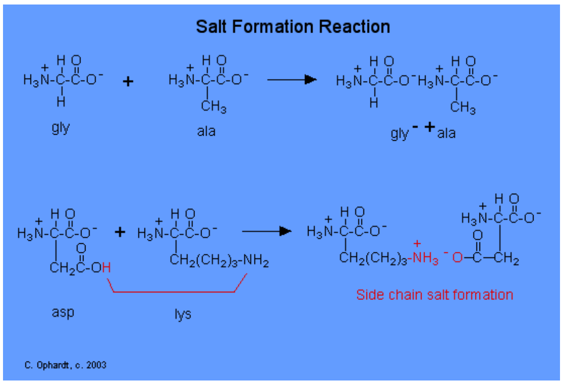


Figure B2. *Non-covalent bond*

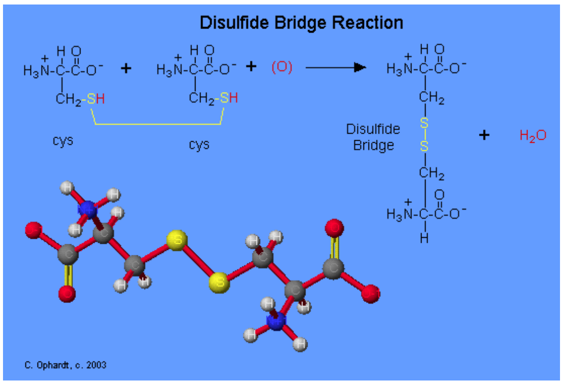


Figure B3. *Covalent bond*

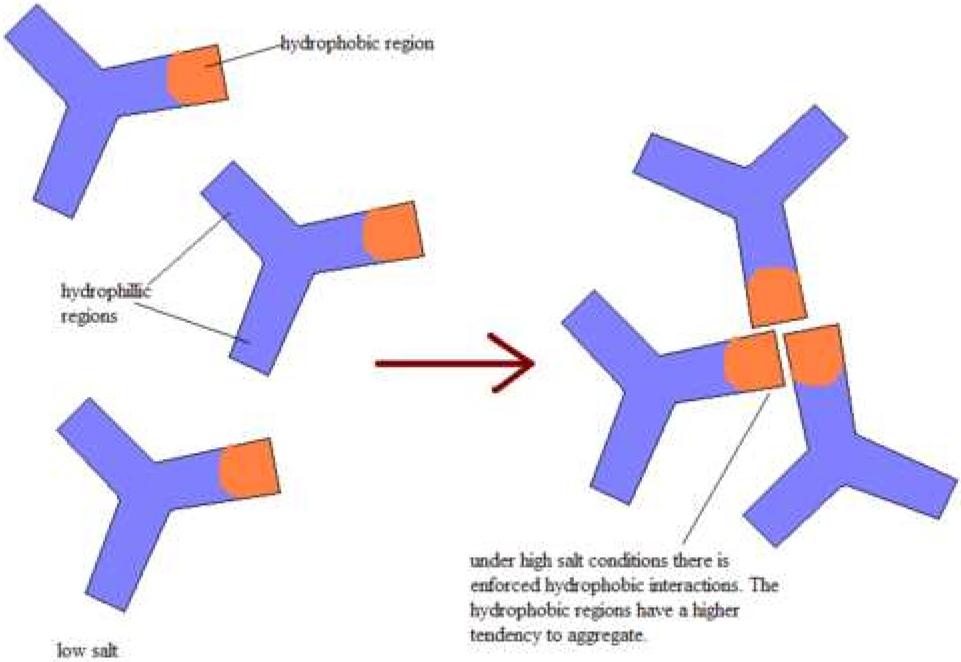

hydrophobic region
hydrophillic regions
low salt
under high salt conditions there is enforced hydrophobic interactions. The hydrophobic regions have a higher tendency to aggregate.

Figure B4. *Hydrophobic interactions: the tendency of hydrocarbons (or of lipophilic hydrocarbon-like groups in solutes) to form intermolecular aggregates in an aqueous medium, and analogous intramolecular interactions*

Hydrophobic effect

The hydrophobic effect represents the tendency of water to exclude non-polar molecules. The effect originates from the disruption of highly dynamic hydrogen bonds between molecules of liquid water by the nonpolar solute. Polar chemical groups, such as OH group in methanol do not cause the hydrophobic effect. However, a pure hydrocarbon molecule, for example hexane, is incapable of forming hydrogen bonds with water. Introduction of hexane into water causes disruption of the hydrogen bonding network between water molecules. The hydrogen bonds are partially reconstructed by building a water "cage" around the hexane molecule, similar to that in clathrate hydrates formed at the lower temperatures. The water molecules that form the "cage" (or solvation shell) have substantially restricted mobilities. This leads to significant losses in translational and rotational entropy of water molecules and makes the process unfavorable in terms of free energy of the system.

The hydrophobic effect can be quantified by measuring the partition coefficients of non-polar molecules between water and non-polar solvents. The partition coefficients can be transformed to free energy of transfer which includes enthalpic and entropic components, ΔG = ΔH - TΔS. These components are experimentally determined by calorimetry. The hydrophobic effect was found to be entropy-driven at room temperature because of the reduced mobility of water molecules in solvation shell of the non-polar solute. However, the enthalpic component of transfer energy was found to be favorable, meaning strengthening of water-water hydrogen bonds in the solvation shell, apparently due to the reduced mobility of water molecules . At the higher temperature, when water molecules became more mobile, this energy gain decreases, but so does the entropic component. As a result of such entropy-enthalpy compensation, the hydrophobic effect (as measured by the free energy of transfer) is only weakly temperature-dependent and became smaller at the lower temperature, which leads to "cold denaturation" of proteins.

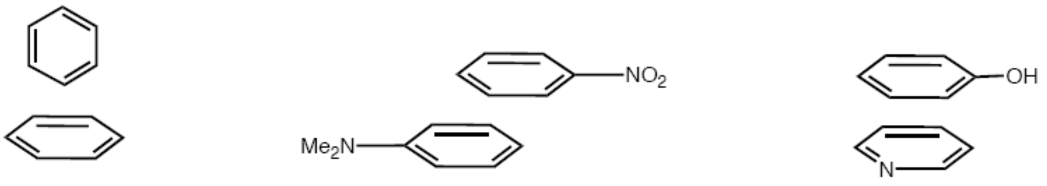


Figure B5. *π-π Interactions: this class of interaction involves direct attraction between arene rings. This was long considered to be a charge transfer phenomenon, but this was later disproved*

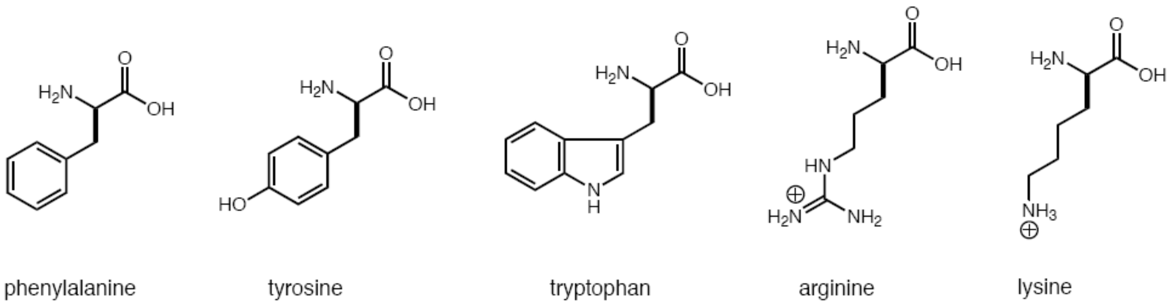


Figure B6. *π-Cation Interactions: Survey of protein database shows that π-Cation stabilization is a major facet of protein structure and enzyme catalysis*