Università degli Studi di Bologna

Dottorato di Ricerca in Scienze Chimiche (XIX CICLO)

CHIM/02

THEORETICAL STUDY OF BIOLOGICAL SYSTEMS

DISSERTAZIONE FINALE

Presentata da:

Bruno Trebbi

Relatore:
Prof. Francesco Zerbetto

Coordinatore di Dottorato:
Prof. Vincenzo Balzani

Bologna 2007

RINGRAZIAMENTI

Ringrazio il professor Francesco Zerbetto (che mi segue fin dalla tesi di laurea) per il suo aiuto, la sua competenza, e per la pazienza che ha sempre mostrato nei miei confronti nei momenti difficili. A quanto pare dovrà rassegnarsi a vedermi in giro ancora per un po'...

Ringrazio il gruppo di ricerca tutto, presente passato e futuro. Quelli che sono qui, e quelli che ora sono sparsi in giro per l'Europa. Grazie per il prezioso aiuto, e spero che ognuno di loro trovi la sua strada.

Ringrazio la mia famiglia, che da sempre mi è vicina e mi aiuta in ogni modo possibile ed immaginabile. Questo dottorato è anche loro.

Ringrazio Michela, per avermi fatto scoprire di nuovo, in un momento davvero buio, che l'amore vero esiste, ed è una cosa meravigliosa.

Ringrazio i miei amici, che da sempre ci sono quando ho bisogno. I veri amici sanno condividere con te i momenti belli, e sanno aiutarti nei momenti di difficoltà.

Ringrazio lo sport, ottima valvola di sfogo contro lo stress della vita quotidiana.
E' bello, quasi a 30 anni, avere ancora la capacità e la voglia di migliorarsi e stupirsi.

Ringrazio Dio, perchè Lui c'è sempre, e sempre ti aiuta, anche e soprattutto quando non Lo ascolti e vai dritto per la tua strada. Grazie a Lui tutti abbiamo una speranza. Persino io.

Bruno

INDEX

PREFACE

The subject of this Ph.D. research thesis is the theoretical and computational study of biological system.
Computational science now is very advanced, and it allows us to perform very sophisticated simulations.
In this work I have applied molecular mechanics models to proteins, investigating about different properties. I report here the more significant results: intra-residue energy distribution of proteins, aromatic stabilization, and configurational temperature.

This thesis is organized into independent chapters. chapter 1 is about theoretical and computational background, while chapter 2 treats about intra-residue energy distribution of proteins. In chapter 3 aromatic stabilization topic is discussed, and in chapter 4 we speak about configurational temperature.
Finally, in chapter 5 there are conclusions and general remarks.

I want to thank Professor Francesco Zerbetto for his support, useful discussion and financial support. I'm also very grateful to all people in lab.

CHAPTER 1
THEORETICAL AND COMPUTATIONAL BACKGROUND

1.1 COMPUTATIONAL CHEMISTRY

The term *theoretical chemistry* may be defined as a mathematical description of chemistry, whereas *computational chemistry* is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Note that the words *exact* and *perfect* do not appear here, as very few aspects of chemistry can be computed exactly. Almost every aspect of chemistry, however, can be described in a qualitative or approximate quantitative computational scheme.

Molecules consist of nuclei and electrons, so the methods of quantum mechanics apply. Computational chemists often attempt to solve the non-relativistic Schrödinger equation, with relativistic corrections added, although some progress has been made in solving the fully relativistic Schrödinger equation. It is, in principle, possible to solve the Schrödinger equation, in either its time-dependent form or time-independent form as appropriate for the problem in hand, but this in practice is not possible except for very small systems. Therefore, a great number of approximate methods strive to achieve the best trade-off between accuracy and computational cost. Present computational chemistry can routinely and very accurately calculate the properties of molecules that contain no more than 10-40 electrons. The treatment of larger molecules that contain a few dozen electrons is computationally tractable by approximate methods such as density functional theory (DFT). There is some dispute within the field whether the latter methods are sufficient to describe complex chemical reactions, such as those in biochemistry. Large molecules can be studied by semi-empirical approximate methods. Even larger molecules are treated with classical mechanics in methods called molecular mechanics.

In theoretical chemistry, chemists, physicists and mathematicians develop algorithms and computer programs to predict atomic and molecular properties and reaction paths for chemical reactions. Computational chemists, in contrast, may simply apply existing computer programs and methodologies to specific chemical questions. There are two different aspects to computational chemistry:

- Computational studies can be carried out in order to find a starting point for a laboratory synthesis, or to assist in understanding experimental data, such as the position and source of spectroscopic peaks.
- Computational studies can be used to predict the possibility of so far entirely unknown molecules or to explore reaction mechanisms that are not readily studied by experimental means.

Thus computational chemistry can assist the experimental chemist or it can challenge the experimental chemist to find entirely new chemical objects.

Several major areas may be distinguished within computational chemistry:

- The prediction of the molecular structure of molecules by the use of the simulation of forces to find stationary points on the energy hypersurface as the position of the nuclei is varied.
- Storing and searching for data on chemical entities.
- Identifying correlations between chemical structures and properties.
- Computational approaches to help in the efficient synthesis of compounds.
- Computational approaches to design molecules that interact in specific ways with other molecules (e.g. drug design).

1.2 MOLECULAR STRUCTURE

A given molecular formula can represent a number of molecular isomers. Each isomer is a local minimum on the energy surface (called the potential energy surface) created from the total energy (electronic energy plus repulsion energy between the nuclei) as a function of the coordinates of all the nuclei. A stationary point is a geometry such that the derivative of the energy with respect to all displacements of the nuclei is zero. A local (energy) minimum is a stationary point where all such displacements lead to an increase in energy. The local minimum that is lowest is called the global minimum and corresponds to the most stable isomer. If there is one particular coordinate change that leads to a decrease in the total energy in both directions, the stationary point is a transition structure and the coordinate is the reaction coordinate. This process of determining stationary points is called geometry optimisation.

The determination of molecular structure by geometry optimisation became routine only when efficient methods for calculating the first derivatives of the energy with respect to all atomic

coordinates became available. Evaluation of the related second derivatives allows the prediction of vibrational frequencies if harmonic motion is assumed. In some ways more importantly it allows the characterisation of stationary points. The frequencies are related to the eigenvalues of the matrix of second derivatives (the Hessian matrix). If the eigenvalues are all positive, then the frequencies are all real and the stationary point is a local minimum. If one eigenvalue is negative (an imaginary frequency), the stationary point is a transition structure. If more than one eigenvalue is negative the stationary point is a more complex one, and usually of little interest. When found, it is necessary to move the search away from it, if we are looking for local minima and transition structures.

The total energy is determined by approximate solutions of the time-dependent Schrödinger equation, usually with no relativistic terms included, and making use of the Born-Oppenheimer approximation which, based on the much higher velocity of the electrons in comparison with the nuclei, allows the separation of electronic and nuclear motions, and simplifies the Schrödinger equation. This leads to evaluating the total energy as a sum of the electronic energy at fixed nuclei positions plus the repulsion energy of the nuclei. A notable exception are certain approaches called direct quantum chemistry, which treat electrons and nuclei on a common footing. Density functional methods and semi-empirical methods are variants on the major theme. For very large systems the total energy is determined using molecular mechanics. The ways of determing the total energy to predict molecular structures are:

- *Ab initio* methods
- Density Functional theory
- Semi-empirical and Empirical methods
- Molecular Mechanics

In the next chapter there will be a brief introduction about Molecular Mechanics methods, the ones most used in this work

1.3 MOLECULAR MECHANICS AND FORCE FIELDS

The microscopic state of a molecular system can be described by defining the position (q_i) and momentum (p_i) of each particle of the system at every time.

Considering the Born-Oppenheimer approximation, the Hamiltonian of a system can be expressed as a function of the nuclear variables, the rapid motion of the electrons having been averaged out. This classical approach requires the use of Force Field (from now on, *FF*) methods, known as

Molecular Mechanics (*MM*), which consider the total potential energy of a chemical structure as a function of the only nuclear atomic positions. Making the additional approximation that a classical description is adequate, we may write the Hamiltonian H of a system containing N particles as a sum of kinetic and potential energy:

$$H(q^N, p^N) = K(p^N) + V(q^N). \tag{1.1}$$

Usually the kinetic energy K takes the form

$$K = \sum_{i=1}^{N} \sum_{\alpha} p_{i\alpha}^2 / 2m_i \tag{1.2}$$

where m_i is the molecular mass and the index α runs over the different *(x,y,z)* components of the momentum of the molecule *i.*

The potential energy V may be divided into terms depending on the coordinates of individual atoms for the given conformation, such as the stretching of bonds, the opening and closing of angles, the rotation about single bonds and the long range interactions. It can be expressed as follows:

$$V(q^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\omega - \gamma)] +$$

$$+ \sum_{i=1}^{N} \sum_{j>i}^{N} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{Q_i Q_j}{4\pi\varepsilon_0 r_{ij}} \right\} \tag{1.3}$$

Equation 1.3 represents the simplest MM Force Field.

As it is shown in Figure 1.1, the mechanical molecular model considers atoms as spheres and bonds as springs.

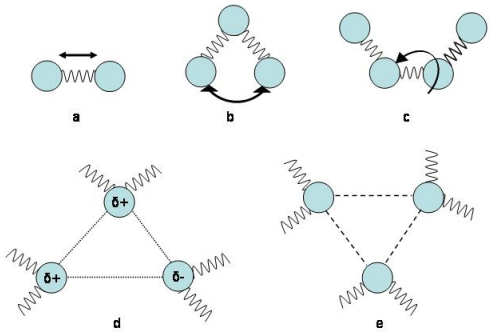


Figure 1.1 *Schematic of a molecular force field: the mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch (a), bend (b), and twist (c). Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii (d,e).*

The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist. In fact, the first term of the potential energy function in Equation 1.3 is similar to the Hooke's law for a spring deformation. It represents the bond stretching and describes the interaction between pairs of bonded atoms by a harmonic potential, increasing in energy as the bond length l_i deviates from its reference value $l_{i,0}$. The second term is the angle of bending θ_i of the molecule, again modelled using a harmonic potential. In both terms, k_i represents the force's constant. The third term is a torsional potential that shows how the energy changes as a bond rotates: the V_n parameter controls the amplitude of the curve, the n parameter controls its periodicity and reflects the type symmetry in the dihedral angle, and γ shifts the entire curve along the rotation angle axis ω.

Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii. Therefore the fourth contribution is the non-bonded term, calculated between all pairs of atoms belonging to different molecules or to the same molecule but separated by at least three bonds. In a simple *FF*, the non-bonded term is usually modelled using a Coulomb potential term for electrostatic interactions, where Q are the charges and r_{ij} the distances, and a Lennard-Jones or Buckingham potential for Van der Waals interactions, where ε_{ij} and σ_{ij} control the depth and position (interatomic distance) of the potential energy well for a given pair of non-bonded interacting atoms.

The *FF*, thus, enables the potential energy of a molecule (or of a system of molecules) to be calculated rapidly and pretty accurately. It also allows describing the energy changes of the molecule caused by internal system changes, like rotations around a bond, as well as the interactions between non-bonded parts of the system. More sophisticated *FF* may have additional terms, but they contain the same four fundamental components.
Few important features characterize a Molecular Mechanics Force Field:

- *The parameter set implemented in the functional form*. Parameters quantitatively define the single energy contributions for each group of interacting atoms and, as a consequence, they govern the computation of the whole energy function.
- *Transferability of parameters*. The same set of parameters can be used to model a series of related molecules, not explicitly included during the parameter optimisation, rather than having to define a new set of parameters for each individual molecule. Transferability has some limitations: the larger the number of parameters that are extrapolated, the lower the accuracy of the force field.
- *The empirical form*. There is not an "a priori" form for a *FF*. The functions of a *FF* very often are meant to offer a compromise between accuracy and computational efficiency: the most accurate functional form may often be unsatisfactory for efficient computation.
- *The Atom Type concept*. It is more that the simple atomic number. It contains information about the hybridization state (i.e. an implicit description of the motion of its electrons) and, sometimes, about the local environment of an atom. When preparing an input for *MM* it is necessary to assign an atom type for each atom in the system.

The parameterization of the *FF* represents the most difficult and time-consuming step in a *MM* calculation. Once the right functional form for describing the system has been chosen, one has to decide which set of parameters to introduce. Derived parameters are expected to be transferable to other classes of molecules. Transferability is one of the most important properties of a force field.

1.4 ENERGY MINIMIZATION METHODS

The most popular application of the empirical potential energy function is to find the geometry of a molecule (or an assemblage of molecules) which corresponds to a minimum of the potential energy function. In MM, the energy of a molecule in its ground electronic state is a function of only the coordinates of its atoms. If nuclei move, the energy changes. Such changes in energy can be considered as displacements on a multidimensional surface, called the Potential Energy Surface

(PES).

The minimization of the potential energy function (i.e., geometry optimization) involves a search for the minimum of a function and usually requires calculations of derivatives of the potential energy function versus independent variables (in our case, coordinates). Most programs use cartesian coordinates as independent variables, however, in some cases, internal coordinates may be used. The derivatives of potential energy are denoted as:

$$g_i = \frac{\partial V}{\partial x_i}; \qquad H_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j} \tag{1.4}$$

where g_i is the *gradient* (i.e., first derivative) of the potential energy V with respect to a cartesian coordinate x_i of an atom; H_{ij}, called *Hessian matrix*, is the second derivative of the energy with respect to the cartesian coordinates. In most modern programs these derivatives are calculated analytically, i.e., the appropriate mathematical formulae for corresponding terms are incorporated into the program. Some older codes compute derivatives numerically by approximating the slope of an energy function (or its gradient in the case of second derivatives) from finite differences. The derivatives are used not only in function minimization but also yield forces acting on atoms (from energy gradients) and normal modes of vibration (from the Hessian matrix).

There are three major approaches to find a minimum of a function of many variables:

- **Search Methods** -- utilize only values of the function itself. They are usually slow and inefficient, but are very simple to program, since deriving cumbersome formulas for derivatives is not necessary. In spite of their inefficiency, the search algorithms are infalliable and always find a minimum. For this reason, they are often used as an initial step, when the starting point in optimization is far from the minimum. Another disadvantage of search techniques is that they are very inefficient for a large number of optimized variables and converge very slowly when the number of variables is more then 10.
- **Gradient Methods** -- utilize values of a function and its gradients. These are currently the most popular methods in molecular mechanics. They offer a much better convergence rate than search methods and do not require a lot of computer memory (only $3N$ first derivatives are needed). However, in some situations they fail to converge to a minimum. The *conjugated gradient* algorithm is considered the most robust in this class.
- **Newton Methods** -- are the most rapidly converging algorithms which require values of function, and its first and second derivatives. The memory required for storing the Hessian

matrix is proportional to N^2 (i.e., prohibitive for large macromolecules). The BFGS algorithm is considered the most refined one.

In general, the minimization methods are iterative. They require on input some initial estimate for the position of the minimum, and provide a better estimate for the minimum as a result. This corrected estimate is used as an input into the next cycle (i.e., iteration) and the process is continued until there is no significant improvement in the position of the minimum.

Most search methods and minimization methods using derivatives are the *descent series methods*, i.e., each iteration results in a solution which corresponds to a lower (or equal) value for the energy function:

$$V\left(x^{(start)}\right) \geq V\left(x^{(1)}\right) \geq V\left(x^{(2)}\right) \ldots \geq V\left(x^{(\min)}\right). \qquad (1.5)$$

As a consequence, these methods can only find the minimum closest to the starting estimate and will never cross to a minimum (however deep) if it is separated from the starting estimate by a maximum (however small). There is no general way of finding a global minimum (i.e., the minimum corresponding to the lowest possible value of the function). A different initial geometry will usually lead to a different final minimum.
Only on very simple molecules will the single geometry optimization yield the global minimum on the first trial. To find a global minimum one has to perform many minimizations and use different initial coordinates for each run.[2]

1.5 MOLECULAR DYNAMICS METHODS

Computer simulation methods allow the analysis of complex systems, by producing replications of the macroscopic system with a handy and manageable number of particles. A computer simulation generates a representative ensemble of possible configurations of these small replications: in this way accurate calculations of structural and thermodynamic properties can be performed, by analysing the mechanical properties of molecules. Therefore the behaviour of the system in time can be studied and properties such as internal energy, entropy, pressure, temperature and so on, can be determined.

MD simulations address numerical solutions of Newton's equations of motion on an atomistic or

similar model of a molecular system. In fact, all of the information needed to calculate the dynamics of a system can be found from the potential energy function *V* of the system.
The force *F* on atom *i* in the system can then be determined from the equation:

$$F_i = -\nabla_i V \tag{1.6}$$

Using the Newton classical approximation, MD simulates the motion of particles in a system they react to forces caused by interactions with other particles. Forces so evaluated are used to determine accelerations. Particle velocities are initially determined by a random distribution, but then they are updated according to the calculated accelerations.
For the continuous nature of the potential functions describing interactions between atoms or molecules, it is necessary to integrate the equations of motion by dividing the calculation into a series of short time steps, which should be at least one order of magnitude shorter than the shortest motion simulated. An important assumption to be made is to consider forces acting on the atoms constant over the time-interval: at each step forces are recomputed and a new set of accelerations, velocities and positions are obtained. Following this technique, MD simulations generate a trajectory of the system describing its evolution over time.

The general property *A* of the system is calculated as an average upon all the *M* visited states:

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A_i \left(q^N , p^N \right) \tag{1.7}$$

where *q* refers to the coordinates and *p* to the linear momenta of the *N* particles constituting the system.
MD simulations can thus be considered as a deterministic method: they provide information about the "real" evolution of the system over time, and they allow to go back over past states of the system as well as to predict future arrangement of its particles. This dynamical view of molecular systems thus provides a useful and important tool for studying time-dependent processes.

1.6 STEPS IN A MD SIMULATION

The first thing before starting with a MD simulation is to decide which *FF* to use to model the

interactions between atoms or molecules in the system.
A simulation can then be described according to four principal points:

1) *Choice of the initial configuration*. This is a crucial moment of the entire simulation. It's very important to set up starting configuration of the system as much as possible similar to the real conformation; in fact, wrong starting coordinates may compromise the whole simulation process. Generally, homogeneous liquids (i.e., composed by molecules of the same type) are described by a standard lattice structure (for example, a face-centred cubic lattice) as starting configuration. The dimensions of the lattice are chosen in such a way to respect as much as possible the real density of the simulated systems. Usually, before proceeding with the simulation, a first minimization of the system energy is required in order to eliminate any term of high energy, which may cause instability in the simulation.

2) *Equilibration phase*. The system is allowed to evolve from the initial configuration until certain stability in the simulation is reached. At this stage, thermodynamic and structural properties, such as energy, temperature, pressure, are monitored: once their values have become stable, equilibration is reached. Order parameters can be also used to check when an equilibration phase can be considered completed.

3) *Production phase*. This is the real simulation stage. The system is set free to evolve and it is possible to calculate reliable properties.

4) *Analysis*. Properties not calculated during the simulation from the molecular mechanics program are evaluated and the configurations produced (and stored) are examined. This phase is important not only to know how the system changes, but also to check if any problems occurred during the simulation after the equilibration step.

When starting an *MD* simulation, the initial velocities of all the molecules must be specified: this usually is done by randomly selecting a set of velocities from the Maxwell-Boltzmann's distribution at the temperature of the simulation.

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T} \right)^{1/2} \exp\left(\frac{-\frac{1}{2} m_i v_{ix}^2}{k_B T} \right) \tag{1.8}$$

The Gaussian distribution of Equation 2.8 gives the probability $p(v_{ix})$ that an atom i of mass m_i, has a velocity v_{ix} in the x direction at the temperature T. Initial velocities are usually adjusted to give a zero total linear momentum:

$$P = \sum_{i=1}^{N} m_i v_i = 0 \tag{1.9}$$

The normal process of equilibration will then redistribute the energy amongst the different degrees of freedom. Precise adjustments to the kinetic temperature are made by scaling velocities during equilibration.

Careful monitoring of the behaviour of properties during the simulation can help to check if problems occur, and in this unfortunately case, the simulation has to be restarted from scratch after removing the cause of the problem.

1.7 FINITE DIFFERENCE METHODS IN MD SIMULATIONS

Finite difference methods are the numerical recipes used in *MD* simulations to integrate equations of motion and to generate trajectories, under the assumption that the energy potential terms are pair wise additive.

If we consider a system of atoms, with Cartesian coordinates $\boldsymbol{r}_i$ and the usual definition of K and V then the equation of motion becomes:

$$m_i \ddot{r}_i = F_i \tag{1.10}$$

where m_i is the mass of atom i and F_i is defined by Equation 1.6.

For a given *FF* characterizing the physical system, the integration method is responsible for the accuracy of the simulation results. If the integration method works correctly, the simulation will provide exact results, within the errors due to the computer finite number representation. However, any finite difference method is naturally an approximation for a system evolving continuously in time. An integration algorithm or integrator is required to have some well defined features such as:

- *Accuracy*. It has to approximate the true trajectory.
- *Stability*. It has to avoid small perturbations generating numerical instabilities.

- *Robustness*. It should allow integrations for relatively long time steps.

A standard method for solution of ordinary differential equations is the finite difference approach. Given the molecular positions, velocities, and other dynamic information at time t, we attempt to obtain the positions, velocities etc. at a later time $t+\delta t$. The equations are solved on a step-by-step basis; the choice of the time interval δt will depend somewhat on the method of solution, but δt will be significantly smaller than the typical time taken for a molecule to travel its own length.
The simplest and most straightforward way to construct an integrator is by expanding positions and velocities in Taylor series. Dividing the simulation in fixed time intervals, δt, the expansion reads:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \frac{1}{6}\delta t^3 b(t) + \ldots \quad (1.11)$$

$$v(t+\delta t) = v(t) + \delta t a(t) + \frac{1}{2}\delta t^2 b(t) + \ldots \quad (1.12)$$

$$a(t+\delta t) = a(t) + \delta t b(t) + \ldots \quad (1.13)$$

where v is the velocity, a the acceleration, and b the third derivate, and so on.
The *Verlet* algorithm[3] is probably the most used method for integrating the equations of motion in MD simulation. This method uses the positions and the accelerations at the time t, and the positions from the previous step, $r(t-\delta t)$, to calculate the new positions at $t+\delta t$. The Verlet algorithm equations are written in the following way:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \ldots \quad (1.14)$$

$$r(t-\delta t) = r(t) - \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \ldots \quad (1.15)$$

By adding the two last equations one obtains:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2 a(t) \quad (1.16)$$

In the Verlet integration algorithm velocities do not appear explicitly, but they can be calculated in

several ways. One of these is the following:

$$v(t) = [r(t+\delta t) - r(t-\delta t)] / 2\delta t \tag{1.17}$$

Implementation of the *Verlet* algorithm is straightforward and the storage requirements are modest and include two sets of positions (*r(t)* and *r(t-δt)*) and the accelerations, *a(t)*. One of its drawbacks is that positions *r(t+δt)* are obtained by adding a small term, $\delta t^2 a(t)$, to the difference of two much larger terms (see Eq. 2.17). This may cause a loss of precision. The *Verlet* algorithm shows other problems, like the difficulty to calculate the velocities, which are not available until the positions have been computed at the next step. In addition, it is not self-starting: the new positions are obtained from the current positions $r(t)$ and the positions from the previous step, $r(t-\delta t)$. At $t = 0$, there is only one set of coordinates and it is necessary to employ some other ways to obtain positions at time, $t-\delta t$.

A large number of variations of the Verlet algorithm have been developed:

- *The velocity Verlet* method[4] evaluates positions, velocities and accelerations at the same time and this does not affect negatively the precision of the calculation:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{1.18}$$

$$v(t+\delta t) = v(t) + \frac{1}{2}\delta t[a(t) + a(t+\delta t)] \tag{1.19}$$

The velocity Verlet algorithm is actually implemented as a three-stage procedure, the new velocities requiring accelerations at the times t and $t+\delta t$. Thus, as first step, positions at the time $t+\delta t$ are calculated, using velocities and accelerations at time t, and then, velocities at time $t+\frac{1}{2}\delta t$ are determined, using the equation:

$$v\left(t + \frac{1}{2}\delta t\right) = v(t) + \frac{1}{2}\delta t a(t) \tag{1.20}$$

The new forces are then computed from the current positions, thus giving $a(t+\delta t)$. In the final step, the velocities at time $t+\delta t$ are calculated using the following relation:

$$v(t+\delta t) = v\left(t+\frac{1}{2}\delta t\right) + \frac{1}{2}\delta t\, a(t+\delta t) \tag{1.21}$$

- The *Beeman's* algorithm[5] uses a more accurate expression for the velocities, and, as a consequence, gives a better energy conservation and the kinetic energy can be calculated directly from the velocities:

$$r(t+\delta t) = r(t) + \delta t\, v(t) + \frac{2}{3}\delta t^2 a(t) - \frac{1}{6}\delta t^2 a(t-\delta t) \tag{1.22}$$

$$v(t+\delta t) = v(t) + \frac{1}{3}\delta t\, a(t) + \frac{5}{6}\delta t\, a(t) - \frac{1}{6}\delta t\, a(t-\delta t) \tag{1.23}$$

All these methods have similar accuracies and are expected to produce identical trajectories in coordinate space.

1.8 MD SIMULATIONS AT CONSTANT TEMPERATURE AND PRESSURE

Molecular dynamics simulations can be performed sampling the phase space of the system considered in ensembles: the most frequently used are the *NVE* or *microcanonical* ensemble, the *NVT* or *canonical* ensemble, the *NPT* or *isothermal-isobaric* ensemble, and the μVT or *grand canonical* ensemble.[1]

The need to maintain the temperature constant during a simulation arises from different reasons. For example, one may wish to know how a system behaves under certain temperature conditions, such as for the unfolding of protein, or in a phase transition or, also, if an annealing process has to be simulated. Moreover, it is worthwhile remembering that the temperature can be considered as an external stimulus affecting the macroscopic behaviour of a given system.

Being the temperature of the system closely related to the time average of the kinetic energy, it can be left unchanged by scaling the velocities[6] of the particles, with a multiplying factor λ, or by coupling the simulated system to an external bath[7] with a constant temperature. In the first case, the relative temperature change is given by the following equations:

$$\Delta T = \frac{1}{2}\sum_{i=1}^{N}\frac{2}{3}\frac{m_i(\lambda v_i)^2}{Nk_B} - \frac{1}{2}\sum_{i=1}^{N}\frac{2}{3}\frac{m_i v_i^2}{Nk_B} \tag{1.24}$$

$$\Delta T = (\lambda^2 - 1)T(t) \tag{1.25}$$

$$\lambda = \sqrt{T_{new}/T(t)} \tag{1.26}$$

In the second treatment, the bath acts as a source of thermal energy, adding or removing heat from the system introducing the possibility to change atomic velocities at each step. The rate of change of temperature is proportional to the difference in temperature between the bath and the system:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{bath} - T(t)) \tag{1.27}$$

The scaling factor for the velocities reads:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right) \tag{1.28}$$

If τ is large, then the coupling is weak. If τ is small, the coupling is strong. When the coupling parameter equals the time step, the algorithm becomes equivalent to the simple velocity scaling method.

In the same way, one may wish to keep the pressure constant during a simulation: this enables the study of certain phenomena such as the onset of pressure induced phase transitions. Many methods used for pressure control are similar to those used for temperature: the pressure is maintained constant by simply scaling the volume, or by coupling the system to an external pressure bath. The rate of the pressure change is given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} - P(t)) \tag{1.29}$$

τ_P is the coupling constant, P_{bath} is the pressure of the bath, and *P(t)* is the actual pressure at time *t*. Introducing the system compressibility, *k*, the volume of the simulation box is scaled by a factor λ,

equivalent to scaling the positions by $\lambda^{1/3}$. Thus:

$$\lambda = 1 - k\frac{\delta t}{\tau_P}(P - P_{bath}) \tag{1.30}$$

$$r_i' = \lambda^{1/3} r_i \tag{1.31}$$

1.9 OPLS FORCE FIELD

Of all the several force field available, the OPLS (Optimized Potentials for Liquid Simulations)[16] is one of the most suitable to describe our kind of system (proteins).
In this FF, the nonbonded interactions are represented by the Coulomb plus Lennard-Jones terms in Equation 2.42, where E_{ab} is the interaction energy between molecules *a* and *b*:

$$E_{ab} = \sum_{i}^{on\cdot a}\sum_{j}^{on\cdot b}\left[\frac{q_i q_j e^2}{r_{ij}} + 4\varepsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r^{12}} - \frac{\sigma_{ij}^{6}}{r^{6}}\right)\right] f_{ij} \tag{1.32}$$

Standard combining rules are used such that $\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{12}$ and $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{12}$. The same expression is used for intramolecular nonbonded interactions between all pairs of atoms (i < j) separated by three or more bonds. Furthermore, f_{ij} = 1.0 except for intramolecular 1,4-interactions for which f_{ij} = 0.5. Nonbonded interactions are also evaluated for intramolecular atom pairs separated by three or more bonds. It was found to be necessary to scale the 1,4-nonbonded interactions to permit use of the same parameters for inter- and intramolecular interactions. Scaling factors f_{ij} = 1/2 for both the Coulombic and Lennard-Jones interactions emerged as the final choice.
The energetics for bond stretching and angle bending are represented by Equations. 1.33 and 1.34.

$$E_{bond} = \sum_{bonds} K_r (r - r_{eq})^2 \tag{1.33}$$

$$E_{angle} = \sum_{angle} K_\theta (\theta - \theta_{eq})^2 \tag{1.34}$$

The last intramolecular term is for the torsional energy (Eq. 1.35),

$$E_{torsion} = \sum_i \frac{V_1^i}{2}[1+\cos(\varphi_i + f_i 1)] + \frac{V_2^i}{2}[1-\cos(2\varphi_i + f_i 2)] +$$

$$+\frac{V_3^i}{2}[1+\cos(3\varphi_i + f_i 3)] \qquad (1.35)$$

where φ_i is the dihedral angle, V_1, V_2 and V_3 are coefficient in the Fourier series, and *f1*, *f2* and *f3* are phase angles, which are all zero for the present system. The total torsional energy, $E_{torsion}$, is then the sum of this series for each dihedral angle.

The general equations of the OPLS force field read:

$$E = E_{bond} + E_{angle} + E_{torsion} + E_{ab} \qquad (1.36)$$

1.10 TINKER: A MOLECULAR MODELING PACKAGE

The computer simulations of collapsing bubbles were performed with TINKER[22,25] a molecular modeling package designed to be a user friendly system of programs and routines for Molecular Modeling Mechanics and Dynamics. It is intended to be enough modular to enable development of new computational methods and enough efficient to meet most production calculation needs. Rather than incorporating all the functionality in one monolithic program, *TINKER* provides a set of relatively small programs that interoperate to perform complex computations. The most important tasks performed by the program are:

1) Build protein and nucleic acid models from sequence.
2) Energy minimisation and structural optimisation.
3) Analysis of energy distribution within a structure.
4) Molecular and stochastic dynamic simulations.
5) Simulated annealing with a choice of cooling schedules.
6) Normal modes and vibrational frequencies.
7) Conformational search and global optimisation.
8) Transition state location and conformational pathways.
9) Fitting of energy parameters to crystal data.
10) Distance geometry with pairwise metrization.

11) Molecular volumes and surface areas.
12) Free energy changes for structural mutations.
13) Advanced algorithms based on potential smoothing.

The basic design the *TINKER* program allows the use of several different parameter sets. At present, the distributed and implemented parameters set are: MM2,[17] MM3,[18] OPLS/OPLS-AA,[19] AMBER-95,[20] CHARMM27.[21]

Many of the various energy minimisation and *MD* calculations can be performed on fully or partial geometries, over Cartesian, internal or rigid body coordinates, and including a variety of boundary conditions and crystal cell types. *TINKER* differs from many other currently available *MM* programs by the possibility given to the user to modify the source code that is extensively commented. The distributed individual routines should be considered as a template for the user who wants to introduce new features to the main program. The core of *TINKER* consists of about 110'000 lines written in Fortran77. Both spherical cutoff images and replicates of a cell are supported by all *TINKER* programs that implement *PBC*. Whenever the cutoff distance is too large for the minimum image to be the only relevant neighbour, *TINKER* automatically switches off the image formalism to use replicate cells.
During the present PhD thesis, we have used the main TINKER programs, with some modifications in order to fit to our type of problem.

1.11 REFERENCES

[1] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Oxford U. Press, New York, 1987.

[2] A. R. Leach, *Molecular Modelling Principles and Applications 2nd Ed.*, Pearson Education, London., 2002.

[3] L. Verlet, *Phys. Rev.* **1967**, *158*, 98-103.

[4] W. C. Swope, H. C. Anderson, P. H. Berens, K. R. Wilson, *J. Chem. Phys.* **1982**, 76, 637-649.

[5] D. Beeman, *J. Comp. Phys*. **1976**, *20*, 130-139.

[6] L. V. Woodkock, *Chem. Phys. Lett.* **1971**, *10*, 257-261.

[7] H. J. C. Berendensen, P. M. Postma, W. F. van Gunsteren, A. Di Nola, J. R. Haak, *J. Chem.*

Phys. **1984**, *81*, 3684-3690.

[8] P. Ewald, *Annalen der Physik* **1921**, *64*, 253-287.

[9] A. R. Leach, *Molecular Modelling Principles and Applications 2nd Ed.*, **2002**, Pearson Education, London, *Chapter 6*.

[10] A. Wallqvist, R. D. Mountain, *Rev. Comput. Chem.* **1999**, *13*, 183-247.

[11] W. L. Jorgensen, J. Tirado-Rives, *Proc. Natl. Acad. Sci.* **2005**, *102*, 6665-6670.

[12] A. Brodsky, *Chem. Phys. Lett.* **1996**, *261*, 563-568.

[13] B.Guillot, *J. Mol. Liquids* **2002,***101*,219-260.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, J. Hermans, in B. Pullman (ed.), *Intermolecular Forces* (Reidel, Dordrecht, 1981), p331.

[15] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem.* **1987**, *91*, 6269-6271.

[16] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

[17] N.L. Allinger, .*J. Am. Chem. Soc.* **1977**, *99*, 8127-8134.

[18] N. L. Allinger, Y. H. Yuh ,J.-J. Lii, *J. Am. Chem. Soc.* **1989**, *111*, 8551-9556.

[19] W.L. Jorgensen, J. Tirado Rivers, *J. Am. Chem. Soc.* **1988**,*110*, 1657-1666.

[20] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

[21] N. Foloppe, A. D. MacKerell Jr., *J. Comput. Chem.* **2000**, *21*, 86-104.

[22] J. W. Ponder, F. J. Richards, *J. Comput. Chem.* **1987**, *8*, 1016-1024.

[23] C. Kundrot, J. W. Ponder, J. W. Richards, *J. Comput. Chem.* **1991**, *12*, 402-409.

[24] M. J. Dudek, J. W. Ponder, *J. Comput. Chem.* **1995**, *16*, 791-816.

[25] *TINKER. Software Tools for Molecular Design*. http://dasher.wustl.edu/tinker

CHAPTER 2

THE INTRA-RESIDUE DISTRIBUTION OF ENERGY IN PROTEINS

2.1 INTRODUCTION

The analysis of the frequency of appearance of specific structural arrangements of amino acid residues in proteins has had a profound impact on the investigation of protein structures and has brought, among the rest, to the development of statistical potentials that are rather accurate in the prediction of folding patterns.[1-6] Such potentials are extremely versatile and effective. They compete directly, or may even be superior to, with those used in molecular mechanics and molecular dynamics simulations.[7-8] Conversion of frequencies of occurrence into potentials implicitly assumes that the sample of protein structures of the database behaves like, or reflects, the dynamics of the proteins, where Boltzmann distribution is valid.

While conformance to the *Boltzmann hypothesis, Bh,* must be understood as a qualitative - although very useful - statement,[9] examples of protein properties, and related energy contributions, that conform to it, are numerous and include hydrophobicity,[10,11] various types of sidechain/side-chain interactions,[12-14] proline-isomerization,[15] hydrogen bonds,[16] internal cavities,[17] interactions at the level of specific atom types,[18-19] and the propensity of the ϕ/φ ratio.[20,21]

Here, we test the conjecture that *Bh* applies to the deformation energy of the individual, naturally occurring amino acids, AA, in a database of highly resolved protein structures of nearly 200 proteins. The deformations are the sum of strain contributions for stretching, bending and torsions, plus variations from equilibrium of other energy terms such as van der Waals and Coulomb interactions. Molecular mechanics is used to calculate the energies of 41672 residues that are assessed in the light of *Bh*, both globally and divided according to the nature of the residue. The picture that emerges is that Boltzmann distribution holds for the intra-residue energy distribution of the single residues. When the focus shifts from the single types of residues to the individual proteins, the energy distribution, often (~50%) takes the form of a Poisson distribution characterized by the same parameters of the entire set of proteins.

2.2 BACKGROUND

Protein structures were obtained in pdb format from the PAPIA (Parallel Protein Information Analysis) service [http://www.cbrc.jp/papia/papia.html]. Only X-ray structures were downloaded, with resolution up to either 1.5 or to 1.6 Å, the minimal number of residues was set to 40, the sequence similarity was set to less than 20%. At the higher resolution, 136 proteins were obtained. Removal from the set of broken chains and chains with one or more unphysically over-distorted residues reduced them to 122. Analogously, at the slightly lower resolution, there were 75 additional structures. The selection procedure is similar to that adopted by Shortle in the very recent investigation of the Boltzmann distribution of the ϕ and φ angles.[21] A choice of resolutions larger than 1.6 Å was deemed to introduce too high an inaccuracy in the calculation of the internal energy of individual residues.

Individual residue energies were calculated with the TINKER program, [22-24] which has found a number of applications in our laboratory,[25-28] using the AMBER/OPLS/UA force field.[29-30] The united atom approach, UA, avoids the difficulty caused by the lack of hydrogen atoms in most of the structures. The initial and final residues were not included in the calculations. For each of the 41672 aminoacids, the energy was calculated with the "group" keyword of the program. For each type of residue, the lowest energy in the data set of the individual amino acids was subtracted. The energy values are therefore distortion energy, which is the sum of torsional, bending and stretching energies of deformation together with the local energy variation (internal to the residue) induced by the change of interatomic distances in the van der Waals and Coulomb terms.

The energies of residues were binned with a step size of 1 kcal mol^{-1}. Apart from being a practical tool, the bin size also reduces the inaccuracy of the energy calculated for each residue due to (i) the uncertainty of the atom positions in the X-ray structures and (ii) the computational model. The unsmoothed, histogram-like, data set, U, were compared with smoothed data set, S, obtained using analytical functions.

To test the hypothesis that smoothed and unsmoothed data are indistinguishable, we used a nonparametric, distribution free statistic. Notice that this distribution is NOT the Boltzmann distribution, but is the distribution of the differences between smoothed and unsmoothed data. The Kolmogorov-Smirnov, K-S, test[31] compared the Normalized Cumulative Distribution Functions, NCDF, (the third kind of distribution to appear in this treatment) of sets U and S. If the absolute value of the maximum difference between the two NCDF, D_{max}, exceeds a certain value, which is a function of the number of bins, the level of significance, p, is low (for

instance, <0.05) and the hypothesis is rejected. A rather similar treatment was recently reported for the characterization of the mass spectra of proteins[32] and the distribution of deformations in molecular crystals.[33]

2.3 RESULTS AND DISCUSSION

The initial conjecture of this work is that also the internal energy of the individual residues - in the protein structural database - follows Boltzmann distribution. The theory of the applicability of the Boltzmann distribution to biomolecules has been discussed by Grzybowski et al.[34] In particular, they recognized that a database of structures frozen in their minimum energy conformation does not represent a proper canonical ensemble of conformational states.[34]
Some care must be exerted to describe the distribution because the levels accessible by a distorted residue are accidentally degenerate. Indeed, a given AA residue can obtain energy E by deforming along its several internal coordinates in more than one way. For a given amino acid, r, the Boltzmann population, $P_r(E_r)$, is proportional to

$$P_r(E_r)\,\mu g_r(E_r)\exp\left(-\frac{E_r}{k_B T}\right) \qquad (2.1)$$

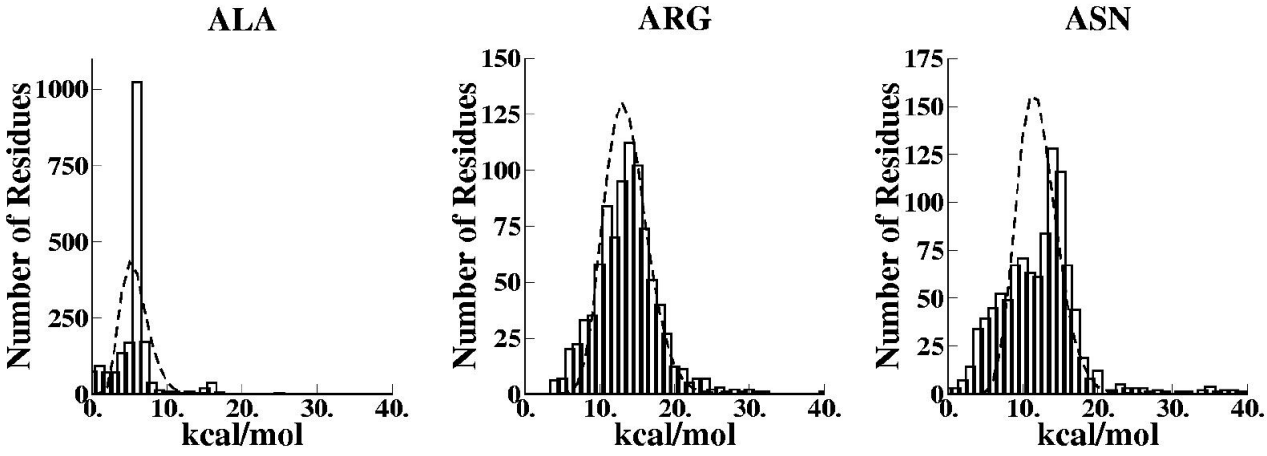
where $g_r(E_r)$ is the density of accessible states at energy E_r. In practice, *g(E)* counts the ways in which different deformations of a residue give a certain energy and grows very rapidly with the energy. Neither the numerical values of *g(E)*, nor the shape of the function are known and depend both on the number of atoms of the residue and on its plasticity. Intuitively, a small moiety deforms in fewer ways than a large one and is therefore characterized by a smaller *g(E)*; analogously, a rigid fragment is more difficult to distort than a floppy one, which must have a larger *g(E)*.
Because of the relative chemical similarity of the residues and although *g(E)* may be an entirely different function for each residue, it was decided to investigate if it could be taken as a simple function of the type

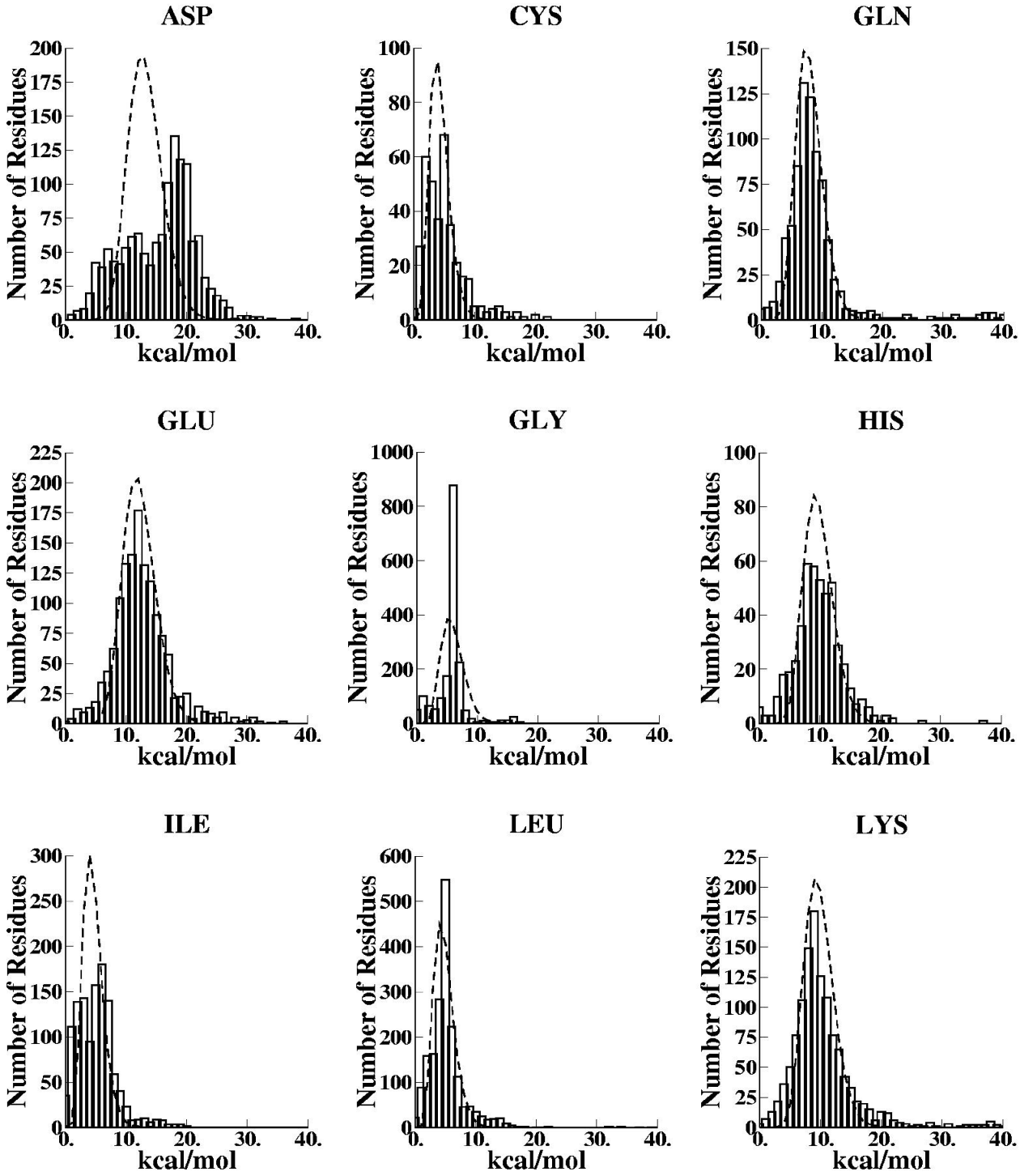
$$g_r(E_r) = E^{\alpha_r} \qquad (2.2)$$

where α_r differs for every one of the 20 naturally occurring amino acids. Equation 2.2 appears in the corresponding expression for the ideal gas[35] and was found to give the best statistical significance compared to other fast growing functions, such as the exponential one.
After calculating the energy of the individual AA for a sample of highly resolved protein structures (see Background section for details), they were binned in steps of 1 kcal mol^{-1}. The

smallest set was 653 residues for tryptophan, the largest sample were the 3650 residues of alanine. They generated 20 unsmoothed distributions that were compared to the function that results by substitution of eq. 2.2 into eq. 2.1, where the coefficient α_r was optimized numerically through the calculation of a grid of points to obtain the "best agreement" between unsmoothed, U, and smoothed, S, data and K_BT is set to room temperature. The role of K_BT in this context may be open to some discussion. Most proteins are crystallized at room temperature. However, some of the structures are measured and resolved at lower temperatures where, if they crystallized, they might assume another folding pattern thereby establishing different inter-residue interactions and, ultimately, different intra-residue energy. A compromise must be made and the use of the higher temperature was deemed more consistent. To check the stability of the procedure, K_BT was varied from the Room Temperature value but this lead to a worse agreement. Interestingly, Grzybowski et al.[34] noticed that a database of structures has a thermodynamic temperature that is given by the derivative of entropy (in their case, the conformational entropy) with respect to the energy.

Figure 2.1 shows the comparison between smooth and unsmooth distributions. The smooth distributions are obtained from the functions that are used in the Kolmogorov-Smirnov test and found to be a statistically significant description of the unsmooth data. In several cases, ARG, GLN, LYS, MET, PRO, the agreement appears quantitative. However, while the figures have an illustrative purpose, the quantitative agreement must not be sought at this level. The practical reason is that differences between U and S may add or subtract, along the distribution, in a "misleading" way.




ASP
CYS
GLN
GLU
GLY
HIS
ILE
LEU
LYS
Number of Residues
kcal/mol
0.
10.
20.
30.
40.

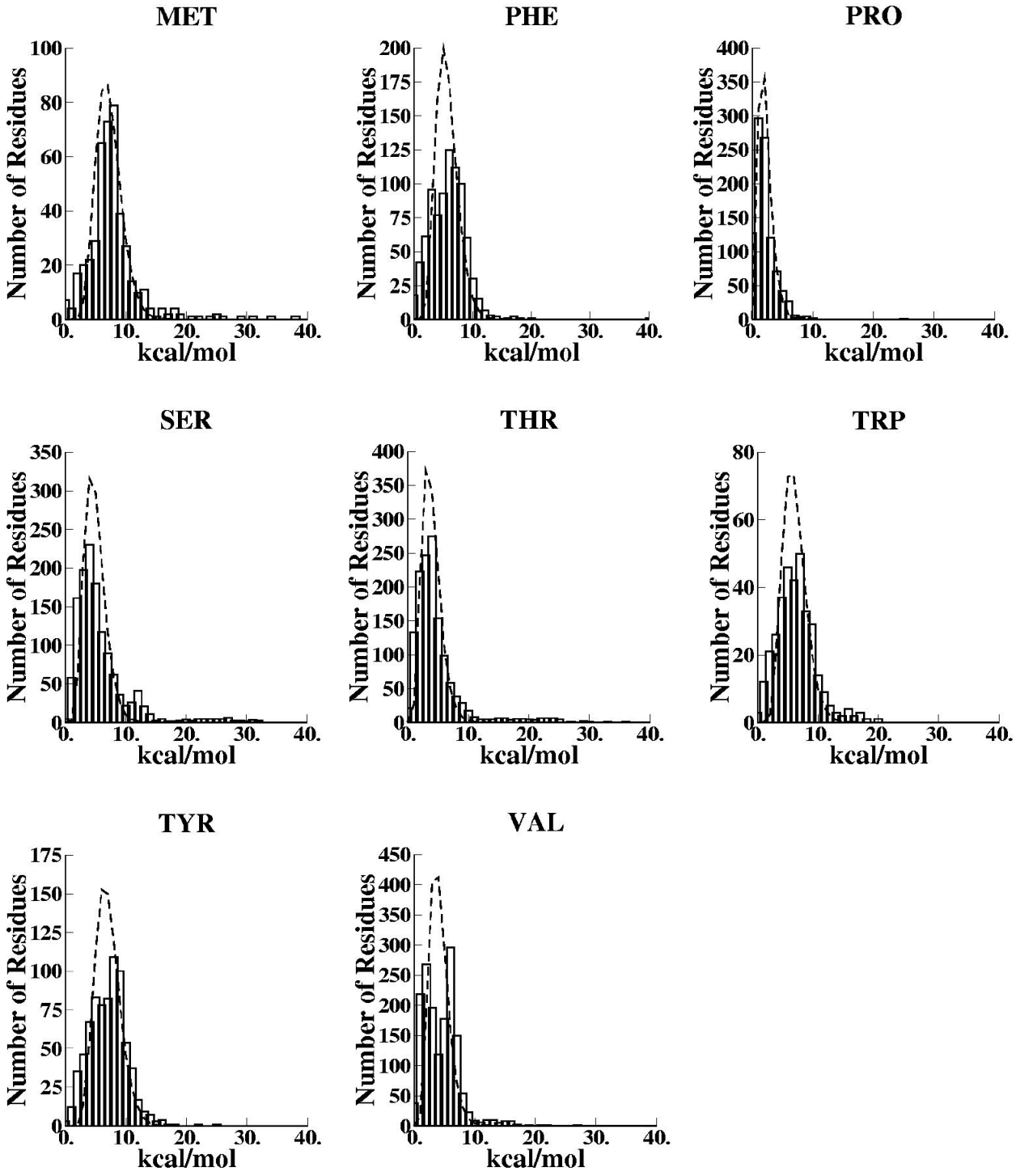
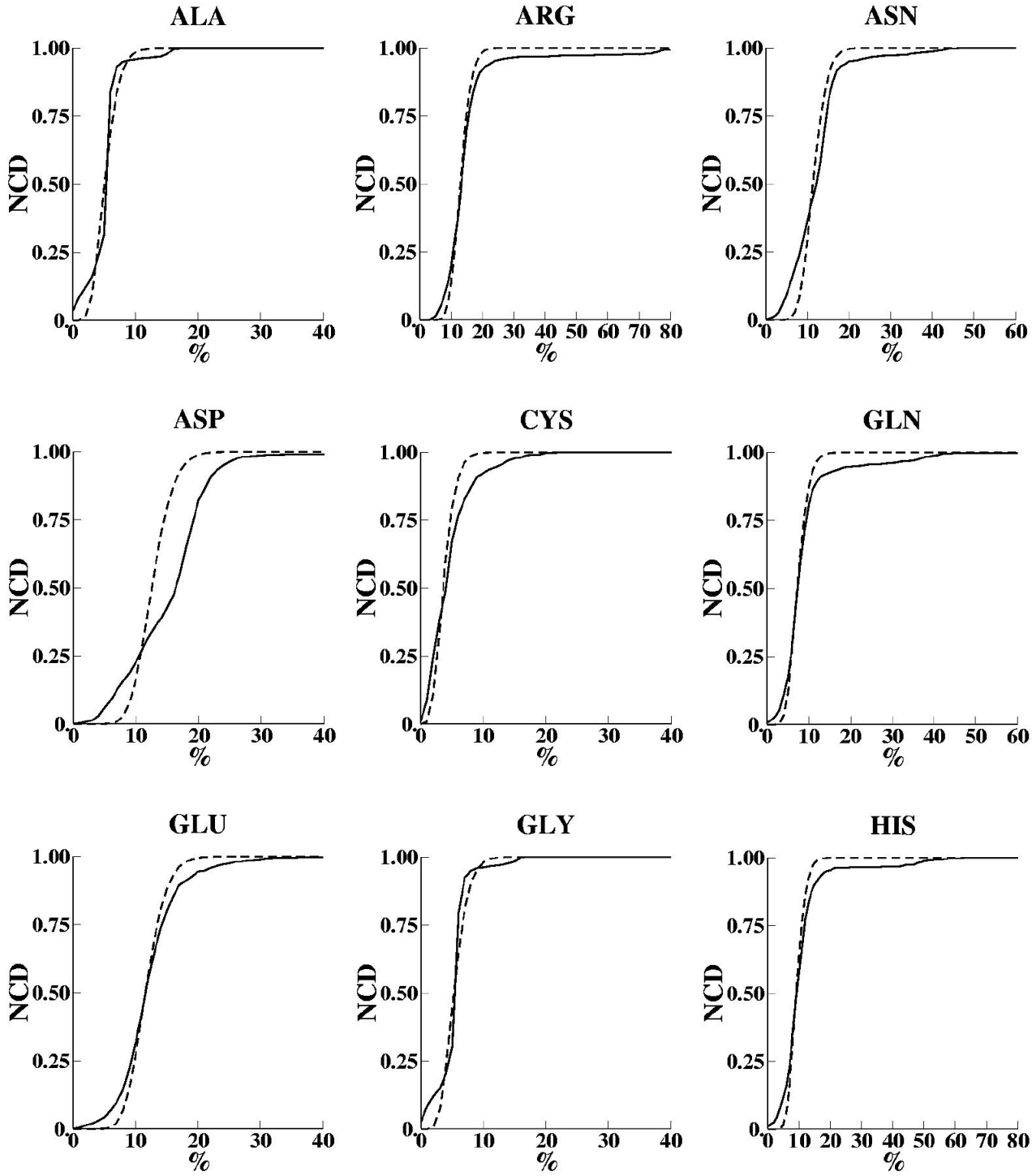
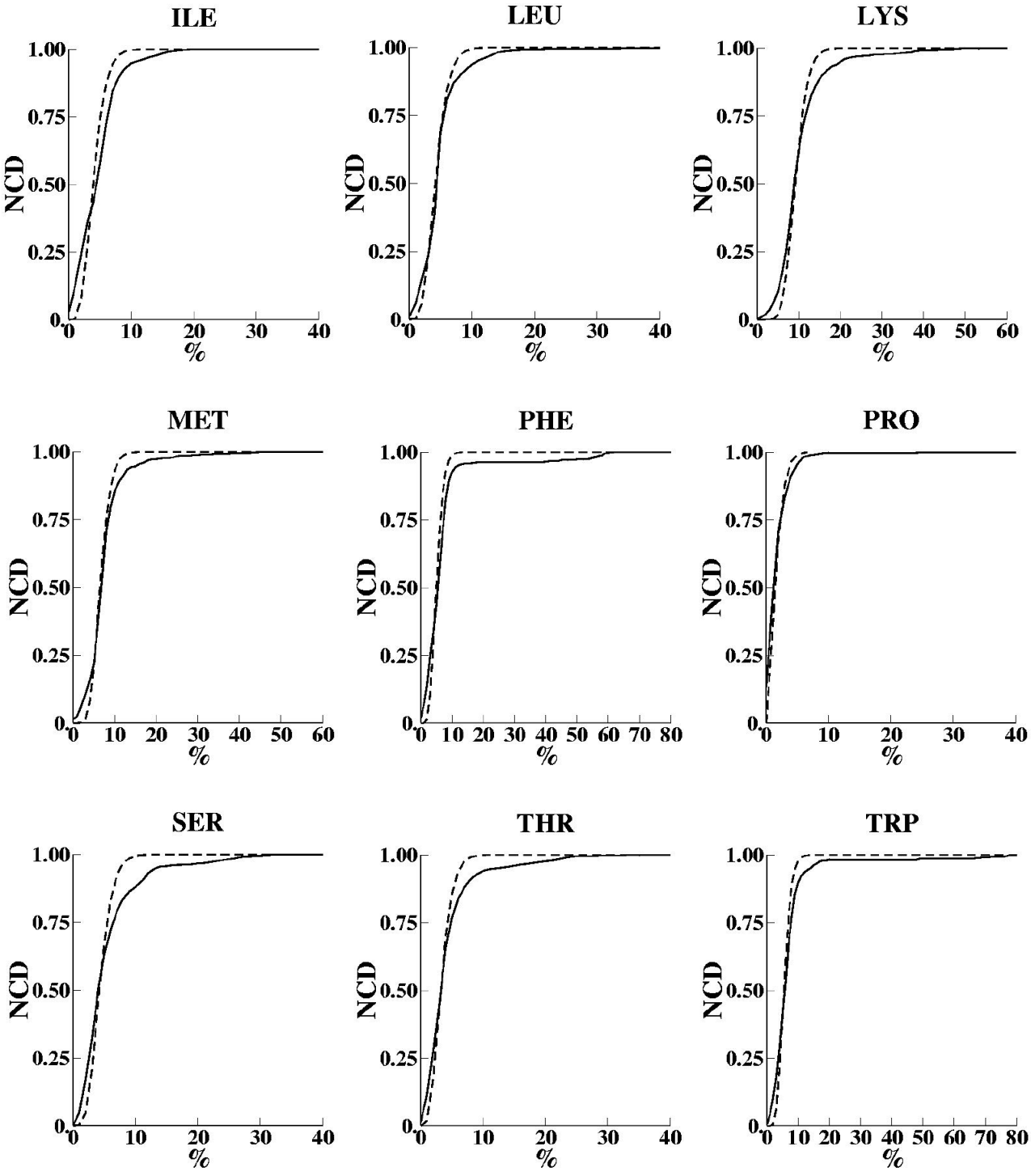


Figure 2.1. *Comparison of the smoothed and unsmoothed energy distributions of the 20 residues.*

The Kolmogorov-Smirnov test[31] on the Normalized Cumulative Distribution, NCD, was devised to appraise quantitative differences between distributions. Figure 2.2 compares such NCD. Visual inspection is rather satisfactory, although it is necessary to determine the largest

difference between the NCD and compare it with tabulated values[31] in order to determine the statistical significance and if U and S are indistinguishable. A similar treatment was recently reported for the characterization of the mass spectra of proteins[32] and the distribution of deformations in molecular crystals.[33]




ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
NCD
%
1.00
0.75
0.50
0.25
0
0 10 20 30 40
0 10 20 30 40
0 10 20 30 40 50 60
0 10 20 30 40 50 60
0 10 20 30 40 50 60 70 80
0 10 20 30 40
0 10 20 30 40
0 10 20 30 40
0 10 20 30 40 50 60 70 80

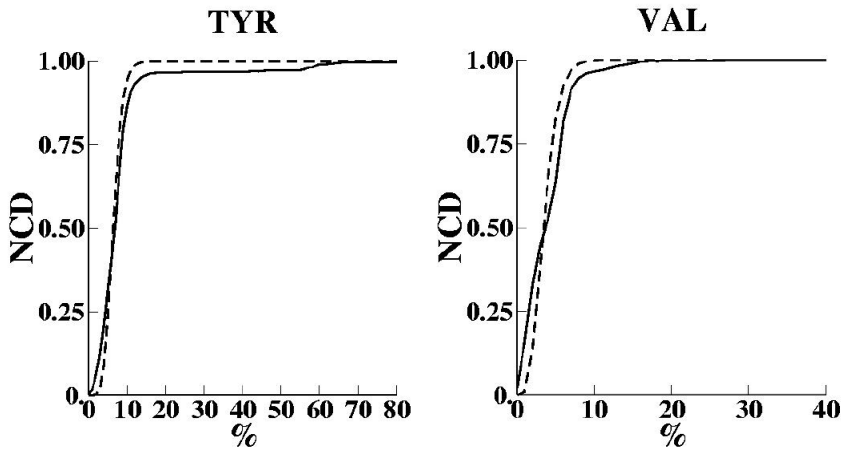


Figure 2.2. *Comparison of the normalized cumulative distributions, NCD, of the 20 residues. Table 2.1 shows a summary of the results. The best α_r values for the individual residues along with the p value. If p>0.05, U and S are considered indistinguishable.*

Residue	α_r/α_r'	p	Residue	α_r/α_r'	p	Residue	α_r/α_r'	p
ARG	21.4/15.41	0.64	THR	5.42/5.31	0.25	PHE	8.36/8.65	0.03
GLN	12.10/8.40	0.59	TRP	9.18/10.81	0.12	ALA	8.61/5.88	0.01
GLU	19.42/11.05	0.55	TYR	10.69/11.39	0.11	ASN	19.0/11.81	0.01
LEU	6.90/6.85	0.39	SER	7.20/5.25	0.08	ILE	6.65/7.38	0.007
MET	10.78/9.74	0.38	PRO	2.60/6.18	0.06	VAL	5.86/7.48	0.0011
HIS	15.39/9.59	0.26	CYS	6.16/6.06	0.04	ASP	21.0/11.88	$10.^{-9}$
LYS	15.37/10.04	0.25	GLY	8.91/4.63	0.03			

Table 2.1. *Residues, their α_r values and relative p values. For α_r', see text below.*

The statistical treatment gives some surprises. For instance, the good visual matches for proline both in figure 1 and 2 are highly penalized in terms of the statistical significance by the deviation that exists in the first bin. In any event, inspection shows that 12 out of 20 natural occurring AA residues have p>0.05. Decreasing the level of confidence to p>0.01 increases the number to 17 AA out of 20. It is tempting to suggest that if more highly resolved structures of proteins were available, all the residues would follow eq. 2.1. At this level of confidence, the distribution of their unsmoothed data set of energies is undistinguishable from that of the smoothed function.

The form of eq. 2.2, that is the energy dependence of the density of states, is critical for the results. On the one hand, the use of $g_r(E_r)$ functions with more than one parameter would likely

make all p_r>0.05. On the other hand, the present α_r values should be justified on a physical basis since even this one-parameter function results in a very steep growth. Physically, the functions measure the deformability/rigidity of each residue at given energy, which, in turn must depend on the vibrations of the system. In fact, a lower frequency implies a softer potential energy curve and a greater tolerance for distortion along the coordinate of the potential. When there are many degrees of freedom, and therefore vibrations, the softer they are the denser the manifold of vibrational levels is at a given energy. This is given by the convolution of the vibrational levels of the first vibration with those of the second, all convolved with the levels of the third one, and so on and so forth. One can therefore attempt to establish whether proportionality between number of (possible) deformations, *g(E)*, and density of vibrational states exists. In order to calculate the density of states, for each residue the geometry of a triplet of AA terminated by two GLYs was optimized and the fundamental vibrational frequencies were calculated, after removal of the GLYs frequencies. The densities of vibrational densities were determined by a time-honored algorithm,[36] that has found a variety of applications in our group,[37-38] and fitted to equation 2 to obtain α_ρ' values (reported in Table 2.1). Figure 2.3 compares the values of α_ρ and α_ρ'. In many cases the ratio of the two exponents is close to one and a correlation is clearly present, although its coefficient is not high, r=0.82, removal of 8 residues, brings it to r=0.91.

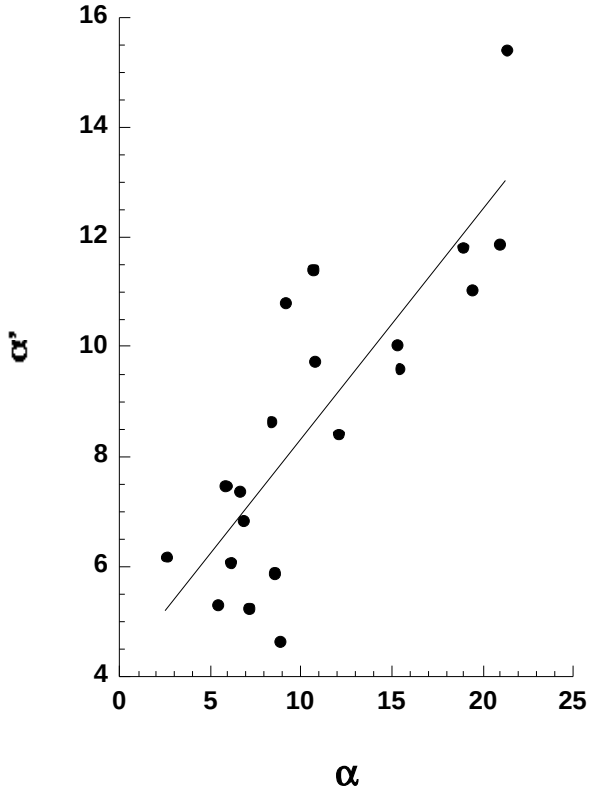


Figure 2.3. *Comparison of the exponents of eq. 2.2 with the exponents obtained from the calculation of the vibrational density of states for the 20 residues.*
In consideration of the simplification made to calculate the α_p' values, we felt that the g(E) functions of the residues were justified. In turn, this implies that the working hypothesis that the energy of distortion AA residues in proteins follows Boltzmann equation holds.

The first practical consequence of these results is that one can define a normalized probability for a single macromolecule. Based on the probability of distortion of the single residues, one can write

$$P_{B,protein} = \frac{1}{N}\prod_{r,i}^{N}\left(E_{r,i}\right)^{\alpha_r}\exp\left(-\frac{E_{r,i}}{k_B T}\right) \qquad (2.3)$$

Where N is the number of residues, $E_{r,i}$ is the energy calculated for the i-th residue of the r-th type and K_BT is the room temperature thermal energy. Figures 2.4a and 2.4b show the results of the application of eq. 2.3 to the 122 proteins of the sample and to the extra 75 proteins with

resolution between 1.5 and 1.6 Å. For convenience, the probability P is given as a logarithm and a subscript B is introduced to denote its Boltzmannian origin (see below). Apart from a few sporadic cases, $\ln(P_B)$ provides a homogeneous descriptor of the structure of existing proteins. Analysis of the handful of structures with low probability did not reveal the presence of any physical reason that should make us disregard them such as steric clashes. However, their clustering at high probabilities is taken as an additional proof that the description embedded in the underlying equations is valid and that the exponents of Table 2.1 are homogeneous.

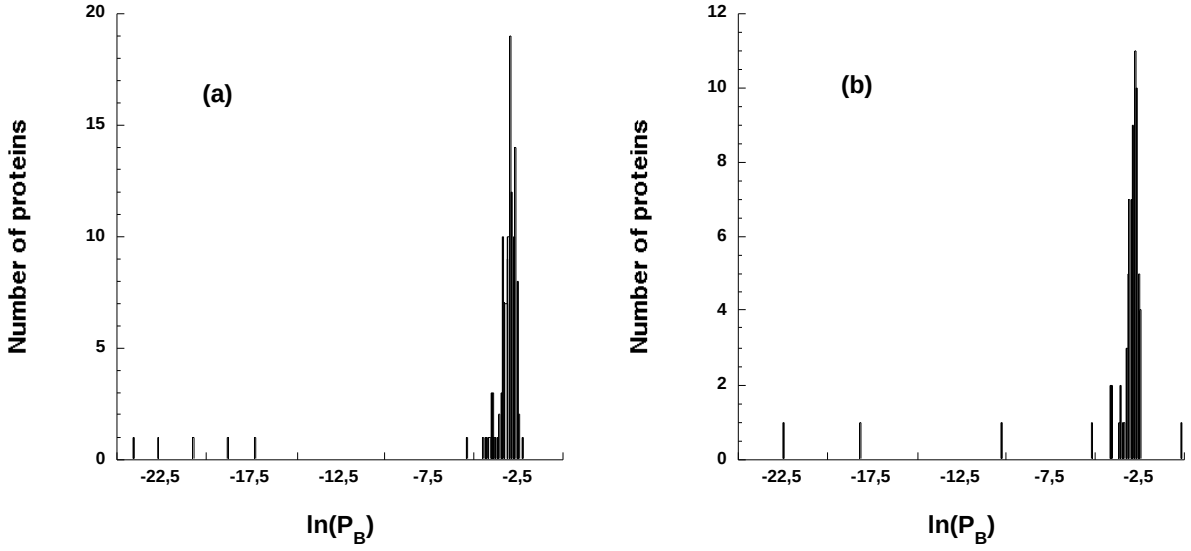


Figure 2.4. *$\ln P_B$, with P according to the Boltzmann-based eq. 2.3, of (a) the 122 proteins of the sample with 1.5 Å resolution; b) the 75 proteins of the sample with resolution between 1.5 and 1.6 Å (two proteins are not present because their value is below –25). The bin-size is 0.1. The subscript B is introduced to denote its Boltzmannian origin.*
The threshold value for $\ln(P_B)$ could be set to –5.0 (10 proteins would remain out of the two samples with this threshold value) and used in the evaluation of the significance of protein structures, for instance in the prediction of protein folding.

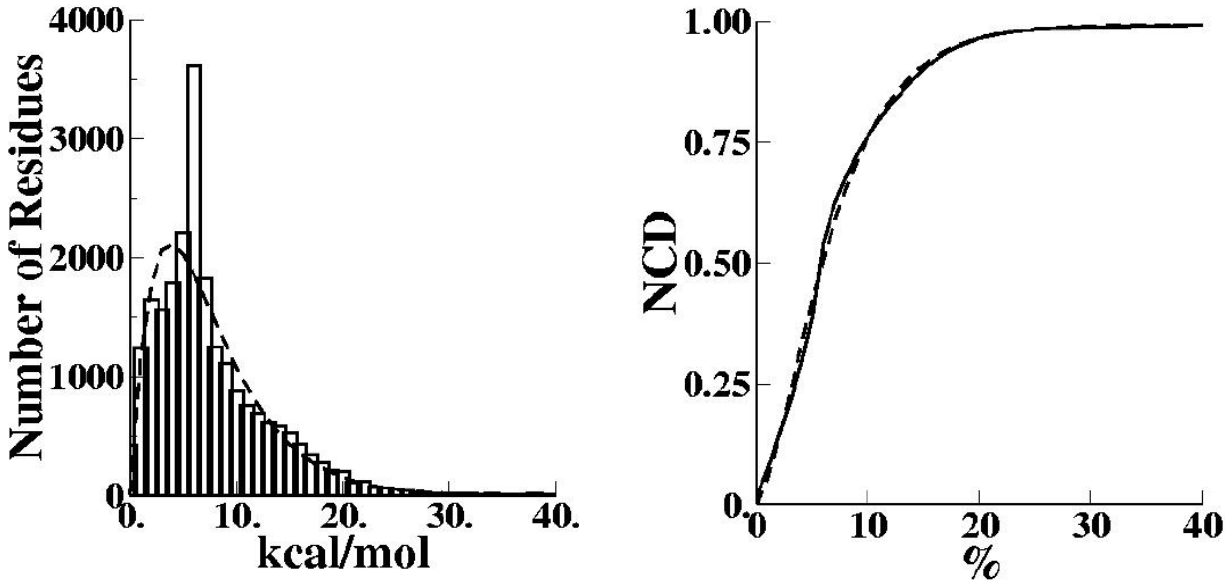
We now take a different approach and consider the energy of distortion of all the individual residues. This is a phenomenological approach that reveals a Poisson-like distribution. The data were therefore smoothed by the function

$$P_P(E) = \lambda E \exp(-\lambda E) \tag{2.4}$$

Where E is the energy of the individual residues not divided according to the type. The best

value for λ is 0.265, which gives p=0.975!! The values are the same, both when one considers the 22970 residues of the proteins with up to 1.5 Å resolution and when one extends the sample to the 41672 residues of the proteins with up to 1.6 Å resolutions. The extremely large p value means that the distribution of the energies of distortion in the database of proteins is indistinguishable by the function of eq. 2.4.

The Poisson distribution arises from counting comparatively rare events occurring in time, space, area, etc. Apart from the well-known example of the recording of radiation by means of a Geiger counter, other phenomena that follow the Poisson distribution involve the presence of errors such as the imperfections on a continuously produced bolt of cloth or the misprints in a book. The sum of the energy of deformation of the individual residues in a protein therefore follows the same distribution of random errors/defects. A result that in retrospect is perhaps not too surprising since the distribution of residues may appear random, despite its serving well-defined biological purposes. The distribution results from the convolution of all the Boltzmann distributions of the 20 residues.



(a) (b)

Figure 2.5. *a) Comparison of the smoothed and unsmoothed energy distributions for all residues; b) Comparison of the cumulative energy distributions.*

Equation 2.4 differs fundamentally from the "Boltzmann probability". Boltzmann equation implies a physical basis for the energy distribution and is based on the ***accessible*** *states/deformations, g(E). On the contrary, eq. 4 is purely phenomenological and is based on the* ***accessed*** *states/deformations.*

The Kolmogorov-Smirnov test was applied to the distribution of energies of the individual residues in each protein. In order to do it, a histogram and a cumulative distribution was generated for the single macromolecules and tested against eq. 4, with λ equal to the value obtained for the whole set. Figure 6 plots the histogram for the logarithm of the probabilities.

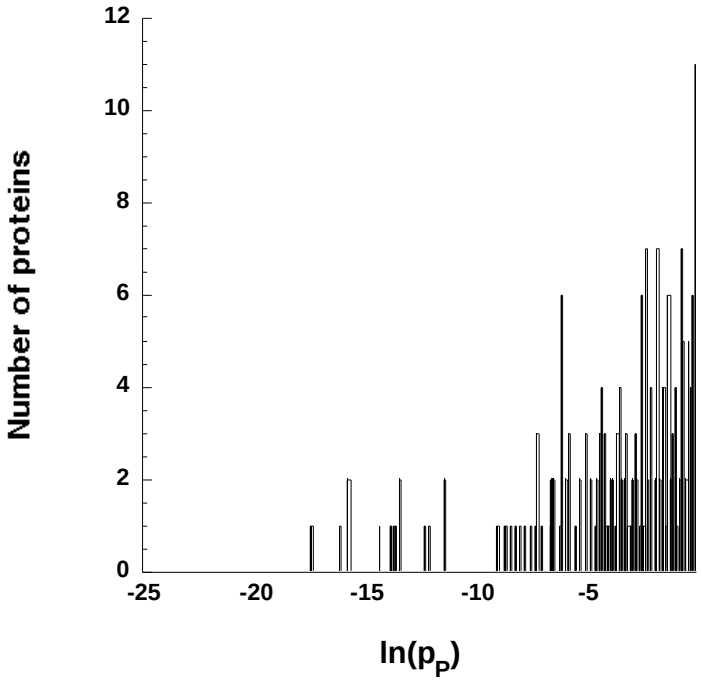


Figure 2.6. *ln(p_P) of the sample of 197 proteins (eight proteins are not shown because their value is below –25). The bin-size is 0.1. The subscript P is introduced to denote the Poisson-like origin.*

Of the proteins in the database, 51.3% have p>0.05 and 81.7% have p>0.001. More than 50% follow eq. 4. If p is set >0.05, the threshold value to consider in figure 6 is –2.99 (it grows to –6.91 if p>0.001).

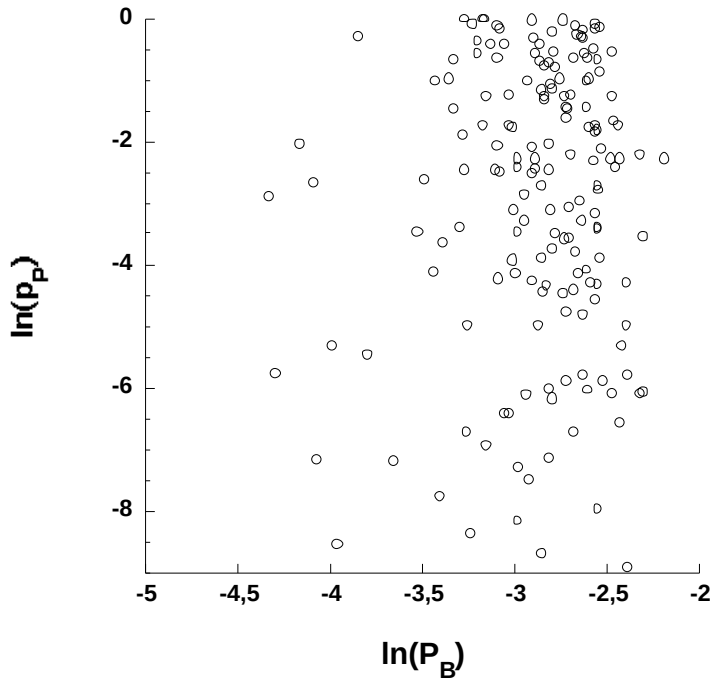


Figure 2.7. *ln(P_B) vs ln(p_P): the two descriptors do not correlate. 11 proteins are outside the plot area.*

The two descriptors, P_B and p_P, share the intra-residue potential energy excess as a common origin. However, their differing nature appears when they are plotted one against the other for all the proteins in the samples, see figure 2.7, where there is no correlation between the two descriptors that are therefore independent. This is not surprising, based, as they are, one on the distortions that can exist in the residues, and the other on the distortions that actually are present in the database.

2.4 CONCLUSION

This work starts from a concept that is becoming common in the analysis of the properties of proteins: Hydrophobicity,[10-11] various types of sidechain/side-chain interactions,[12-14] proline-isomerization,[15] hydrogen bonds,[16] internal cavities,[17] interactions at the level of specific atom types,[18,19] and the propensity of the ϕ/φ ratio[20,21] have all been found to

follow the Boltzmann distribution, or in other words to conform to the Boltzmann hypothesis. In principle, this is not entirely obvious *a priori*. The protein structures deposited in the databases are obtained for conformations at, or close to, minima of the potential energy surface. Each degree of freedom of each molecule/structure should be populated according to Boltzmann distribution. However, it is worth noticing two features:

i) the role of Boltzmann law appears readily out of the databases,

ii) several properties are independent of the others. In other words, some degrees of freedom can be separated (adiabatically) from the others so that the role of the Boltzmann distribution emerges from the analysis of their structures .

The addition of the internal degrees of freedom of the 20 types of residues to the seven other cases mentioned above contributes to further the idea that the databases should be explored not only in the search of general rules for the structural parameters, but also from the point of view of basic physical laws.

The second part of the work considers the excess energy of all the residues inside the protein. Since the energy levels of the 20 different residues are different from one another, the distribution of the energies may not be of Boltzmann type. The distribution is actually due to the sum of the distortion energies of the"i" residues (here i=41672) each belonging to one of the twenty types of residues, r. In practice

$$\sum_{i(r)}^{41672} E^{\alpha_{i(r)}} \exp\left(-\frac{E_{i(r)}}{k_B T}\right) = \lambda E \exp(-\lambda E) \qquad (2.5)$$

the resulting function is empirical, but not unexpected since both sides of equation 5 contain exponential functions and it seems reasonable that with the proper values of the parameters the equality holds. In practice, the sum on the left can be considered a power expansion of the right-hand side of the equation. Perhaps more unexpected is that individual proteins, or at least more than 50% of the high-resolution structures of 197 proteins selected here, conform to the empirical distribution of energies. One can envisage some interesting follow-ups to this finding. The first could be the application of a test based on the validity of equation 4 to the structures that are obtained from X-ray diffraction experiments, a test that may expedite the refinement. Alternatively, a similar test could be used to assist the prediction of the folded structures of proteins out of the many that can be generated computationally.

List of symbols

There are four symbols related to the letter "p/P" with the meaning reported below:

p is the level of significance that a hypothesis is statistically valid; here, it is used in

conjunction with the Kolmogorov-Smirnov test to ascertain if a set of energies of distortion is described by a function whose parameters are varied to obtain the highest statistical significance.

$P_r(E_r)$ it is the probability that a residue is distorted and has energy Er; this probability is Boltzmann multiplied the density of states; each type of residue, r, has a different density of states; the function for the density of states is determined using the Kolmogorov-Smirnov test on the distribution of distortion energies of the database of high resolution structures.

P_B it is product of the $P_r(E_r)$ of the residues along a protein

p_P is the level of significance obtained by the Kolmogorov-Smirnov test that the distortion energy of the residues of a single protein follows Poisson distribution of eq. 4 with λ=0.265.

2.5 REFERENCES

[1] Pohl, F.M. Empirical protein energy maps, *Nat. New Biol.* **1971**, *234*, 277–279.

[2] Sippl, M.J. Knowledge-based potentials for proteins, *Curr. Opin. Struct. Biol.* **1995**, *5*, 229–235.

[3] Vajda, S.; Sippl, M.; Novotny, J. Empirical potentials and functions for protein folding and binding, *Curr. Opin. Struct. Biol.* **1997**, *7*, 222–228.

[4] Rojnuckarin, A.; Subramaniam, S.; Knowledge-based interaction potentials for proteins, *Proteins*, *Struct. Funct. Gen.* **1999**, *36*, 54-67.

[5] Zhang, L.; Skolnick, J. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci.*, **1998**, *7*, 112-122.

[6] Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comp.-Aid. Mol. Des.*, **1993**, *7*, 473-501.

[7] Kuszewski, J.; Gronenborn, A.M.; Clore, G.M. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci*., **1996**, *5*, 1067–1080.

[8] Moult, J. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.*, **1997**, *7*, 194–199.

[9] Finkelstein A.V.; Badretdinov, A.Y.; Gutin, A.M. Why do protein architectures have Boltzmann-like statistics? *Proteins*, **1995**, *23*, 142–150.

[10] Rose, G.D.; Geselowitz, A.R.; Lesser, G.J.; Lee, R.H.; Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science*, **1985**, *229*, 834–838.

[11] Casari, G.; Sippl, M.J. Structure-derived hydrophobic potential, A hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, **1992**, *224*, 725–732.

[12] Miyazawa, S.; Jernigan, R.L. Estimation of effective inter-residue contact energies from crystal structures, Quasi-chemical approximation. *Macromol.*, **1985**, *18*, 534–552.

[13] Kannan, N.; Vishveshwara, S. Aromatic clusters, A determinant of thermal stability of thermophilic proteins. *Protein Eng*., **2000**, *13*, 753–761.

[14] Butterfoss, G.L.; Hermans, J. Boltzmann-type distribution of side-chain conformation in proteins. *Protein Sci.*, **2003**, *12*, 2719-2731.

[15] MacArthur, M.W.; Thornton, J.M. Influence of proline residues on protein conformation. *J. Mol. Biol.*, **1991**, *218*, 397–412.

[16] Sippl, M.J.; Ortner, M.; Jaritz, M.; Lackner, P.; Flockner, H. Helmholtz free energies of atom pair interactions in proteins. *Fold. Des*., **1996**, *1*, 289–298.

[17] Rashin, A.A.; Rashin, B.H.; Rashin, A.; Abagyan, R. Evaluating the energetics of empty cavities and internal mutations in proteins. *Protein Sci*., **1997**, *6*, 2143–2158.

[18] Samudrala, R.; Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol*., **1998**, *275*, 895–916.

[19] Lu, H., Skolnick, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **2001**, *44*, 223–232.

[20] Shortle, D. Composites of local structural propensities, Evidence for local encoding of long range structure. *Protein Sci.*, **2002**, *11*, 18–26.

[21] Shortle, D. Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci*., **2003**, *12*, 1298-1302

[22] Dudek, M.J.; Ponder, J.W. Accurate modeling of the intramolecular electrostatic energy of proteins. *J. Comp. Chem*., **1995**, *16*, 791-816,

[23] Kundrot, C.E.; Ponder, J.W.; Richards, F.M. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. *J. Comp. Chem.*, **1991**, *12*, 402-409.

[24] Ponder, J.W.; Richards, F.M. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comp. Chem*., **1987**, *8*, 1016.

[25] Biscarini F, Cavallini M, Leigh DA, León S, Teat SJ, Wong JKW, Zerbetto F. The Effect of Mechanical Interlocking on Crystal Packing, Predictions and Testing. *J. Am. Chem. Soc.*, **2002**, *124*, 225-233.

[26] León, S.; Leigh, D.A.; Zerbetto, F. The effect of guest inclusion on the crystal packing

of p-tert-butylcalix[4]arenas. *Chem. Eur. J.*, **2002**, *8*, 4854-4866.

[27] Höfinger, S.; Zerbetto, F. On the cavitation energy of water. *Chem. Eur. J.*, **2003**, *9*, 566-569.

[28] Teobaldi, G.; Zerbetto, F. Molecular dynamics and implications for the photophysics of a dendrimer-dye guest-host systems. *J. Am. Chem. Soc*., **2003**, *125*, 7388-7393.

[29] Weiner, S,J,; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S. jr; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc*., **1984**, *106*, 765-784.

[30] Maxwell, D.S.; Tirado-Rives, J.; Jorgensen, W.L. A comprehensive study of the rotational energy profiles of organic systems by ab initio MO theory, forming a basis for peptide torsional parameters. *J. Comp. Chem*., **1995**, *16*, 984-1010.

[31] Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. Numerical Recipes in Fortran. **1992**, Cambridge University Press.

[32] Blank, P.S.; Sjomeling, C.M.; Backlund, P.S.; Yergey, A.L. Use of cumulative distribution functions to characterize mass spectra of intact proteins. *J. Am. Soc. Mass. Spectry*, **2002**, *13*, 40-46.

[33] Brancato, G.; Zerbetto, F. On the Distribution of Local Molecular Symmetry in Crystals. *J. Phys. Chem. A,* **2000**, *104*, 11439-11442.

[34] Grzybowski, B.A.; Ishchenko, A.V.; DeWitte, R.S.; Whitesides, G.M.; Shakhnovich, E.I. Development of a Knowledge-based potential for crystals of small organic molecules: calculation of energy surfaces for C=O···H-N hydrogen bonds, *J. Phys. Chem. B*, **2000**, *104***,** 7293-7298.

[35] L.D. Landau, E.M. Lifshitz, Statistical Physics, Butterworth Heinemann, 1980.

[36] Stein, S.E.; Rabinovitch, B.S. Accurate evaluation of internal energy level sums and densities including anharmonic oscillators and hindered rotors. *J. Chem. Phys*., **1973**, *58*, 2438-2445.

[37] Zerbetto, F. Carbon Rings Snapping. *J. Am. Chem. Soc*., **1999**, *121*, 10958-10961.

[38] Heine, T.; Zerbetto, F. Dynamics of carbon clusters, Chemical equilibration of rings and bicyclic rings. *Chem. Phys. Letters*, **2002**, *358*, 359-367.

CHAPTER 3

AROMATIC STABILIZATION OF PROTEINS

3. 1 INTRODUCTION

The origin of the additional stability of thermophilic proteins has received considerable attention both experimentally and theoretically.[1-7] The picture that has emerged is that different protein families adapt to high operating temperature using different structural tools, and that proteins from extreme and moderate thermophiles are stabilized by different mechanisms.[1] Perhaps, the rule observed most consistently in the structures of thermophilic proteins is an increase in the number of ion pairs with increasing growth temperature. Other parameters tend to show only qualitative trends.[1] It has been suggested that the presence of extra salt bridges (and hydrogen bonds) results in a lower heat capacity of unfolding than in mesophilic proteins. Higher folding stability and lower heat capacity can both be modeled by a simple approach.[2]

In addition to Coulomb interactions, adaptation to high temperatures involves a number of subtle co-operative effects, often specific to a given protein family. These include (i) minimization of surface energy, (ii) hydration of non-polar surface groups, (iii) burying of hydrophobic residues, (iv) optimization of core packing, (v) hydrogen bonds, and (vi) optimization of weak protein-protein and protein-solvent interactions. This complex picture is further complicated by the fact that high melting temperature is not always synonymous with greater thermodynamical stability.[3]

One contribution that has come under scrutiny as a source of additional stability of thermophilic proteins is the aromatic electrostatic interaction, leading to so-called aromatic clusters. A graph spectral method was used[4] to identify aromatic clusters for a dataset of 24 protein families for which the crystal structures of thermophilic and mesophilic homologues were available. For 17 different thermophilic protein families, the analysis showed the presence of additional aromatic clusters, or enlarged aromatic networks, absent in the corresponding mesophiles. These clusters were often located close to the active site of the thermophilic enzyme. A geometrical analysis of the packing geometry of the pairwise aromatic interaction showed a preference for T-shaped orthogonal packing.[4] However, a local increase of the energetic stability via improved packing does not unequivocally favour a given mutation, because it may imply concomitant limitations on motion. Rigid structures imply higher vibrational frequencies, which, in turn, imply smaller entropy and militate against decrease of the free energy. In the simplest (harmonic) approximation, entropy is

associated with mobility, and the qualitative expectation is that a strongly stabilizing geometrical motif gives a lower entropy and hence a poorer free energy.

We therefore decided to investigate the *entropic* contribution of the mutated aromatic fragments in thermophilic proteins using the dataset of proteins identified by Kannan and Vishveshwara.[4]

3.2 BACKGROUND

All Molecular Dynamics calculations were carried out with the TINKER program,[8-10] which has found a number of applications in our laboratory,[11-15] using the AMBER/OPLS/UA force field.[16,17] Only the clusters of the mutated residues in the thermophilic and the mesophilic proteins were allowed to undergo dynamics, subject to the interaction with the rest of the protein. For each cluster, 420 ps of dynamics were run, with the initial 20 ps sufficient for equilibration.

In order to calculate the entropy, a computer program was written based on the approach of Schäfer, Mark and van Gunsteren, based on the equation[18]

$$S=\frac{1}{2}k_B \ln\left|1+\frac{k_B T e^2}{h^2}M^{\frac{1}{2}}sM^{\frac{1}{2}}\right| \qquad (3.1)$$

where e is the base of natural logarithms, $\boldsymbol{M}$ is the diagonal matrix of atomic masses, and $\boldsymbol{\sigma}$ is the covariance matrix of the atomic position fluctuations

$$s_{ij}=\left\langle \left(x_i-\langle x_i\rangle\right)\left(x_j-\langle x_j\rangle\right)\right\rangle \qquad (3.2)$$

the other symbols have their usual meanings. The larger the mobility of a cluster of atoms, the greater is the entropy calculated from eq. (1 and, in practice, if the position of any particular atom fluctuates greatly, its entropic contribution is large.

To evaluate the vibrational motions of a cluster as a single unit, every picosecond, we calculate

$$O=\left(\sum_i^{atoms}\left(x_i(t+Dt)-x_i(t)\right)^2+\sum_i^{atoms}\left(y_i(t+Dt)-y_i(t)\right)^2+\sum_i^{atoms}\left(z_i(t+Dt)-z_i(t)\right)^2\right)/N \qquad (3.3)$$

where N is the number of atoms in the cluster. $\boldsymbol{O}$ is a measure of the overall motion of the cluster. Its Fourier Transform gives the frequency of the motion. Three frequency ranges were explored 0-10 cm^{-1}, 0-30 cm^{-1}, and 0-50 cm^{-1}. After the Fourier transform, we take the integral of the vibrational amplitudes, I, in absolute value over this range of frequency. The result is conveniently expressed in ppm.

3.3 RESULTS AND DISCUSSION

Full calculation of the entropy for a large set of proteins is daunting as positional fluctuations may

not converge within the time of a molecular dynamics (MD) run. A simplified, reduced approach is in order. It was decided to investigate the vibrational freedom of aromatic clusters in thermophilic protein and compare it with the motion of the equivalent set of residues in the mesophile. In this way, an entropy can be assigned to the cluster. In the MD calculations, every residue of the clusters undergoes dynamics, subject to interaction with the remainder of the protein, which is held frozen. Advantages and disadvantages of calculating entropy as the sum of the contributions of individual residues are critically discussed in ref. [19]. Qualitatively, since entropy is an extensive property, freezing the main body of the protein amounts to neglect of the (second order) effect of fragment motion on that of a much larger object. It is reasonable to expect for the thermophilic and the mesophilic fragments embedded in the protein a similar accuracy. Relative values, or entropy ordering, should be predicted accurately, while the absolute values could be inaccurate.

Overall, 16 pairs of proteins and 36 clusters were investigated, see Table 3.1. The systems are taken from Table II of ref. [4].

Protein	Cluster	Thermophile residues	Mesophile residues	Protein	Cluster	Thermophile residues	Mesophile residues
1. Neutral protease (1THL /1NPC)	1	TYR93	ILE94	9. Reductase (1EBD /1LVL)	1	PHE209	TYR210
		TYR151	ASN152			PHE358	TYR355
		TRP115	TRP116		2	TYR321	ALA318
	2	TYR28	TYR29			TYR339	PRO336
		TYR24	LEU24		3	TYR59	ARG81
2. Lactate dehydrogenase (1LDN /1LDM)	1	PHE156	PHE170			PHE194	LEU193
		TRP 187	TRP201			PHE79	ILE84
		PHE216	HIS228			TYR189	TYR190
	2	TYR266	TYR278		4	PHE115	VAL120
		PHE300	LEU313			PHE134	CYS135
	3	PHE115	PRO129	10. Triose phosphate isomerase (1BTM /1TIM)	1	TYR165	TYR163
		PHE119	ILE133			TYR209	TYR207
	4	PHE103	PHE117			PHE221	LEU219
		PHE136	LEU150		2	PHE67	LYS67
		TYR131	TYR145			TYR73	PHE73
		TRP134	TRP148		3	TRP9	TRP11
		TYR261	MET273			PHE21	LEU23
		TYR272	PHE285			PHE242	PHE239

		PHE315	PHE329	11. Xylanase (1YNA /1XYN)	1	TYR13	TYR5
	5	PHE51	LEU65			TRP9	*
		PHE23	ILE37			TYR170	ASN157
	6	PHE16	ALA30			TYR171	TYR158
		TYR234	SER246			TYR72	LEU62
3. Phosphofructo kinase (3PFK /2PFK)	1	PHE230	LEU231			TYR87	TYR77
		TYR196	PHE197			PHE92	ASN82
	2	TYR38	TYR39			TYR76	TYR66
		TYR69	GLY70			TRP78	TRP68
4. Ribonuclease H (1RIL /2RN2)	1	PHE7	PHE8			PHE133	PHE121
		TYR67	SER68			TRP137	ILE125
	2	TYR72	TYR73		2	TYR26	TYR17
		TRP104	TRP104			TYR14	ASP6
		PHE77	ILE78			TRP16	ASN8
		PHE118	TRP118			TYR34	PHE24
		PHE120	TRP120	12. Glycosyltransferase (1XYZ /2EXO)	1	PHE205	PHE202
		TRP81	TRP81			TYR228	PHE222
		TRP85	TRP85			PHE237	ILE231
		TRP90	TR90			PHE187	TYR184

5. Malate dehydrogenase (1BMD /4MDH)	1	TRP184	TRP184			PHE277	VAL270
		PHE192	TYR192		2	TRP288	TRP281
		TYR272	TYR278			TRP280	TRP273
		TYR280	TYR286			PHE293	PHE286
		PHE282	PHE288			TYR296	GLU289
		PHE302	PHE308	13. Triacylglycerol acylhydrolase (1TIB /1LGY)	1	PHE51	ILE19
		TRP213	TRP218			PHE66	TYR62
		PHE218	PHE223			TYR16	PHE169
		PHE196	ASN196			PHE13	GLN193
		TYR214	LEU219			PHE169	PHE13
	2	TYR18	TYR18			TYR194	PHE257
		PHE22	TYR22			PHE10	ILE10
	3	PHE62	LEU62			PHE262	TYR256
		TYR141	SER141			PHE7	VAL171
		PHE152	PHE152			TYR261	ILE48
6. Hydrolase (2PRD /1INO)	1	PHE57	TYR57			TYR171	TYR16
		TYR32	ILE32	14. Pyrophosphatase (2PRD /1OBW)	1	PHE57	TYR57
7. Phospho glycerate kinase (1PHP /3PGK)	1	TYR303	PHE322			TYR32	ILE32

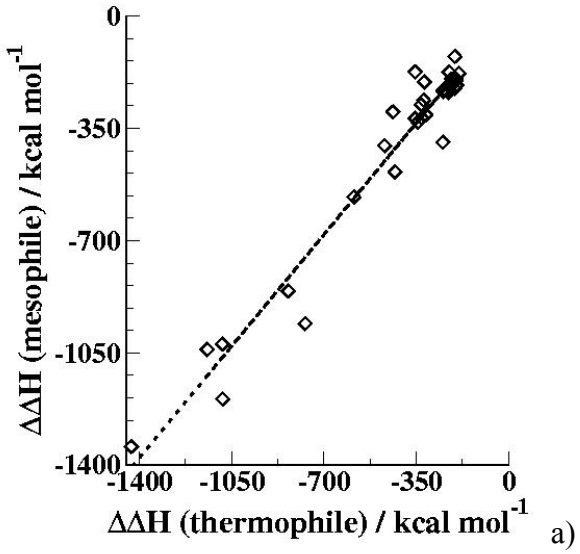
		TYR261	VAL279	15. Carboxypeptidase (1OBR /2CTC)	1	PHE272	TYR265
	2	PHE225	PHE240			PHE266	TYR259
		PHE249	LEU267			PHE274	PHE267
		PHE260	VAL278			PHE233	LEU219
8. Subtilisin (1THM /1ST3)	1	TYR174	ARG164			PHE230	LYS216
		TYR171	TYR161			TYR212	TYR204
		TYR175	TYR163			TYR216	TYR208
		TRP199	GLY189			TYR214	TYR206
	2	TYR196	TYR186			TYR151	ALA141
		TYR265	LEU256			TRP264	TRP257
	3	TYR210	GLN200			PHE174	ASN171
		TYR7	SER3			TYR149	TRP147
		TYR218	TYR208	16. Ornitine carboxypeptidase (1AIS /2OTC)	1	PHE21	GLU38
						TRP168	LEU197

Table 3.1. *Pairs of thermophilic and mesophilic proteins together with their pdb codes (the first is for the thermophile, the second the mesophile) and their clusters investigated in this work.*

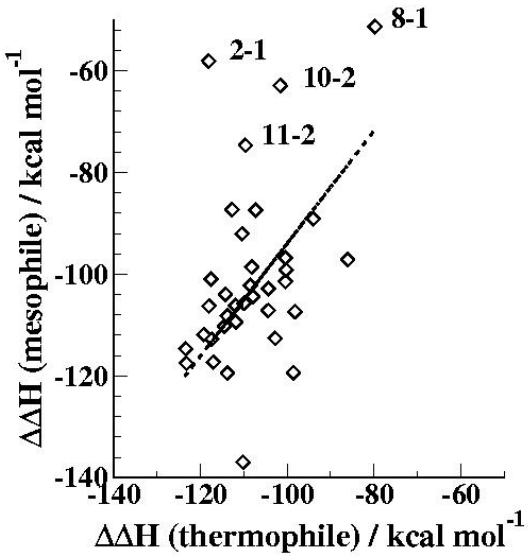
Free-energy comparison involves both enthalpic and entropic factors. In order to compare proteins with and without aromatic clusters it is necessary to define differential properties. We first examine ΔΔH, which are given by the average, over each Molecular Dynamics run, of the energy of interaction between each cluster and the rest of the protein. The term ΔΔH is preferred over ΔH since the latter entails the contribution from the formation of the covalent bonds, which is not considered here (or is effectively subtracted).

Figure 3.1 compares these differential enthalpies for the fragments in the mesophilic and thermophilic proteins in the data set. The range of values covered by the ΔΔH's is substantial. However, once each value is divided by the number of residues in the cluster, the average is 106 kcal mol^{-1}. This value is similar to that of a single CC carbon bond, but is due to, and includes, all

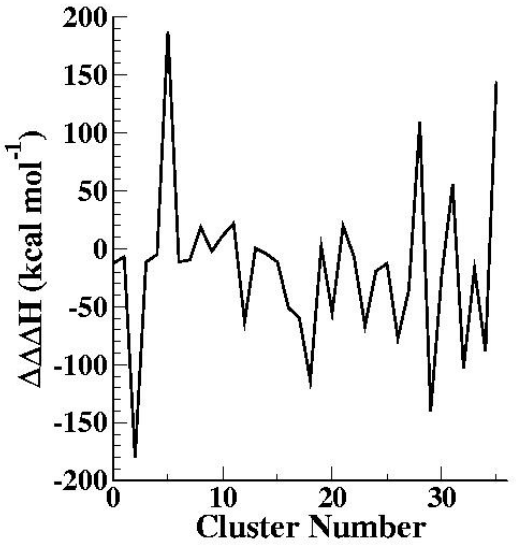
the van der Waals and Coulomb interaction between a single residue in the cluster and all the other residues in the protein. Figure 3.1 demonstrates a linear correlation between the two sets, with correlation factor r=0.98. Some of the correlation is due to the (trivial) correlation in the number of residues of the mesophilic clusters and their thermophilic counterparts. However, a much weaker correlation, r=0.47, is found when ΔΔH is divided by the number of residues in the cluster, see figure 3.1b. As mutations typically involve only a few residues out of the many in the protein, such a linear relationship is unexpected. Significantly, however, the quantities ΔΔH give no additional stabilization of thermophiles over mesophiles, beyond that of the inherent internal energy of the local aromatic interaction. Figure 3.1c and 3.1d show that for the majority of clusters there is no net enthalpic advantage due to the thermophilic mutations. Indeed, 12 clusters are stabilized, 6 are destabilized, and half of them are neither stabilized nor destabilized.



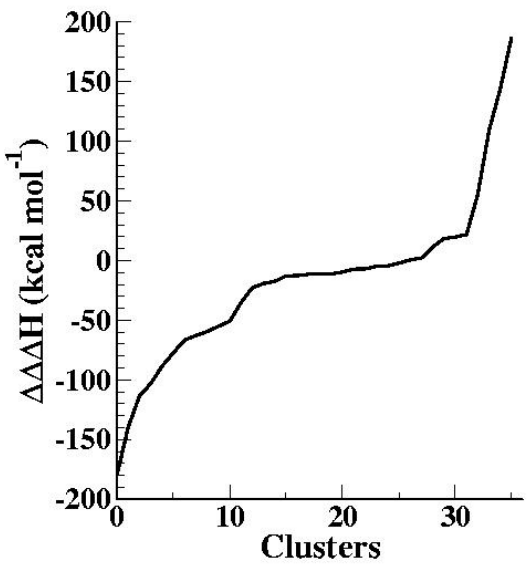
a)



b)



c)



d)

Figure 3.1. *Comparison of differential enthalpies of interaction for fragments in thermophilic and mesophilic proteins. ΔΔH accounts for the interaction of a given fragment with the remainder of the protein, after subtraction of the internal energy at 0 K, i.e. the stabilization inherent in formation of the cluster. ΔΔΔH is the enthalpy difference between thermophiles and mesophiles. Each datapoint refers to one cluster in a thermophile/mesophile pair. (a) is the energy is per cluster, the best-fit line (dashed) corresponds to ΔΔH(mesophile)= 23.71 kcal mol^{-1} +1.04 ΔΔH(thermophile), with r=0.98; b) the energy per cluster has been divided by the number of residues in the cluster, r=0.47; c) ΔΔΔH of the enthalpies of each cluster with the residues in the same order of the original database as in Table 3.1; d) ΔΔΔH ordered increasingly.*

During the Molecular Dynamics runs, entropy builds up until convergence is reached. Figure 3.2 shows two examples of the convergence of entropy (the first and the last cluster of the set). Similar plots for all the other clusters have been calculated.

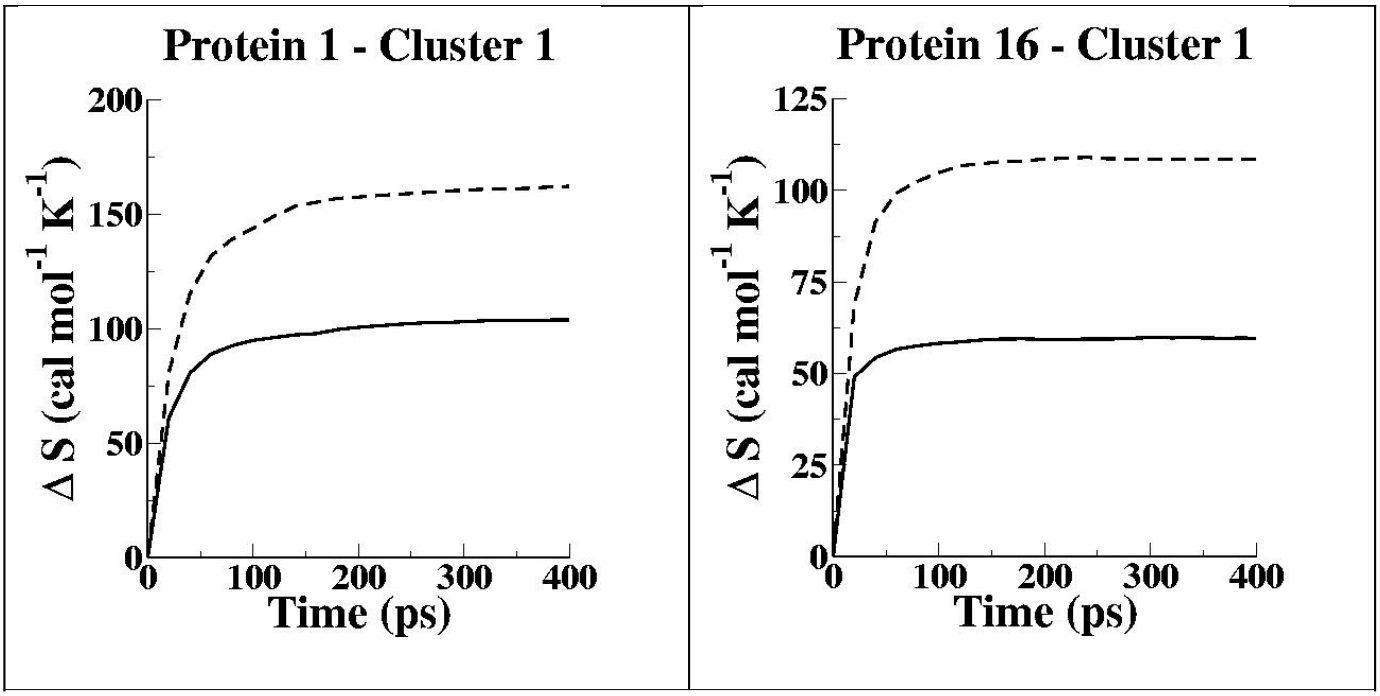


Figure 3.2. *Convergence in time of the entropy for the first and last cluster of the 36 investigated. The solid line is for the mesophiles, the dashed line is for the thermophiles.*

Figure 3.3a and 3.3b compare the entropic stabilization, TΔS, with the enthalpic stabilization energy for thermophiles and mesophiles. Within each group there is a good linear correlation between the two quantities, although the two slopes differ. At 298 K, that is the temperature used in the plots, the enthalpic factor substantially exceeds the entropic one. The correlation between entropy and enthalpy in the two sets of clusters suggests that two components of free energy are governed by the same factors.

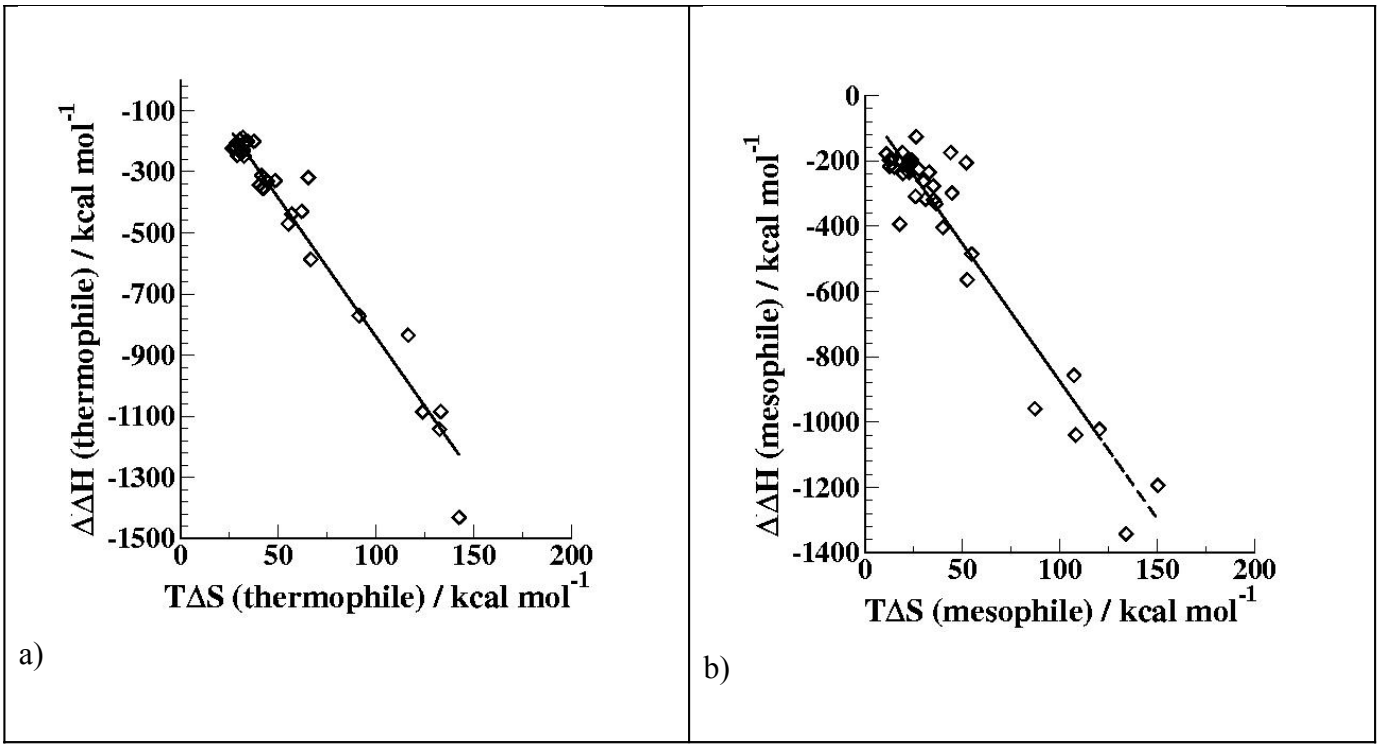


Figure 3.3. *Comparison of the entropic, TΔS, and enthalpic stabilization: a) thermophiles, b) mesophiles. T was set to 298 K. The best-fit line corresponds to ΔH(thermophile)= -9.08 TΔS + 67.41, with r=0.96; and H(mesophile)= -8.41 TΔS –33.32, with r=0.91.*

Figure 3.4 compares the *entropies*, S, of aromatic clusters and the equivalent fragments in the thermophilic and mesophilic proteins. There is another linear correlation between the two sets, S(mesophile)=-40.42 cal mol^{-1} K^{-1} + 1.047 S(thermophile), with a correlation factor r=0.98. Once again, correlation of motional entropy of these residues within an unchanged bulk protein is not expected *a priori*.

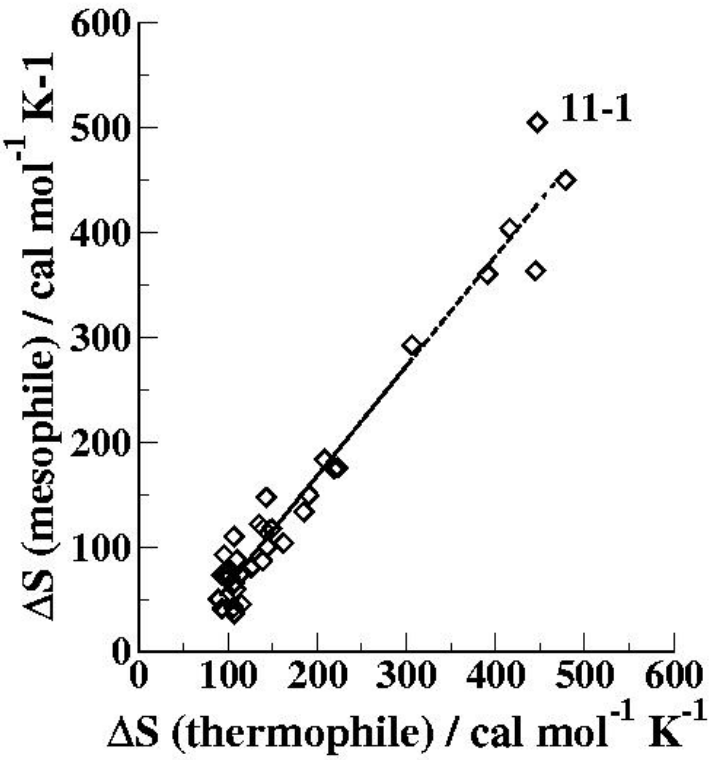


Figure 3.4. *Comparison of the entropic contributions of the aromatic clusters and equivalent fragments of thermophilic and mesophilic proteins. The best-fit line (dashed) corresponds to* $S(mesophile)=-41.42\pm7.82$ *cal* mol^{-1} K^{-1} $+ 1.047\pm0.037\ S(thermophile)$, *with* $r=0.98$.

The fit indicates a systematic entropic advantage introduced by the "aromatic" mutations. Out of the 36 clusters, only three have greater entropy in the mesophilic proteins. This is better appreciated in figures 3.5a and 3.5b where the substantial entropic advantage of the thermophilic mutations is readily perceived.

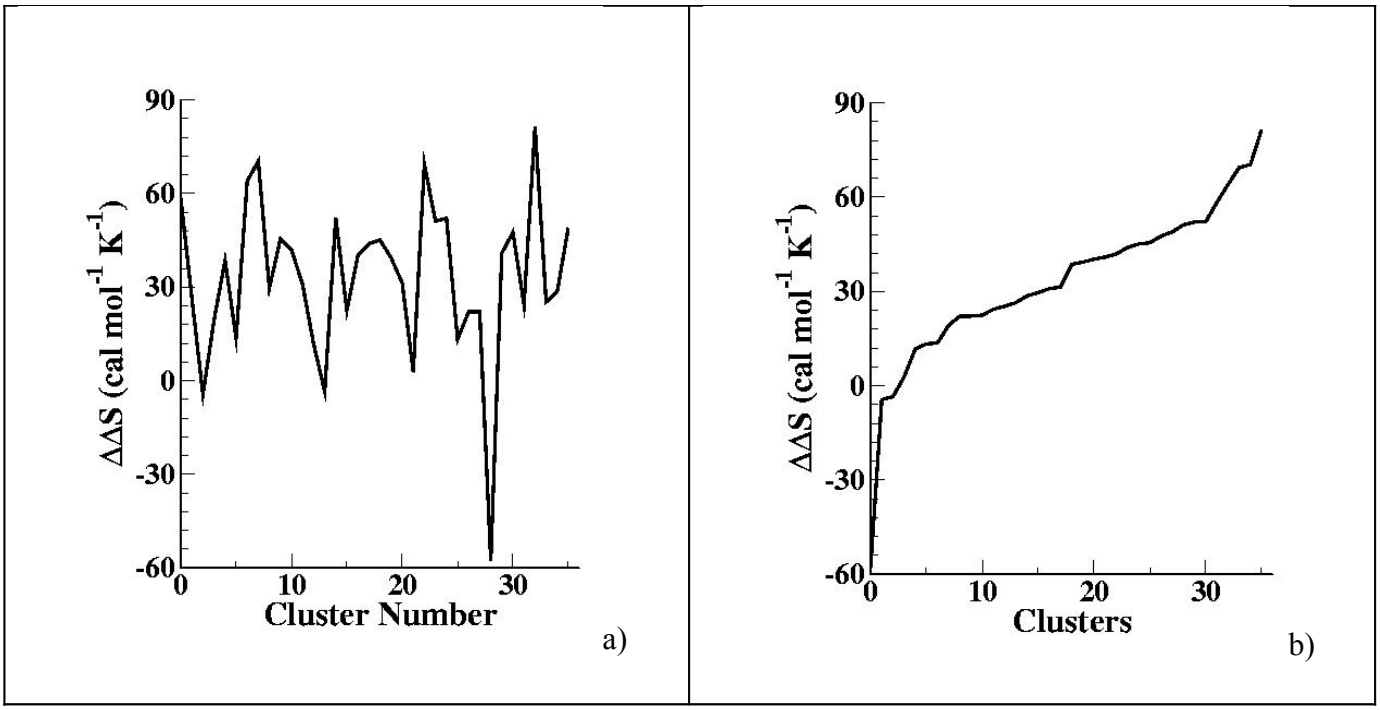


Figure 3.5. *Comparison of differential entropies of the fragments in the clusters of thermophilic and mesophilic proteins: a) ΔΔS for each cluster with the residues in the same order of the original database as in Table 3.1; db) ΔΔS ordered increasingly.*

The correlation evidently has a positive slope, slightly in excess of unity, and an intercept indicating a negative entropy of ~20 R (where R is the universal gas constant). These two features of the correlation shed light on two different aspects of the cluster motion. First, the slope is qualitatively consistent with the notion that aromatic clusters are locally more rigid: their internal motions have higher frequencies and contribute less to the entropy than motion of the corresponding mesophile fragments. Thus, if the intercept of the linear fit were not non-zero, the presence of aromatic clusters would be an entropic disadvantage, with mesophilic proteins having greater entropy than the thermophiles. However, the intercept is large and negative. Several simple approaches were tested with the intent of explaining the entropic advantage. Correlation of the entropic values with the largest root mean square deviation from the equilibrated structure did not show any systematic trends. Nor did a similar correlation with the number of conformers, n, (actually with log n) detected during the Molecular Dynamics run. The best explanation we were able to find is based on the hypothesis that the advantage arises from a systematic difference in low-frequency, high-amplitude motions undergone by the cluster units. In the aromatic case, several residues are tightly coupled together, and will move together; in the mesophiles the equivalent residues are less strongly interacting and can be expected to move more independently. Such motions of the relatively rigid aromatic-cluster subunits are expected to be highly anharmonic and to lead ultimately to a higher

entropy. This hypothesis can be tested: integrated amplitudes, I, of low-frequency vibrations of clusters in thermophiles and the corresponding fragments in mesophiles were computed and are compared in Figure 3.6.

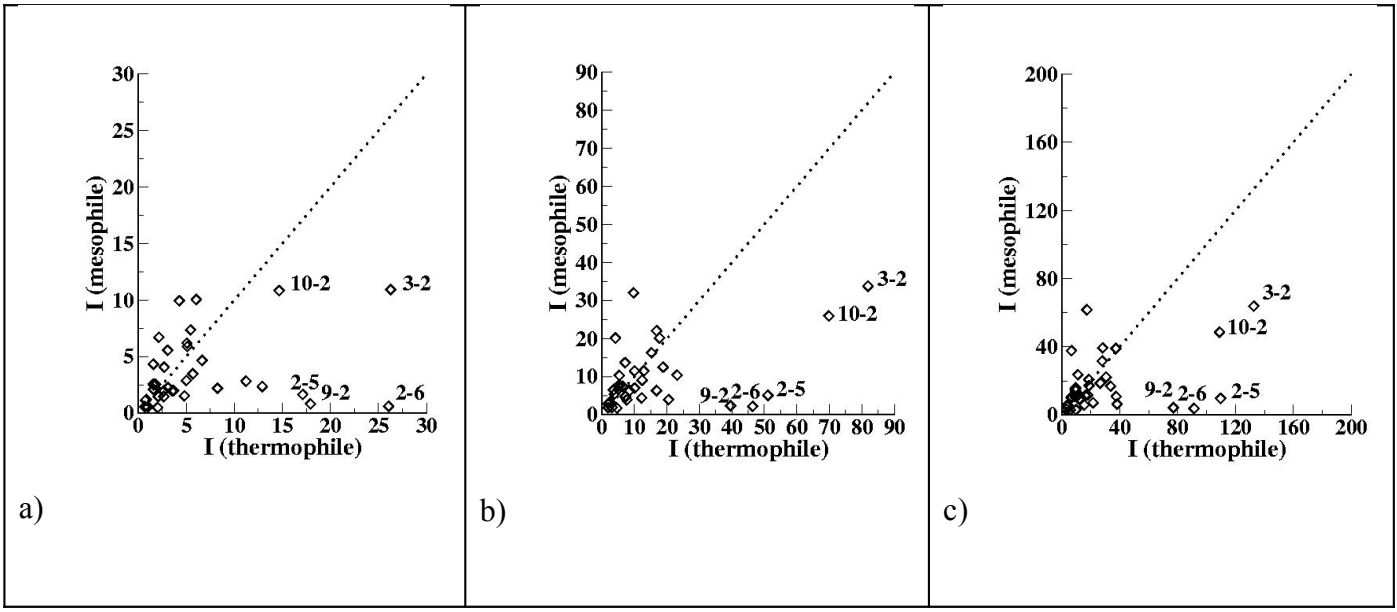


Figure 3.6. *Comparison of integrated amplitudes of whole-cluster motions in thermophile and mesophile pairs. The dashed line has unit slope and is used to indicate the divide between cases where the amplitude is larger in the thermophile (the majority) from those where it is larger in the mesophile: a) the cutoff has been set to 10 cm^{-1}; b) the cutoff has been set to 30 cm^{-1}; c) the cutoff has been set to 50 cm^{-1}.*

The figure shows that most thermophile aromatic clusters have significantly greater integrated amplitude in their low frequency region than do the corresponding fragments in mesophiles, a result that is not sensitive to the cutoff. This then, is a source of entropic advantage.

3.4 CONCLUSION

In conclusion, aromatic fragments in thermophilic proteins tend to generate larger entropy via their overall low-frequency motion. This feature indicates one direction for exploration in connection with rational design of ultrastable proteins. Finally, it may be noted that the aromatic residues present in the 36 clusters are tryptophan, tyrosine, and phenylalanine. The latter two (together with cysteine) have substantially increased their frequency of occurrence with respect to ancient proteins over the last three billion years.[20] The present work suggests that one advantage of their presence is greater stability arising from their entropic contribution.

3.5 REFERENCES

[1] Szilagyi, A.; Zavodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey, *Struct. (London)* **2000**, 8, 493-504.

[2] Zhou, H.-X. Toward the physical basis of thermophilic proteins: linking of enriched polar interactions and reduced heat capacity of unfolding, *Biophys. J.* **2002**, *83,* 3126-3133.

[3] Ladenstein, R.; Antranikian, G., Proteins from hyperthermophiles: stability and enzymic catalysis close to the boiling point of water, *Advances in Biochemical Engineering/Biotechnology 61(Biotechnology of Extremophiles)* **1998**, 37-85.

[4] Kannan, N.; Vishveshwara, S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins, *Prot. Engin.* **2000**, *13*, 753-761.

[5] Dalhus, B.; Saarinen, M.; Sauer, U. H.; Eklund, P.; Johansson, K.; Karlsson, A.; Ramaswamy, S.; Bjork, A.; Synstad, B.; Naterstad, K.; Sirevag, R.; Eklund, H, Structural basis for thermophilic protein stability: structures of thermophilic and mesophilic malate dehydrogenases, *J. Mol. Biol.* **2002**, *318*, 707-721.

[6] Cowan, D. A. Thermophilic proteins: stability and function in aqueous and organic solvents. *Comparative Biochemistry and Physiology, Part A: Molecular & Integrative Physiology* **1997**, *118A*, 429-438.

[7] Jaenicke, R.; Bohm, G. 1998, The stability of proteins in extreme environments. *Current Opinion in Structural Biology,* **1998**, *8*, 738-748.

[8] Dudek, M.J.; Ponder, J.W. Accurate modeling of the intramolecular electrostatic energy of proteins. *J. Comp. Chem* **1995,** *16*, 791-816.

[9] Kundrot, C.E.; Ponder, J.W.; Richards, F.M. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. *J. Comp. Chem.* **1991,** *12*, 402-409.

[10] Ponder, J.W.; Richards, F.M. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comp. Chem*. **1987,** *8*, 1016.

[11] B. Trebbi, M. Fanti, I. Rossi, F. Zerbetto, The intra-residue distribution of energy in proteins, *J. Phys. Chem. B*, **2005**, *109*, 3586-3593.

[12] Biscarini F, Cavallini M, Leigh DA, León S, Teat SJ, Wong JKW, Zerbetto F. The Effect of Mechanical Interlocking on Crystal Packing, Predictions and Testing, *J. Am. Chem. Soc.* **2002,** *124*, 225-233.

[13] León, S.; Leigh, D.A.; Zerbetto, F. The effect of guest inclusion on the crystal packing of p-

tert-butylcalix[4]arenas, *Chem. Eur. J.* **2002**, *8*, 4854-4866.

[14] Höfinger, S.; Zerbetto, F. On the cavitation energy of water, *Chem. Eur. J*. **2003**, *9*, 566-569.

[15] Teobaldi, G.; Zerbetto, F. Molecular dynamics and implications for the photophysics of a dendrimer-dye guest-host systems, *J. Am. Chem. Soc*. **2003**, 125, 7388-7393.

[16] Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S. Jr; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins, *J. Am. Chem. Soc*. **1984**, 106, 765-784.

[17] Maxwell, D.S.; Tirado-Rives, J.; Jorgensen, W.L. A comprehensive study of the rotational energy profiles of organic systems by ab initio MO theory, forming a basis for peptide torsional parameters, *J. Comp. Chem.* **1985**, *16*, 984-1010.

[18] Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix, Chem. Phys. Lett. 1993, 215, 617-621; H. Schäfer, A. E. Mark, W. F. van Gunsteren, Absolute entropies from molecular dynamics simulation trajectories, *J. Chem. Phys.* **2000**, *113*, 7809-7817.

[19] Schäfer, H.; Daura, X.; Mark, A.E.; van Gunsteren, W.F. Entropy Calculations on a Reversibly Folding Peptide: Changes in Solute Free Energy Cannot Explain Folding Behavior, *Proteins: Struct., Funct., Gen.* **2001,** *43*, 45–56.

[20] Brooks, D.J.; Fresco, J.R. Increased Frequency of Cysteine, Tyrosine, and Phenylalanine Residues Since the Last Universal Ancestor, *Molec. & Cell. Proteomics* **2002,** *1*, 125–131.

CHAPTER 4
CONFIGURATIONAL TEMPERATURE

4.1 INTRODUCTION

Temperature, the macroscopic expression of kinetic energy, reflects the dynamics of molecular ensembles. Because of the energy flow between degrees of freedom, temperature must be obtainable in terms of the geometrical deformations accessed by the particles. Indeed, in recent years models describing a configurational temperature, T_{conf}, have appeared.[1-5] They offer the possibility of evaluating temperature from potential energy derivatives at the (instantaneous) structure of the molecular system. While a variety of applications have been discussed,[1-4] only a single experimental application has been presented so far.[5]

4.2 BACKGROUND

The configurational temperature is given by the first and the second derivatives of the potential energy

$$k_B T_{conf} = \frac{\sum_j \langle |\nabla_j U|^2 \rangle}{\sum_j \langle \nabla_j^2 U \rangle} \qquad (4.1)$$

where k_B is Boltzmann's constant, $\nabla_j U$ the gradient, $\nabla_j^2 U$ the Laplacian of the potential energy U, with respect to the position of the j-th particle. The summation may be over all particles in the system or restricted to a single species, or even to an individual particle. This equation is also known as a hypervirial relation.[1,2] It was used by Rugh[3] as an independent estimate of temperature in simulations and recommended as a diagnostic test for lack of equilibrium by Butler et al.[4]

4.3 RESULTS AND DISCUSSION

Equation 4.1 is here used initially to assess its accuracy in a practical implementation by performing molecular dynamics simulations at constant kinetic temperature on Crambin, a small protein with

46 residues and 327 heavy atoms. Figure 1 compares at 50, 100 and 300 K configurational and kinetic temperatures, T_{kin}, in a molecular dynamics run of 100 ps. The kinetic temperature was maintained constant by coupling to a bath.[6] This algorithm is a common feature of computer packages that perform molecular dynamics and we are only interested in showing the reliability of the present approach to calculate T_{conf}. Other expressions for the temperature exists,[7] but as illustrated in figure 4.1, equation 4.1 suffices for the present purposes.

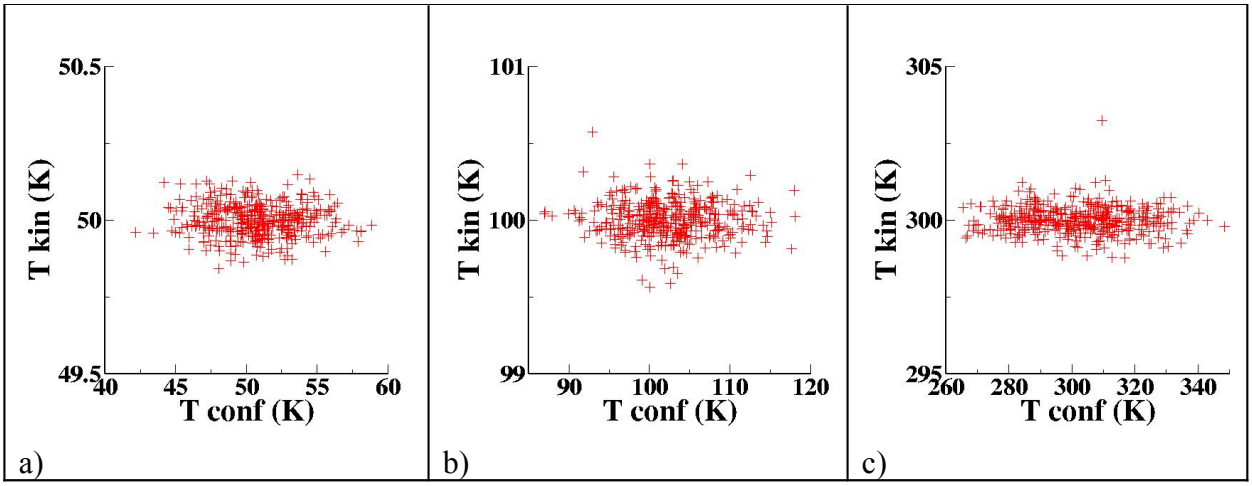


Figure 4.1. Molecular dynamics simulation of Crambin: a) 50K, $<T_{kin}>=50.00\pm0.05$ K, $<T_{conf}>=50.76\pm2.87$ K; b) 100K, $<T_{kin}>=99.98\pm0.50$ K, $<T_{conf}>=102.33\pm5.48$ K; c) 300K, $<T_{kin}>=299.93\pm1.63$ K, $<T_{conf}>=301.50\pm16.52$ K;

Throughout the simulations, the configurational temperature maintains values very similar to T_{kin}. Notice that the standard deviation of T_{conf} for the simulation at 100 K is 5.48 K. The approach requires a sufficiently large number of degrees of freedom. In other MD runs - not shown - we observed that it is already successful for anthracene and heptane.

Crystallographic structures are continuously deposited. The Brookhaven database [8] contains more than 40,000 protein and DNA structures. Protein structures were obtained in pdb format from the PAPIA (Parallel Protein Information Analysis) service [9]. Only X-ray derived structures were considered, with a minimal resolution of 2.0 Å. The minimal number of residues was set to 40, the sequence similarity was set to less than 20%. A total of 935 structures was downloaded; 85% of them provide the temperature of the experiment, that we call T_{PDB}.

Of these 935 structures we calculated configurational temperature. Gradients and Laplacians were calculated with the TINKER program[19-21] which has found a number of applications in our

laboratory.[22-25] Trivial hot spots such as atoms with dangling bonds were removed from the sums in equation. Force field used was OPLS[17].

We noticed the presence of about 60 outliers. Outliers are dued to errors in atom coordinates and lack of chains often present in the deposited structures. In corrispondence of the chain problems T usually becames enormous and affects all the calculation.
We have developed an algorithm to solve this problem. Protein chains were divided into small blocks of 20-atoms size. Configurational temperature was calculated for all blocks. We applied standard statistical method to treat outliers, discarding critical values over $<T>+7\sigma$: then a new average value was calculated.
Average of the temperature of the experiment, $<T_{PDB}>$, is 130.85 K. Average of configurational temperature, $<T_{conf}>$, is 134.99 K. There is a good agreement because experimental data and our calculations.

Experimentally, the majority of the structures are measured at 100 K. This temperature prevents radiation damage to the crystal and allows the collection of more data than from an unfrozen crystal. It also avoids the water transition that takes place in hydrated proteins, at 220 K, with the sudden onset of anharmonic and liquid-like motion.[10-16]
In order to make the statistical treatment straightforward, it was decided to consider only the structures recorded at 100 K. Only structures with similar average gradients were considered The final sample was 538 proteins. In this case $<T_{PDB}>$, is obviously 100 K, and we found an average value of $<T_{conf}>=107.45$. Again there is a good agreement between experimental and calculated data.

4.4 CONCLUSION

We have calculated configurational temperature for a database of experimental structures. To our knowledge, it's the first time that a such thing is done. We show a good agreement beetween out calculations and experimental data, that is an encouraging starting point.
But a variety of issues remain open. For instance, one could ask how the present approach performs with other structural databases such as the Cambridge one; or how the protein structures obtained from NMR measurements behave with respect to T_{conf}; or if the structures are equilibrated or present hot spots. I believe that the concept of configurational temperature coupled to data mining in structural databases will deliver a host of important information in the close future and will be

useful in structural refinement, and I hope to continue my work about it.

4.5 REFERENCES

[1] Hirschfelder, J. O. *J. Chem. Phys.* **1960**, *33*, 1462-1466.

[2] Powles, J. G.; Rickayzen, G.; Heyes, D. M. *Mol. Phys.* **2005**, *103*, 1361-1373.

[3] Rugh, H. H. *Phys. Rev. Lett.* **1997**, *78*, 772-774.

[4] Butler, B. D.; Ayton, G.; Jepps, O. G.; Evans, D. J. *J. Chem. Phys.* **1998**, *109*, 6519-6522.

[5] Han YL, Grier DG Configurational temperature of charge-stabilized colloidal monolayers *Phys. Rev. Lett.,* **2004**, 92 (14)

[6] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, **1984**, 81, 3684-3690

[7] Owen G. Jepps, Gary Ayton, and Denis J. Evans, Phys. Rev. E 62, **2000**, 4757-4763

[8] http://www.rcsb.org/pdb/Welcome.do

[9] http://www.cbrc.jp/papia/papia.html

[10] Parak, F. & Knapp, E. W. (1984) *Proc. Natl. Acad. Sci. USA,* **1984,** 81, 7088–7092.

[11] Doster, W., Cusack, S. & Petry, W. (1989) *Nature,* **1989** 337, 754–756.

[12] Doster, W., Cusack, S. & Petry, W. *Phys. Rev. Lett.,* **1990,** 65, 1080–1083.

[13] Rasmussen, B. F., Stock, A. M., Ringe, D. & Petsko, G. A., **1992,** *Nature* 357, 423–424.

[14] Paciaroni, A., Bizzarri, A. R. & Cannistraro, S. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* ,**1999** , 60, R2476–R2479.

[15] Caliskan, G., Kisliuk, A. & Sokolov, A. P. *J. Non-Crystalline Solids,* **2002***,* 307–310, 868–873.

[16] S.-H. Chen, L. Liu, E. Fratini, P. Baglioni, A. Faraone, Mamontov *Proc. Natl. Acad. Sci. USA,* **2006**, 103, 9012–9016.

[17] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr. and P. Weiner, "A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins", J. Am. Chem. Soc., **1984**, 106, 765-784

[18] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins,

J. Phys. Chem. B **1998**, 102, 3586 – 3616.

[19] Dudek, M.J.; Ponder, J.W. Accurate modeling of the intramolecular electrostatic energy of proteins. *J. Comp. Chem*., **1995**, *16*, 791-816,

[20] Kundrot, C.E.; Ponder, J.W.; Richards, F.M. Algorithms for calculating excluded volume and its derivatives as a function of molecular conformation and their use in energy minimization. *J. Comp. Chem.*, **1991**, *12*, 402-409.

[21] Ponder, J.W.; Richards, F.M. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comp. Chem*., **1987,** *8*, 1016.

[22] Biscarini F, Cavallini M, Leigh DA, León S, Teat SJ, Wong JKW, Zerbetto F. The Effect of Mechanical Interlocking on Crystal Packing, Predictions and Testing. *J. Am. Chem. Soc.*, **2002**, *124*, 225-233.

[23] Hoefinger, Siegfried; Zerbetto, Francesco. Simple models for hydrophobic hydration. Chemical Society Reviews **2005**, 34(12), 1012-1020.

[24] Trebbi, Bruno; Dehez, Francois; Fowler, Patrick W.; Zerbetto, Francesco. Favorable Entropy of Aromatic Clusters in Thermophilic Proteins. Journal of Physical Chemistry B **2005** 109(38) 18184-18188.

[25] Lugli, Francesca; Hoefinger, Siegfried; Zerbetto, Francesco. The Collapse of Nanobubbles in Water. Journal of the American Chemical Society **2005**, 127(22), 8020-8021.

CHAPTER 5

CONCLUSIONS AND GENERAL REMARKS

In this Ph.D research thesis I have applied computational models to proteins, investigating about different properties, reporting the more significant results: intra-residue energy distribution of proteins (i), aromatic stabilization (ii), and configurational temperature (iii).
Results are encouragins, so one can say that computational methoods, and in particular molecular mechanics, are very suitable to describe biological system like proteins.

In (i) we analized the energy distribution of a protein database, finding that they follow Boltzmann's law.
In (ii) we found that aromatic fragments in thermophilic proteins tend to generate larger entropy via their overall low-frequency motion. This feature indicates one direction for exploration in connection with rational design of ultrastable proteins.
In (iii), finally, we have calculated configurational temperature for a database of experimental structures, showing a good agreement beetween out calculations and experimental data.

For more details, please refer to single chapters' conclusion paragraph.