0.4

ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

# The Role of Native State Topology in Protein Folding and Dynamics

Thesis submitted for the degree of
*Doctor Philosophiæ*

**Candidate:**
Giovanni Settanni

**Supervisors:**
Prof. Amos Maritan
Prof. Paolo Carloni

October 2001

SISSA - SCUOLA
INTERNAZIONALE
SUPERIORE
STUDI AVANZATI
TRIESTE
Via Beirut 2-4

TRIESTE

SISSA ISAS

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

# The Role of Native State Topology in Protein Folding and Dynamics

Thesis submitted for the degree of
*Doctor Philosophiæ*

**Candidate:**
Giovanni Settanni

**Supervisors:**
Prof. Amos Maritan
Prof. Paolo Carloni

October 2001

# Table of contents

# Introduction

Proteins are heteropolymer chains of amino acids universally produced in living cells after transcription of genetic materials (DNA, mRNA). In eucariotic cells the transcription takes place in the ribosomes (giant molecules composed of protein and RNA chains). During transcription amino acids are assembled one after the other to form the protein. After transcription the proteins migrate to their specific cell compartment according to their function.

Most of the proteins assume a defined three-dimensional structure in order to accomplish their function. The process that leads the protein to its defined three-dimensional structure is called protein folding. Protein folding can take place directly after transcription or during migration of proteins. It can be spontaneously driven or catalyzed by other agents (chaperons).

After reaching their own cell compartment and their own three-dimensional structures, proteins interact with the components of their environment (that can be other proteins, nuceleic acids or any other molecule present in around) and carry out the most diverse functions: from catalysis or inhibition of various chemical or physical reactions, to the transport of other molecules, electrons, protons, excitons, to the stabilization of the architecture within the cell.

A fundamental issue in order to understand protein function is represented by the folding process, which leads the protein from a conformation with no apparent order to a somewhat defined three-dimensional structure. One of the very first characteristics of the folding process to be identified was the one-to-one correspondence between the precise amino acid sequence that constitutes a protein and the three-dimensional structure that it assumes in solution (1). The presence of this privileged state, which can be (almost) spontaneously and reversibly reached, led to consider the proteins as thermodynamic stable systems with a free energy minimum in their native conformation. It was immediately clear that the space of possible protein conformations is simply too large to be explored through random diffusion in the search for the free energy minimum (2). The very first hypothesis to explain this paradox pointed to the presence of a privileged pathway from the unfolded conformations to the native, correctly folded conformation.

On a parallel way, the theoretical study of model heteropolymers led to the conclusion that sequences folding in a unique native state represent an exception in the set of the possible sequences. Physics of random heteropolymers presents many aspects that are similar to Physics of spin glasses. Indeed simple model studies (3) indicate that, heteropolymers display many equivalent free energy minima and usually do not reach a unique native conformation. Thus, natural protein sequences should have been the product of an evolutionary process that has led to the selection of amino acid sequences that have a unique native state.

The lack of non-native metastable conformations and of local free energy traps in many real proteins (that is the lack of "frustration" according to the language of spin glass theory) led to the formulation of models able to reproduce this aspect. Possible source of frustration in the folding of proteins can lay in interactions between amino acids; in principle not all the interactions between amino acid in the native conformation are favorable to its formation; some of them may be unfavorable, i.e. there may be an electrostatic repulsion between two neighboring amino acid or an hydrophobic amino acid may prefer a more buried conformation. However all these unfavorable contributions are counterbalanced by overall favorable interactions; the protein sequence folds notwithstanding their presence and frustration is reduced as much as possible (4).

On the other hand, a sort of frustration can also lay in the structure of the protein itself. Indeed, the connectivity of the peptidic chain poses several constraints on the possible pathways from the unfolded to the folded state. Thus, the number of pathways that really connect unfolded conformations to the folded one varies according to the structure of the native state. This type of frustration, denoted as topological frustration, remains even if interactions between amino acids have been optimized to fully favor the native conformation. It has a markedly entropic nature(5). The topological frustration does not depend on the details of the sequences being folded but more in general will play the same role on all the sequences that have the same native state.

In the past few years an increasing amount of experimental evidence has supported this view. Perhaps the best example is provided by the close conformational similarity of the transition states of proteins having structurally-related native states (and yet poor sequence similarity) (6, 7). Also, the pioneering work of Plaxco and coworkers (8) has revealed how simple topological properties of native states, such as the contact order, have a strong impact on folding rates.

A simplified but useful point of view is to regard the folding process from a mere geometric perspective, where a random loose conformation (denatured state) undergoes a series of changes that take it through states with an increasing similarity with the native state. The topology of the latter ultimately dictates the series of obligatory rate-limiting steps encountered along the possible folding pathways. In principle, detailed chemical interactions may well play a role as important as topology itself (intended as pure configurational entropic bias), thereby complicating this simple picture. Although this is certainly true for some individual proteins, the range of validity of the topological picture appears to be surprisingly wide. A confirmation of this view is provided by the remarkable accord of the key folding stages predicted by topology-based models and the available experimental results (9, 10) as will be shown in this thesis. Indeed, the success of the topological picture itself helps to explain why the folding process is not too sensitive to the detailed chemical composition of most residues in a protein.

In fact, the dependence of the folding process on the detailed chemistry is much more subtle than that given by the native state topology. Although for many proteins mutations at various sites are not relevant to the folding mechanism there do exist sites at which the chemical details appear to be very important. The existence of those key sites can be proved directly from the emergence of the conserved residues (11) in homologous proteins as well as those involved in formation of the transition states (12). From a theoretical viewpoint it can be anticipated that those key residues take part to contacts that are crucial for the folding process. In other words, the dramatic configurational entropy reduction encountered in the folding bottlenecks must be accompanied by a fine-tuned interplay of chemical interactions to lower the free energy barrier and thus avoiding permanent trapping in metastable states. The establishment of the key contacts leads to a rapid formation of further interactions. Interestingly, it has been shown that those key contacts can be identified by just knowing the topology of the native conformation. Thus, the topology itself also dictates the impact of chemistry on the folding process.

As we mentioned before, many proteins need to reach their native conformation in order to accomplish their biological function. In fact, the native conformation encloses some characteristics that are necessary to the function of the protein. A specific surface charge distribution, the presence of binding sites for interaction with other molecules represents aspects that a fixed native conformation can supply. However, native conformations of proteins are far from being fixed. They retain an intrinsic flexibility that in some cases may be strictly related to the function of the protein. Several instances of this kind of functionality have been reported (13, 14). The flexibility properties of a protein are still connected to its structural topology. Furthermore, they depend on the intra molecular interactions as well as on protein-solvent interactions. In order to study the influence of the flexibility on the protein function in comparison to the influence of other factors like the electrostatics interactions or the solvent mediated interactions a high level of detail is required. Classical molecular dynamics, which describes the motions of the protein atoms and of the surrounding solvent molecule, is the tool of choice. The trajectory has to be subsequently analyzed in order to extract the collective motions that may characterize the protein.

The content of this thesis is organized as follows. In the first chapter, the main computational technique used in this work are presented along with a detailed description of the topological approach to the study of protein folding, which has been developed during these years at SISSA. The second chapter is devoted to the study of the folding of intracellular antibodies a class of proteins particularly interesting because of their possible use in functional genomics experiments performed on living cells. These proteins have been selected in SISSA labs as blocking agents of tau protein (15), a molecule involved in the onset of Alzheimer's neurodegeneration. The third chapter is devoted to the study of the folding characteristics of the prion protein that is considered the pathogenic agent of a class of neurodegenerative diseases known as prion diseases. The fourth chapter is dedicated to the study of the dynamic characteristics of the complex formed by the Nerve Growth Factor (NGF) and its extracellular receptor TrkA. Those proteins are also involved in the complex phenomenology that characterizes the Alzheimer's disease.

# Chapter 1 - Topological Approach to Protein Folding

In this chapter we describe the methods that we used to carry out our work. More detailed descriptions of the standard methods are referenced in the text.

In the case of a random heteropolymer, there is little structural correlation among low energy conformations (16). Foldable sequences have exactly the opposite property. That is, interactions between amino acids that favor the folded conformation simultaneously disfavor the unfolded ones. Both effects are equally important, since folding is not only determined by properties of the folded state but also by the energetic difference between the folded and unfolded ensembles of states.

We now imagine the following “ideal” protein model. (For simplicity, we consider the case where energetic terms include only interactions between non-covalent beads in spatial contact, i.e., folding is driven by tertiary interactions.) In such a situation, a good order parameter to measure the degree of nativeness of any configuration is the fraction of native contacts Q (contacts that exist in the native conformation). Q varies between 0 and 1 with the native conformation having Q = 1. Given a desired structure, how does one determine a good sequence that folds to it? A homopolymer with attractive interactions between residues would favor all the native contacts, but it would also favor all other contacts with the same attractive interaction. Good folding sequences have to favor the native interactions or tertiary contacts but also disfavor the non-native contacts.

If we had unrestricted flexibility in designing the sequence, we would make sure that only the native interactions were attractive. In addition, well-designed sequences would have most of this attractive energy equally divided among all contacts. Ideally designed sequences should have the energy of their conformations proportional to Q (5, 17-19). Real sequences have additional roughness introduced by attractive non-native contacts and/or repulsive native ones. This correlation between energy and structure provides a uniform bias toward the native fold for all configurations with the same degree of similarity to the folded state. Random heteropolymers have no correlation between energy and Q and therefore have a landscape where the roughness is dominant over the bias to any individual conformation.

Given a desired structure, what would be the folding mechanism for an ideal sequence as described above, i.e., one where only the native contacts are attractive with equal strength? Would all the pathways towards the folded state be equivalent as just described, independent of the shape of the native structure? Equivalently, is the order in which the native contacts are formed in a folding pathway irrelevant? If such a situation were true, the order of native contact formation would not be determined by the topology of the native structure, and there would be an absence of geometrical constraints in the connectivity of locally similar conformational states. We know, however, that this “ideal” situation does not apply to real proteins. The origin for these “topological” differences among pathways (or contact order formation) is due to the polymeric nature of the chain (4). Because of the folded shape of the protein, the geometrical accessibility of different native contacts is different (20), and therefore some are more easily formed than others, i.e. contacts between residues which are close to each other along the sequence have a higher probability to form a contact. In addition, some contacts may be topologically required (or at least be more likely) to be formed before others during folding. In this thesis, we focus our discussion on the nature of the transition state ensemble for real proteins and how energetic or topological effects may determine their structural details.

In order to elucidate the importance of the topological features of a protein a simplistic representation of its three dimensional structure is adopted (9, 10, 21). The C$_\alpha$ atom positions are only retained to represent the amino acids. Then, the contacts between amino acids are determined by simply checking the relative distance of the C$_\alpha$'s. If the distance is less than a cut off of 6.5Å (usually this can be set from 6.5 and 8.0 (22)) then the amino acids are considered in contact, no contact otherwise.

### *Molecular dynamics simulations*

This standard tool of scientific computations has been widely used in this work. It has been exploited as a sort of driver for the exploration of phase space of our model proteins and for the determination of the kinetic pathways described by them.

A molecular dynamics simulation involves the solution of the Newton's equation of motion for a system of $N$ particles of masses *{m$_i$}* and positions *{***r**$_i$*}*.

$$m_i\ddot{\mathbf{r}}_i = -\nabla_{\mathbf{r}_i} E \qquad i = 1,\ldots,N \qquad \text{Eq. 1.1}$$

The potential energy $E$ acting on each particle can be described as a function of the positions of all the particles in the system: $E = E(\{\mathbf{r}_i\})$.

The energy function for our system is a standard in biopolymer MD simulations (23, 24). It is composed by a "bonded" and a "non-bonded" term and written as:

$$E = V_B + V_{NB} \qquad \text{Eq. 1.2}$$

$V_B$ represents a potential between close-in-sequence C$_\alpha$'s and is supposed to mimic the peptide bond and the geometrical constraints of close-in-sequence amino acids (i.e. Ramachandran angle propensities).

Its detailed form is:

$$V_B = \quad g\cdot V_p + h\cdot V_a + k\cdot V_d \qquad \text{Eq. 1.3}$$

$$V_p = \quad \sigma\sum_{i=1}^{N-1}\left(r_{i,i+1} - r_{i,i+1}^{(n)}\right)^2$$

$$V_a = \sigma\sum_{i=1}^{N-2}\left(\theta_{i,i+1,i+2} - \theta^{(n)}_{i,i+1,i+2}\right)^2$$

$$V_d = \sigma\sum_{i=1}^{N-3} 1-\cos\left(\tau_{i,i+1,i+2,i+3} - \tau^{(n)}_{i,i+1,i+2,i+3}\right) \ .$$

where $r_{i,j}$ is the distance between residue i and j, $\theta_{i,j,k}$ is the angle with the j-th amino acid as vertex and the i-th and the k-th as edges and $\tau_{i,j,k,l}$ is the dihedral generated by the i-th, the j-th, the k-th and the l-th amino acid. The $n$ as superscript means the native state value. $\sigma$ represents a suitable scale factor to fix the temperature scale. As formulas show, the minimum of the potential energy is set in the native state geometry, independently of the coupling constants $g$, $h$ and $k$. Tests have been done with different sets of coupling regimes in order to check the robustness of the analysis (see Chapter 2).

The non-bonded term represents, in a very simplified way, the interactions that take place between far-in-sequence residues and that keep the protein in its native state(17). The interactions are present for each pair of amino acids (but the sequence neighbors) and are accounted for by a Van der Waals-like potential of the type:

$$V_{NB} = \sigma\sum_{i<j-3}^{N} V_{i,j} = \sigma\sum_{i<j-3}^{N}\left(5\cdot\left(\frac{r^{(n)}_{i,j}}{r_{i,j}}\right)^{12} - 6\cdot\left(\frac{r^{(n)}_{i,j}}{r_{i,j}}\right)^{10}\cdot\Delta_{i,j}\right) \qquad \text{Eq. 1.4}$$

where $\Delta_{i,j}$, the contact map, is 1 or 0 if the i-th and j-th $C_\alpha$'s are in contact or not in the native state respectively. This means that every pair of contacting $C_\alpha$'s in the native state interacts through an attractive potential whereas non-contacting $C_\alpha$'s feel just hard-core repulsion. Low values of $V_{NB}$ correspond to the protein in the folded conformation while high values correspond to the unfolded state.

Given the energy function for the molecular interactions, the trajectory of the system is in principle determined by its initial conditions: atomic positions and velocities. In practice, the Newton's equations of motions cannot be solved analytically. The integration of the equations of motion is achieved using finite difference methods. The time variable is discretized by assuming that no variations of the force acting on the atoms occur in the time step δt. Various algorithms have been proposed for the numerical integration, differing for accuracy and computational speed (25-27).

We used the velocity Verlet algorithm that gives the positions **r** and velocities **v** at time $t+\delta t$ using the following formulas (26):

$$\begin{cases} \mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) \\ \mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t[\mathbf{a}(t) + \mathbf{a}(t+\delta t)] \end{cases} \qquad \text{Eq 1.5}$$

where **a** is the acceleration.

This algorithm has also the property to preserve the phase space measure (25).

MD runs for folding models of have been carried out with a time-step of 0.005.

The temperatures during the runs have been kept constant using the Berendsen algorithm (26) with a coupling constant set to 0.01. The method was first illustrated by Berendsen *et al.* (28). The system is virtually coupled with a thermal bath at temperature $T_0$. For each atom *i*, a Langevin equation replaces the equations of motion to describe the coupling between the two systems:

$$m_i\mathbf{a}_i = \mathbf{F}_i + m_i\gamma\left(\frac{T_0}{T} - 1\right)\mathbf{v}_i \qquad \text{Eq. 1.6}$$

where T is the instant temperature of the system derived from the average kinetic energy, and γ is the friction coefficient.

At each time step $\delta t$ , Eq. 6 is equivalent to rescale the velocities by a factor $\lambda = 1 + \frac{\delta t}{2\tau}\left(\frac{T_0}{T} - 1\right)$, where $\tau = (2\gamma)^{-1}$ leading to en effective temperature $T(t+\delta t) = T(t) + \frac{\delta t}{2\tau}(T_0 - T(t))$.

The motion of the center of mass is usually removed every 1000 steps to prevent an erroneous computation of the total kinetic energy. For each model protein that we studied, we have usually performed a scan of the whole range of temperatures: from a temperature where the native state is fully stable to temperatures well above the complete unfolding transition temperature. The temperature of each successive run has been increased by a fixed amount. Then, a cooling process was also carried out using a reverse procedure and was used to assess the reliability of the configurational sampling. Each run consisted of a number of time steps that allowed for a good exploration of the whole phase space, as will be illustrated below.

### *Analysis of trajectories*

The Swendsen-Ferrenberg (29) method has been applied to Van der Waals energy data ($V_{NB}$) from simulations.

This method allows obtaining the density of states of the system combining the trajectories at different temperature. The data from the simulation are represented as histograms. The $i$-th histogram, $n_i(E)$, is the number of times the energy $E$ was observed in the $i$-th run. $N_i$ is the number of bins in the $i$-th histogram.

The density of states is:

$$W_i(E) = \frac{n_i(E)}{N_i}\exp\left(\frac{E-F}{k_B T_i}\right) \qquad \text{Eq. 1.7}$$

Ferrenberg and Swendsen make the ansatz that the density of states can be expressed as a weighted sum:

$$\begin{aligned} W(E) &= \sum \omega_i W_i(E) \\ \sum \omega_i &= 1 \\ \delta^2 W(E) &= \sum \omega_i^2 \delta^2 W_i(E) \\ \delta^2 W_i(E) &= N_i^{-2}\exp\left(\frac{2E-2F_i}{k_B T_i}\right)(1+2\tau_i)n_i(E) \\ n_i(E) &= N_i W_i(E)\exp\left(\frac{F_i-E}{k_B T_i}\right) \end{aligned} \qquad \text{Eq. 1.8}$$

where $\tau_i$ is the correlation time for the $i$-th run.

The method is based on the iterative solution of the following:

$$W(E) = \frac{\sum (1+2\tau_i)^{-1} n_i(E)}{\sum N_i (1+2\tau_i)^{-1}\exp\left(\frac{F_i-E}{k_B T_i}\right)} \qquad \text{Eq. 1.9}$$

$$\exp\left(-\frac{F_i}{k_B T_i}\right) = \sum W(E)\exp\left(-\frac{E}{k_B T_i}\right)$$

The 200'000 equilibration steps at the beginning of each run have been discarded. The snapshots for the analysis are collected every 500 steps to allow proper decorrelation. That method helps in interpolating data from runs at different temperatures and in building a temperature independent density of states. With that, we can, in principle, access all thermodynamic quantities, such as the mean potential energy, specific heat, etc. In particular, we used the peaks in the specific heat to pinpoint the folding transition temperatures.

A more detailed analysis has been carried out at single contact level, starting from data collected in the previous step. We measured the temperature dependence of the energy associated with each contact near transitions, through the single contact specific heat:

$$\frac{d\langle V_{ij}\rangle}{dT} = \left(\langle V_{ij}V_{NB}\rangle - \langle V_{ij}\rangle\langle V_{NB}\rangle\right)/kT^2 \qquad \text{Eq. 1.10}$$

This quantity also measures the energy change due to a mutation in the strength of a contact.

Indeed if $V_{NB}$ is modified multiplying each $V_{ij}$ by $\varepsilon_{ij}$ than:

$$\left.\frac{d\langle V_{NB}\rangle}{d\varepsilon_{hk}}\right|_{\varepsilon_{ij}=1} = \left.\frac{d\left\langle \sigma\sum_{i<j-2}\varepsilon_{ij}V_{ij}\right\rangle}{d\varepsilon_{hk}}\right|_{\varepsilon_{ij}=1} = \qquad \text{Eq. 1.11}$$

$$= \sigma\left(-\left(\langle V_{hk}V_{NB}\rangle - \langle V_{hk}\rangle\langle V_{NB}\rangle\right)/(kT) + \langle V_{hk}\rangle\right) = \sigma\left(-T\cdot\frac{d\langle V_{hk}\rangle}{dT} + \langle V_{hk}\rangle\right)$$

$\varepsilon_{ij}$'s are dummy coefficients that represent the change of strength in the contact between amino acid $i$ and $j$ and are eventually set to 1 to get the original energy function. The variation of the average total potential energy of the protein model is directly related to (minus) the specific heat contribution of the modified contact. This means that, as far as the modifications are not very large, the stabilization of a contact with high specific heat would increase the stability of the model more than the same modification on a contact with small specific heat. The formula above, thus, represents the susceptibility of the system to a change in the strength of a contact. It can be directly related to experimental data on the folding capabilities of mutated protein sequences..

A measure of the role played by a given amino acid at *i*-th position along the chain during the (un)folding is obtained by:

$$c_i \equiv \sum_j \frac{d\langle V_{ij} \rangle}{dT} \Delta_{ij} \qquad \text{Eq. 1.12}$$

Which can be considered as the contribution to the specific heat related to the *i*-th amino acid position.

The single residue specific heats (SRSH) have been used to identify folding and unfolding phases and to pinpoint residues that play a relevant role in the process.

# Chapter 2 - Intracellular Antibodies

The organization of this chapter is the following. In the first part we present an introduction to the biology of intracellular antibodies. This part is meant to expose the actual knowledge of the subject, the open issues and the motivations of our work. In the second part of the Chapter, the results obtained exploiting the topological model of the Ig variable domain are presented and discussed; in the third part an analysis of the intrabody sequences selected by our collaborators of the Biophysics and Molecular Genetics sector is performed and the results are discussed. Finally, a conclusive part is dedicated to the analysis of the convergent results from the two approaches.

## Biological Introduction

Antibodies are secreted by plasma cells and have evolved to act in a variety of compartments of the mammalian body, outside of cells. The demands on stability have kept a selection pressure on immunoglobulin domains, to retain disulfide bonds in all germline immunoglobulin genes. In fact, all antibody variable domains, the antigen-binding domains, contain one conserved disulfide bond linking two pairs of conserved cysteine residues. The disulfide bonds form during the process of secretion from the cell, in the endoplasmic reticulum. Recent advances in the field of recombinant antibodies have allowed engineering a wide variety of recombinant forms of antibody domains, e.g. such as the single-chain Fv (scFv) fragments (30, 31), which consist only of the variable domains connected by a linker. These simpler antibody domains can be expressed in a variety of hosts, ranging from bacteria to mammalian cells, but still exploiting disulfide formation in the host secretory pathway. However, it has been shown that, by suitable engineering, it is also possible to redirect the expression of antibodies or antibody domains away from the secretory pathway, to other intracellular compartments (32-34).

This exploits the use of intracellular targeting signals that normally dictate the intracellular fate and traffic of proteins. Antibody domains, equipped with targeting sequences "borrowed" from other proteins, have indeed been successfully targeted to novel intracellular compartments (33), where antibodies are not normally found, such as the cytoplasm, the nucleus, the mitochondria (Fig. 1). Since from the initial studies, it was clear that if a sufficient functional expression, i.e. a correct folding, of such intracellular antibodies (intrabodies) could be achieved, this would enable them to bind to their target protein in the appropriate intracellular compartment and evoke specific biological effects. In fact, intracellular antibodies have been shown to act as "protein knock-out" reagents, by inactivating the recognized protein (Fig. 2). Thus, proteins involved in signal transduction (p21 ras, (35, 36)), plant viral pathogenesis (37), human

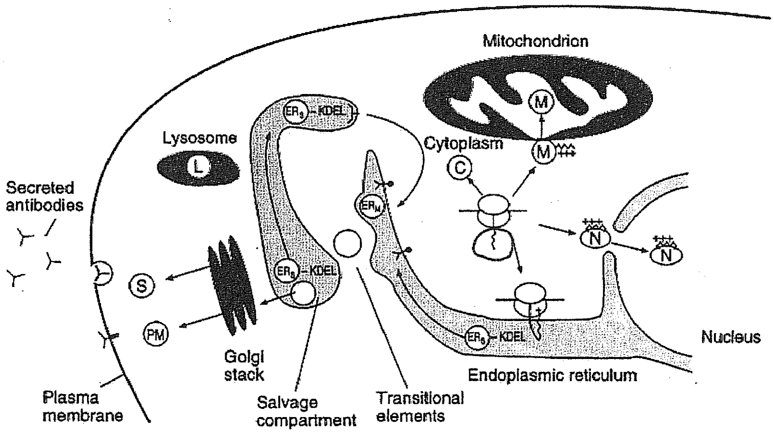


Figure 1. Antibody expression and targeting: according to the signal sequence (red circled) the antibody sequences can be directed along the natural secretory pathway and secreted or can be directed to different cell compartments like the nucleus (N), the cytoplasm (C) or the mitochondria (M).

virus replication (38-40) have been successfully blocked by this emerging technology. This has an immediate application for research, were tools to inactivate proteins are fundamental in the process of elucidating protein function. In perspective, the intrabody technology has been proposed to have broad therapeutic applications, possibly in a gene therapy setting (41).

More recently, the intrabody technology has been proposed as a potential tool of choice in functional genomics and functional proteomics (42). The completion of the sequencing of the human genome calls for methods to assess the function of genes, coming from genome sequencing projects or from proteomic screens. Protein knock out by the intrabody technology promises to become an essential tool in the endeavour to facilitate the discovery of gene and protein function, provided its present bottlenecks are solved.

As discussed above, in principle antibody domains can be directed to all intracellular compartments, by encoding the corresponding targeting sequence attached to that encoding for the antibody. Among the different intracellular locations, the cytoplasm is a cross road, since while proteins targeted to the secretory pathway are cotranslationally targeted to the endoplasmic reticulum, proteins targeted to other intracellular compartments are first synthesized in the cytoplasm, as also the proteins resident of the cytoplasm itself. For antibodies, expression in the cytoplasm is the most difficult task, because of its reducing environment (43, 44). This reducing potential prevents the formation of disulfide bonds, including the conserved intradomain disulfides in antibody domains (45, 46). Indeed, it was demonstrated that scFv fragments expressed in the cytoplasm do not form the disulfide bonds (33, 47, 48). The intradomain disulfide contributes 4-5 kCal/mol to the stability of antibody domains (49, 50).

Therefore, antibody fragments expressed in a reducing environment are strongly destabilized, compared to the same molecules containing disulfides, and a smaller fraction of these antibody domains is likely to fold to the correct native structure. This fact is believed to be responsible for the fact that not all antibodies perform equally well, when expressed in the cell cytoplasm (51). Indeed, while a number of cytoplasmically expressed antibody fragments were reported to show specific and well controlled biological effects (see (42) for a review of representative examples), the average scFv fragment isolated from the corresponding monoclonal antibody or from a phage library will fold ineffi-

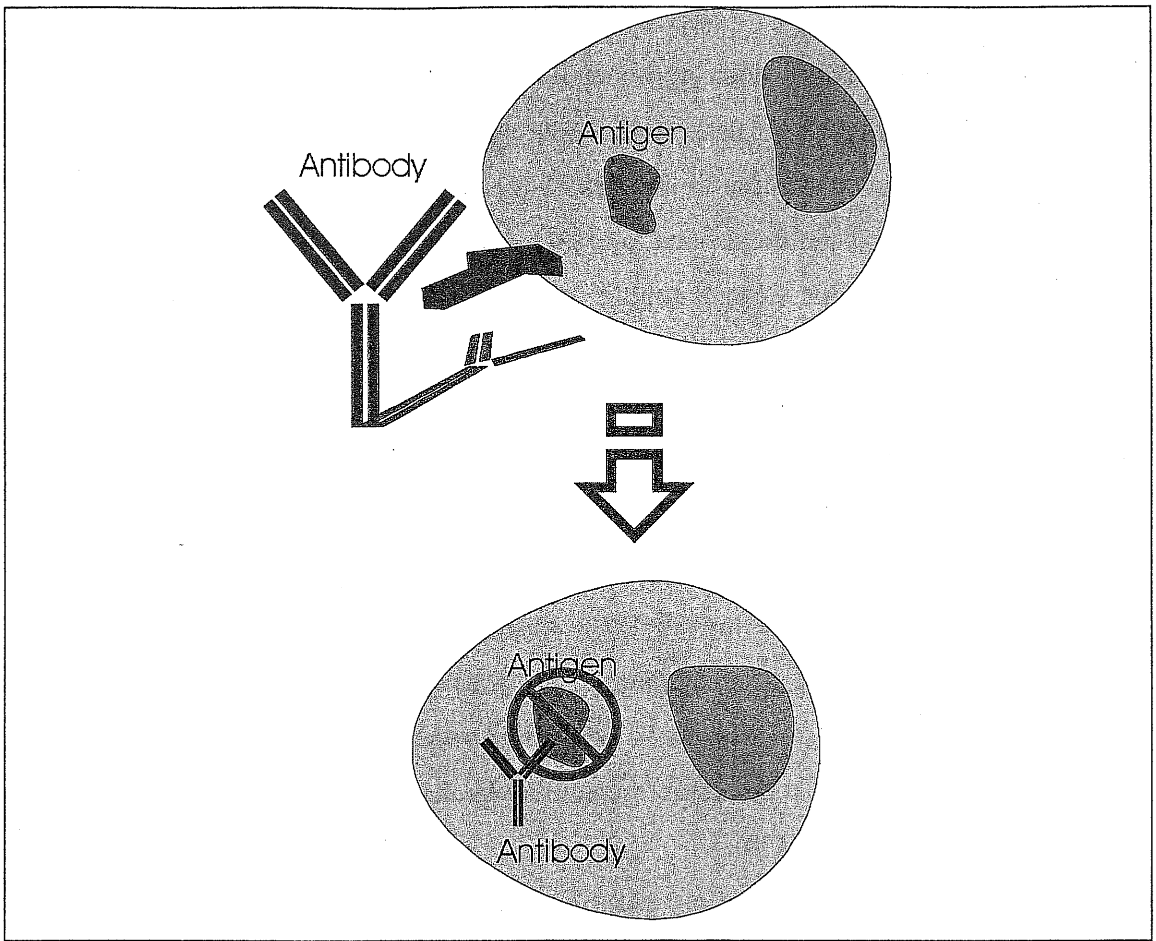


Figure 2. "Protein Knock-out" by intrabody technology; the great binding capabilities of antibodies, normally exploited for extra-cellular antigens and in *in vitro* experiments, could be used in internal cell compartments as soon as a good technique was available for their intracellular expression. The importance of such a technique in functional genomics  is evident.

ciently in the cell cytoplasm. Therefore, for many applications, antibody of interest may require further ad hoc optimization by protein engineering, to make it more stable (52). Moreover, if intrabody technology is to hold its promises as the technology of choice for high throughput functional genomics  , a reliable and/or a priori predictable access to these molecules is needed.

Inspired by the study of a small number of naturally occurring antibodies which, due to somatic mutations, lack the intradomain disulfide bond and yet appear to fold correctly (53), tolerating the absence of this stabilizing bond, it is now believed that the overall stability of antibody domains is contributed in a "distributed" way by many residues, with the disulfide bond being one of the different concurring stabilizing factors (54). Thus, the overall folding stability of different antibody domains covers a wide range, and those antibodies naturally falling near the upper stability range will tolerate the absence of the disulfide bond, while those near to the lower stability limit will not. Consistent with this prediction, it was found that antibodies of the latter class could be engineered to tolerate the absence of the disulfide bond by mutations elsewhere in the framework region (55-57). According to this view, it should be possible to design suitable selection procedures to isolate from natural repertoires of antibodies those that are able to bind antigen under conditions of intracellular expression. Such a selection procedure has been recently described and shown to allow the isolation of functional intrabodies (ITT or intrabody trap technology, (48, 58).

The alternative approach is to investigate the limiting factors for antibody folding stability. An understanding of the principles underlying, and the interplay between, antibody stability, affinity and their performance as cytoplasmically expressed intrabodies would greatly facilitate the rational engineering of superior antibody frameworks as hosts for CDR (complementarity determining regions) grafting (52), or the rational improvement of individual antibodies (59).

Immunoglobulin Fv domains represent an excellent testing table for the hypothesis mentioned previously and can greatly help in clarifying the range and limits of applicability of a description based on the topological assumption. They present a high degree of sequence diversity and high structure conservation (fig 1) so that in principle we can assume that the detailed amino acid chemistry has been incorporated in the native state topology through an averaging process. Moreover the importance of the folding stability of this particular protein family in the reducing environment of the cytoplasm, the role of the disulfide bond and the behavior of the domain after its removal represent aspects highly worth to be studied. For this reason we performed our calculations without considering the disulphide bridge as a covalent bond. As a comparison we have also performed calculations on titin, whose native state topology is similar to the one of Ig variable domain.

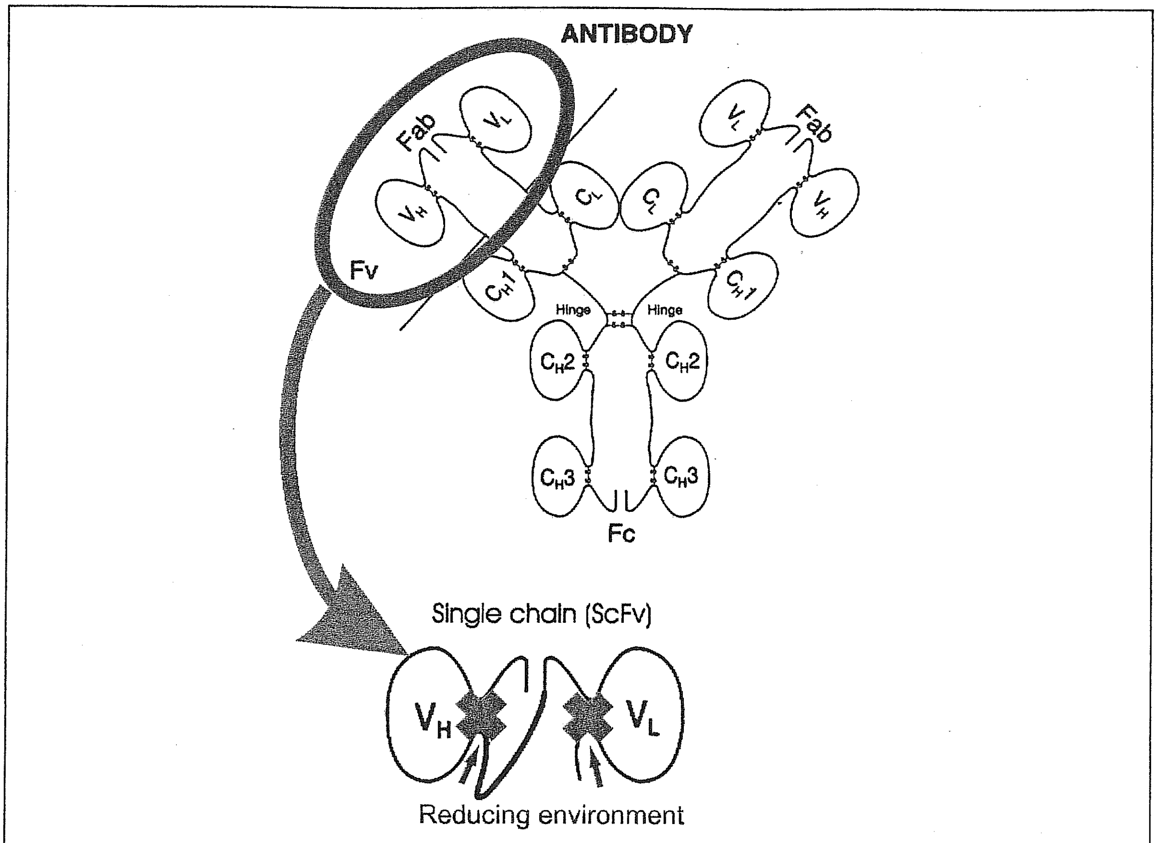


Figure 3. Full antibody and single chain fragment (ScFv). Antibody are multi-chain and multi-domain protein molecules. The component domains share a very similar structure. Each domain is stabilized by a disulphide bond. In intracellular expression conditions the disulphide bond does not form because of the cytoplasmic reducing environment.

### *Topological aspects of Ig variable domain*

In this section of the chapter an analysis of the topological features of the immunoglobulin variable domain is presented through molecular dynamics simulations of topology based coarse-grained models of the protein. The simulations are performed in the absence of the disulphide bond. The results demonstrate that the topology of the native state of the proteins contains information about relevant events of the folding pathways and their bottlenecks. The role of the geometrical position of each amino acid in the folding process is evaluated by monitoring its contribution to the specific heat at the folding transition. This measure is, thus, basically derived from the average behaviour of the system during several folding and unfolding events. Then the relationship between topologically relevant amino acid positions and the possible stabilizing mutations is presented.

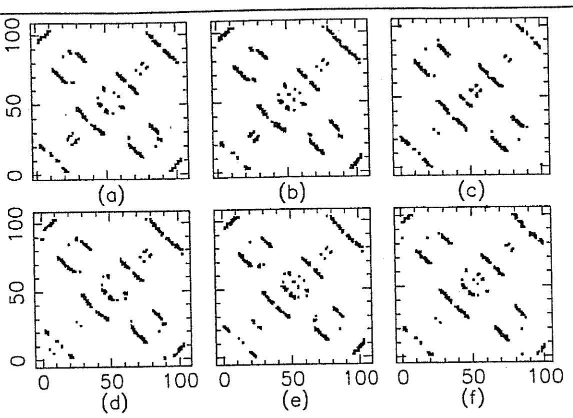


Figure 4. Contact maps of some immunoglobulin variable domains from the heavy chain and from the light chain after being structurally aligned to our model map (a). The pdb ids of the structures from which maps have been extracted are: 1mfc (b), 1fig (c), 2fb4 (d). The high structural homology is evident.

The 109 C$_\alpha$ atom positions, which have been used to represent the amino acid positions in the simplified model of Ig domain, are extracted from the crystallographic structure of an immunoglobulin variable domain (deposited at pdb with id 1mco). The choice of a particular Ig domain is irrelevant because, with the definition of contacts here adopted, many different domains both from VL and from VH region present essentially the same contact map $\Delta_{i,j}$, as shown in Fig. 4, apart from insertions and deletions in loop regions. The structural alignments have been obtained by maximizing the overlap between the pairs of structures defined as:

$$\mathcal{O} = \sum_{i<j-2} \Delta_{i,j}^{(1)} \cdot \Delta_{i,j}^{(2)} \qquad \text{Eq. 2.1}$$

where $\Delta_{i,j}^{(x)}$ 's are the aligned contact maps of the two proteins.

Three coupling regimes have been considered: in the first one, which will be denoted by **A,** the parameters used for Eq. 1.3 are $g = 50, h = 0, k = 0$, focusing on non-local interactions only; in the second regime (**B**) the settings are $g = 50, h = 9, k = 0$, partially turning on local (angular) interactions; in the third case (**C**) $g = 50, h = 9, k = 9$ have been set, to have structure-determining local interactions (that was not the case in the previous regime). The ratio between coupling parameters has been chosen to roughly match vibration frequencies of the corresponding terms in all-atom MD simulations (data not shown). For **A** and **B** *kT* was increased from 0.01σ to 0.72σ with steps of 0.01σ. For **C** *kT* was increased from 0.7σ to 1.06σ with steps of the same size as above. The masses of the amino acids have been set to 100 arbitrary units.

### *General temperature dependent behavior of the system*

The model system has shown different heating behaviors according to the coupling regime governing the bonded interactions. (But nevertheless the role of different positions along the chain doesn't depend on that).

In the case in which only the elastic bonds were present and angle and dihedral terms were set to 0 (A), the system has experienced a multiple phase folding process. The spe-

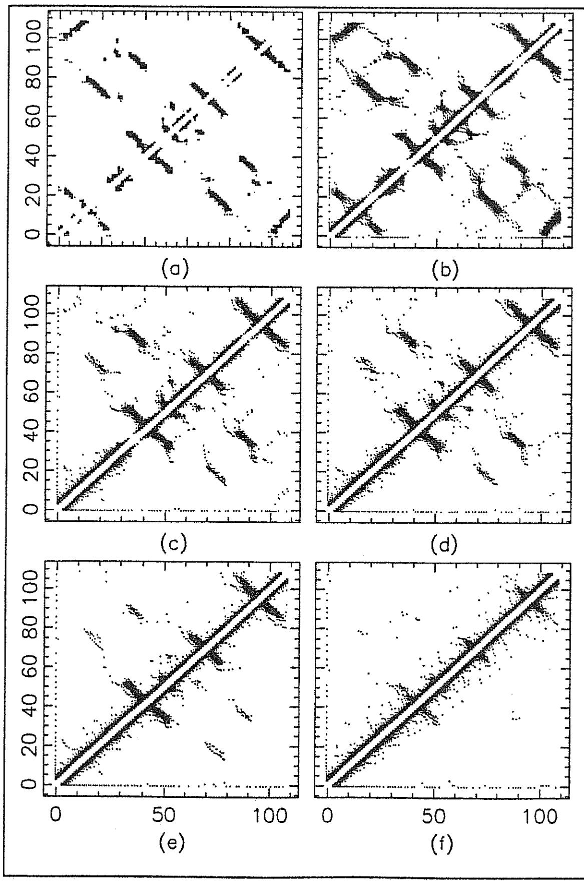


Figure 5. Time average contact maps of the protein, within regime **A**. The averages are computed at *kT* = 0.01σ (a), *kT* = 0.33σ (b), *kT* = 0.42σ (c), *kT* = 0.46σ (d), *kT* = 0.52σ (e), *kT* = 0.72σ (f). The partial unfolding of the peptide is evidenced by the residual structures present at temperatures in between the two folding temperatures (c and d) as obtained by the two peaks of the specific heat for regime **A** (see fig. 6). In regimes **B** and **C** average contact maps were as in (a) below folding transition temperature and as in (f) above it.

cific heat shows two evident peaks as is shown in fig 6, revealing two main events in the process. Mean contact maps, drawn at different temperatures (fig 5), clearly represent a partially structured configuration in-between the completely folded and completely unfolded configurations. In the partially structured configurations three out of the seven native beta strands are broken apart. The contacts that break apart in correspondence of the low temperature peak are between strands 3-8 and 20-25, and between 3-8 and 98-109 (see the contact maps in fig 5). Strands 15-25 detach from 68-80 gradually increasing the temperature from the low temperature peak to the high temperature peak. The strands between 85-93 and 35-40 break apart in correspondence of the high temperature peak.

The coupling regime with elastic bonds and angle interactions (**B**) shows a sharper behavior. Only a single peak is present in the specific heat, indicating that the transition between the folded and the unfolded conformation occurs in a single step, even though a smooth tail makes its appearance. In that case, all the strands of the domain break apart at the same temperature and only folded or unfolded conformations are seen (fig 5a and 5f). The extent to which the residues cooperatively contribute to the transition (60) is, thus, higher in this case with respect to the previous one, since the transition occurs in a single step instead of two.

Finally, when dihedral interactions are also turned on (**C**), the behavior is even sharper. The specific heat shows again a single peak but now with no tail, cooperativity being maximal. The transition temperature is much higher than in the previous cases, because the angular and torsional energy terms together make the system more stable. Indeed they are sufficient to enforce the native topology even without the Van der Waals term.

The energy density of states computed during a heating simulation is almost identical to that computed during a cooling run. This indicates that the trajectories we considered are long enough for a good sampling of the configurational space of our model.

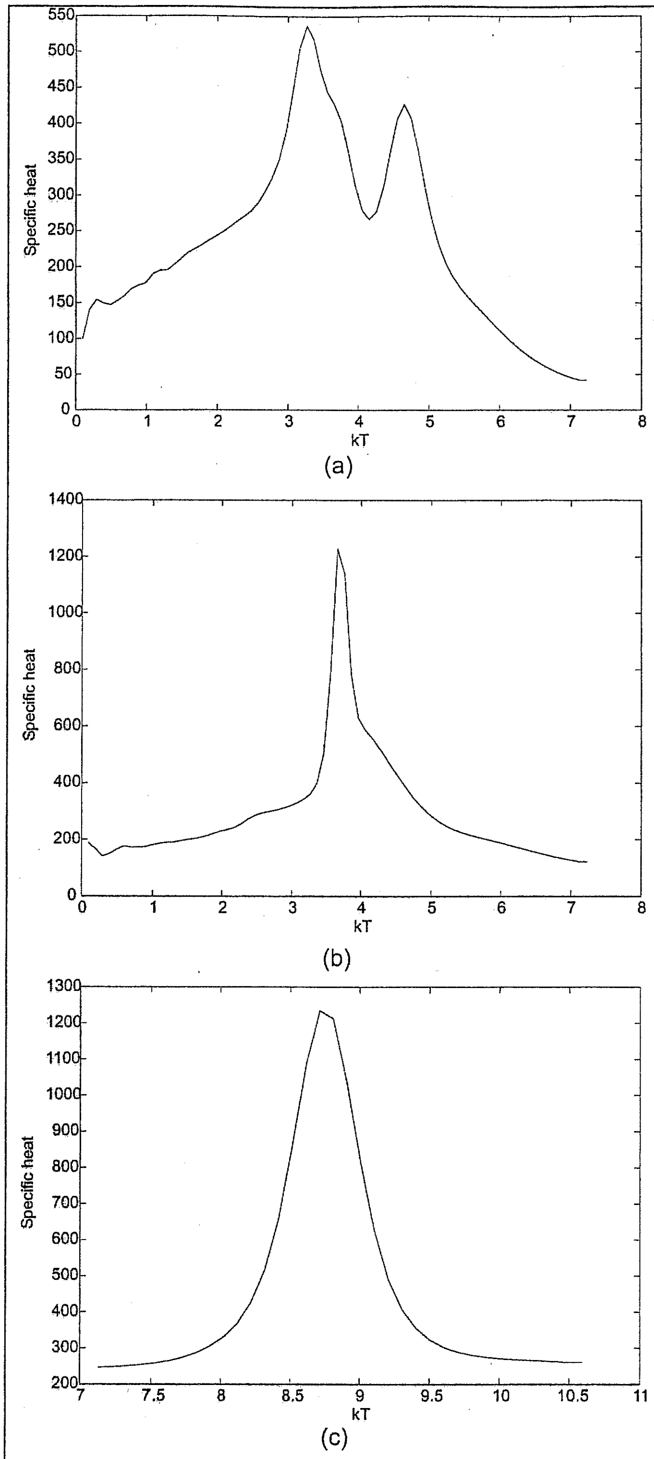


Figure 6. Specific heat of the protein computed with the Swendsen-Ferrenberg method for regime **A** (a), regime **B** (b) and regime **C** (c). The peaks indicate folding transitions. The cooperativity of the folding transition is much higher in (b) and (c) than in (a) where residual structure is present at temperatures comprised between the two peaks (see fig. 5).

### Behavior of residues and contacts at transition

The specific heat of every native contact and the SRSH ($c_i$, see Eq. 1.13) for every amino acid position are measured. Two typical temperature-dependent behaviors of SRSH are presented in fig. 7. As shown by the definition (Eq. 1.10), the contact specific heat also measures the correlation between the formation of the contact (represented by the associated Van-der-Waals energy $V_{ij}$) and the folding of the whole protein (represented by the total Van-der-Waals energy $V_{NB}$).

The formation of contacts linking residues with high SRSH peak takes place, on an average over several folding and unfolding processes, in correspondence with the folding of the whole protein. Thus, it is expected that the more the residue is involved in the folding transition the higher is the peak of the SRSH (Fig. 7). The measurement is performed for the trajectories from the different coupling regimes. In fig 8, the profiles of the SRSH peaks are shown for every amino acid position as computed in the three cases. The figure tells us that their mutual correlation is quite high, 0.87 on average. Another very simple measure of the topological importance of an amino acid is given by its degree of burial defined as the number of contact that the residue forms in the native state conformation or, alternatively, its exposed surface area. However this measure does not contain important topological information like chain connectivity or contact order (the maximal dis-

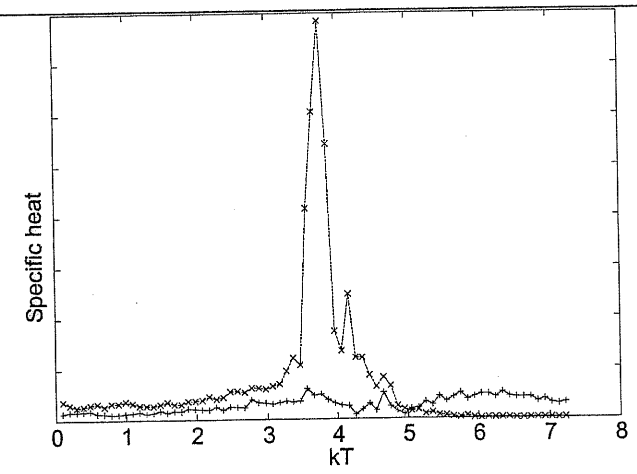


Figure 7. Typical temperature dependent contributions to total specific heat of a relevant amino acid (dashed line) and a scarcely relevant one (continuous line) within regime **B**. The relevant amino acid is strongly involved in the folding transition as showed by the height of the peak in correspondence of the total specific heat peak (fig. 6). For regime A and C the behavior is practically the same.

tance along the sequence of contacting amino acids).

Indeed, the correlation of our results with the degree of burial (i.e. the number of native contacts of each amino acid) is lower than the mutual correlation showed above (0.79 on average). If the degree of burial is computed as the fraction of buried surface of each amino acid then the correlation become even worse.

### *Comparison with data from multiple sequence analysis*

A large amount of immunoglobulin variable domain sequences is available from standard databases. They are easily alignable thanks to their high sequence identity. An aligned database has been created by Kabat and coworkers (45). From there, we obtained a set of aligned VL and VH immunoglobulin sequences. We collected data about the frequency to find a specific amino acid in a specific position of the aligned sequences. We estimated the degree of variability of the amino acids for each position of the alignment as the site-specific entropy:

$$S_i = -\sum_{j=1}^{N} p(i, A_{ij}) \log(p(i, A_{ij})) \qquad \text{Eq. 2.2}$$

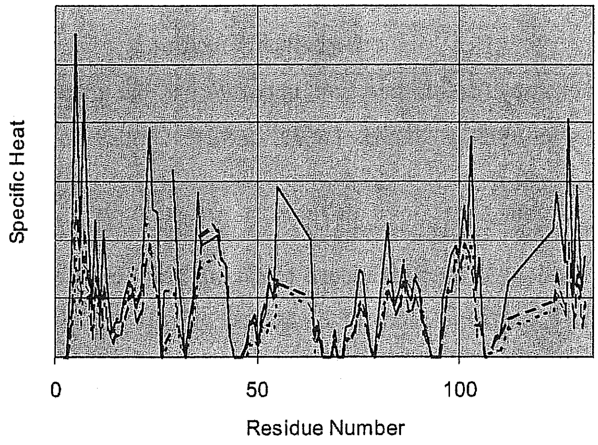


Figure 8. Single residue contributions to specific heat peak for each amino acid ($c_i$). Regime **A**(continuous line), regime **B** (dashed line), regime **C** (dotted line). The high correlation in the profiles (0.87) supports the idea that topology acts similarly on the folding transition notwithstanding the way it is enforced in the model.

*i* is the site index (that runs from the N-terminus to the C-terminus of the Ig domain), *N* is the total number of sequences in the database, $A_{ij}$ is the amino acid type in the *i-th* position of the *j-th* sequence and $p(i,A)$ is the frequency to find the amino acid *A* at *i-th* position of the sequences in the database.

Since these sequences are evolved under the natural selective pressure for stability and functionality, they are supposed to show conserved amino acids in positions that are relevant for their stability (i.e. the cysteine residues are very well conserved). However, these sequences are not tested for intracellular expression. Our per-amino-acid specific heat is

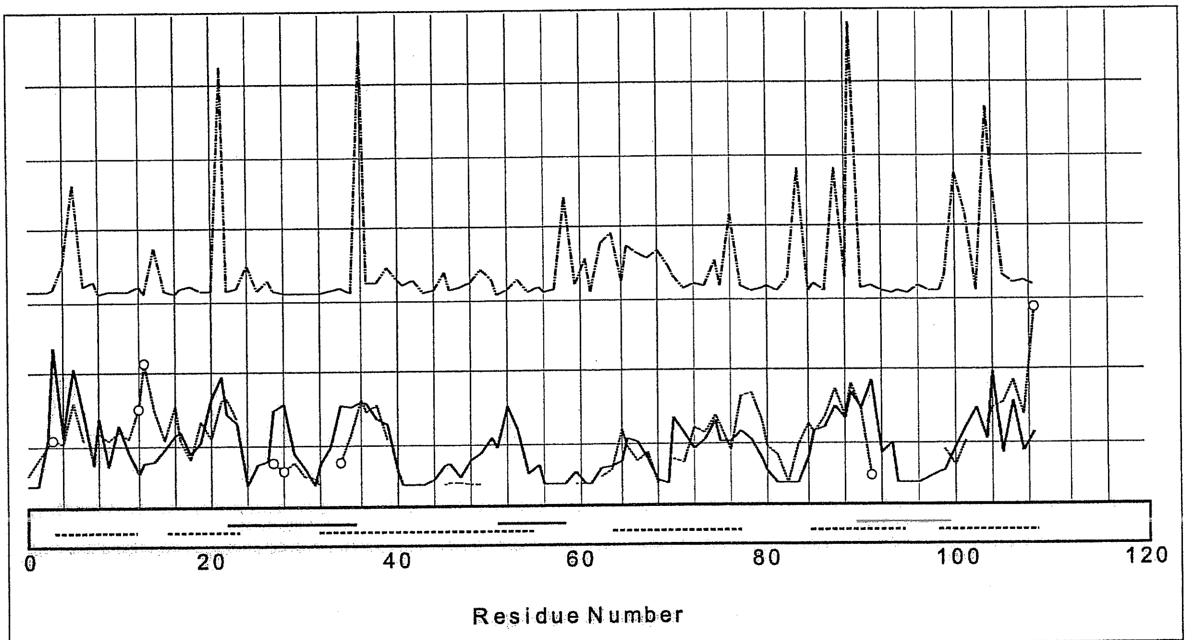


Figure 9. Inverse site entropy (dot-dashed line, upper part of the graph) as computed from the VL domain database (Eq. 2.2), model average SRSHs for Ig variable domain (continuous line) and SRSHs for titin (dashed line), from our calculations. Large values of inverse site entropy correspond to highly conserved amino acids. Large values of SRSH correspond to residues geometrically important for the folding process according to our calculations. The coincidence of the highest peaks of the different profiles (six peaks of inverse entropy out of nine coincide with the highest peaks in SRSH of Ig variable domain and titin) enlightens the very important role the corresponding amino acids play in the stabilization of the native state, according to their topological location. The same or a better match by random guess of the sites with highest conservation would occur with a probability of 0.016%. The graph also shows a good agreement in the SRSH profiles computed on titin and Ig variable domain. The disagreement is localized on eight sites only (white circles). Removing them from the comparison the correlation coefficient between the two sets is 0.79. On the bottom of the graph the black continuous stripes indicates the CDRs of Ig variable domain, while the dotted stripes mark residues in strand conformation.

compared with the variability associated with each position of the alignment and a striking correlation is found between the most relevant sites and the most conserved positions (fig 9). Among the nine most conserved positions of the alignment, six belong to the fifteen highest SRSH peaks (average over model A, B and C). Another one is nearest neighbor of an SRSH peak. Only residues 82 and 98 (Kabat VL numeration) are outside topologically relevant regions. The probability to obtain the same or a better match (6 or more matches) by chance, i.e. picking up randomly two sets of respectively nine and fifteen sites, would be 0.016%[1]. Thus, highly conserved amino acids are located in the topologically relevant positions.

### *Comparison with data from site directed mutagenesis*

These results are compared also with data from the literature in which site directed mutagenesis of the Ig domain was correlated experimentally to the ability of the corresponding Ig domain to fold under reducing conditions (47, 55, 61, 62). These data identify sites that have positive or negative effects on the overall stability of the domain in the absence of the disulphide bond. The two cysteine residues are included to these sites because of their obvious relevance. These thirteen sites of mutations are compared with the thirteen sites showing the highest SRSH peaks (average over model A, B, C). Tab. 1 presents the results of the comparison. Sites are indicated according to Kabat standard numeration and aligned to our sequence. The match between the key sites (i.e. the one with the highest SRSH peak) and the relevant mutations is very good. An estimate of that can be given by measuring the probability to obtain the same match by chance. The match with the experiments (five matches out of thirteen hot sites versus thirteen experimental point mutations on a sequence of 109 amino acids) would be equal or larger than the pre-

[1] If $n_1$ is the number of relevant mutations and $n$ the sequence length, the probability to guess at least $q$ of the $n_1$ by a random choice of $n_2$ sites is:

$$\sum_{m=q}^{\min(n_1,n_2)} \frac{\binom{n_1}{m}}{\binom{n}{n_2}} \binom{n-n_1}{n_2-m} \qquad \text{Eq. 2.3}$$

sent results, only in 3.5% of the possible random choices of the key sites. Furthermore, the use of the number of contacts for guessing the relevant mutations (higher is the number of contacts, more relevant is the mutation) gives a poor match with experimental data (38% of the random picks match better).

```
CDR                                                H1
Seq. num.        1         10        20        30          40         50
                 .....................C........{.....}..............{..
                    | +    -         |+      |      |  ||           -
-------------------------------------------------------------------------
         H2                                             H3
             60        70        80           90          100        110
.abc.............}.................abc.........C..{........}...........
+     -                    -      -             +-|              |  |
=========================================================================
CDR                                        L1
Seq. num.    1         10        20               30          40
             ......................C{....abcdef.....}.................
               |  |                |+        |       - | ||
-------------------------------------------------------------------------
     L2                                          L3
  50           60        70         80         90           100
{........}.............................C.{......ab..}.........a.
 -+                                        +  |               |  |
```

Table 1. The immunoglobulin variable domain sequence VH (upper squares) and VL (lower squares). In the first row of every square, the CDR positions are indicated. In the second and third rows, the standard Kabat sequence numeration. In the forth row the experimental relevant mutations are indicated by a (-) and the calculated topologically relevant sites with a (|). When they overlap, a (+) is used. The positions of the cysteines are indicated with a C on the third row.

## Comparison with data from titin

Titin is a protein involved in muscular elasticity; one of its domains shares the generic Ig topology. However, it differs from the Ig variable domain topology for several aspects; it is 89 residue long, so 20% shorter than the Ig variable domain. It also lacks an anti-parallel beta strand present in the Ig variable domain between residue 48 and 58 (standard Kabat VL numeration). This means that the quality of the alignment is worse than for VH-VL alignments. However it represents a useful comparison to test the degree of variability of the topological measures developed in this thesis on proteins belonging to a similar folding family. SRSH have been computed for titin (the Ig like domain pdbid 1tit)(63) following the same procedure adopted for the Ig variable domain (Model B has been used for the molecular dynamics). The results have been compared after alignment of titin with Ig variable domain maximizing Eq. 2.1.

The SRSH profiles for the aligned residues of titin have been reported in fig 9. Since the alignment of the two structures is not very good, in order to get a more sensible comparison between the two SRSH profiles, data have been smoothed averaging the value $c_i$ at position $i$ with the half weighted neighboring $c$'s. This lead to a correlation of 0.56 instead of the 0.47 as one would obtain without the smoothing procedure. However, the profiles clearly disagree in only 8 localized positions. These can be identified as the sites where the absolute value of the difference in SRSH overcomes a fixed threshold. By removing them the correlation coefficient between the two profiles becomes 0.79. The eight sites showing low agreement are localized in the top and bottom loops of the structures and are organized in three subgroups. One subgroup contains three residues and is localized on the non-CDR loops (see fig 10 as a reference); on titin, these residues interact with three C-terminal residues that are not present in the Ig variable domain; this perturbation may be the cause of this disagreement. The second subgroup is composed of three contacting residues two of which belong to the CDR1. This region is poorly aligned with very high rmsd (6.0 Å versus 2.8 Å of the whole 83 aligned residues); that might be the reason for the discrepancy in SRSH. The last subgroup is composed of two contacting residues; in Ig variable domain, one of the residues is also in contact with the strand that is missing in titin, the other one is located at the beginning (N-ter) of CDR3 which is 8 residue long in Ig variable domain while its titin counterpart is only 2 residue long.

The two SRSH profiles from titin and Ig variable domain agree in many regions. In particular they agree where the conservation in VL sequences is the highest. Setting the same threshold used to identify the fifteen highest SRSH peaks in the comparison with VL sequence entropy (see above), sixteen SRSH peaks are identified on titin structure. Seven of them coincide with Ig variable domain SRSH peaks and six of these correspond to VL sequence entropy peaks too. A comparison with titin sequence conservation has less statistical significance than with VL domain sequence conservation because only tenths of similar sequences have been retrieved using standard alignment algorithms and sequence databases, that are very few with respect to the thousands of sequences in the Kabat database used to compile the VL inverse entropy profiles.

### *Discussion*

The comparison between the phase diagrams obtained in the different coupling regimes shows that activation of angular and dihedral energy terms brings to an amplification of the cooperative behavior of the system. Dihedral and angular interactions are local multibody interactions. In this model their presence determines the disappearance of intermediate states between the folded and unfolded conformation and the enhancement of an all-or-none behavior in which all the residues cooperatively promote one of the two states. Indeed the increase in the stiffness of the chain due to the angular and dihedral terms determines the narrowing of the allowed conformational space and the exclusion of the region occupied by the intermediate state in model **A**.

Experimentally measured thermal denaturation of ScFv domains (47) is in accordance with results from regime **B** and **C** showing a two state transition from the folded to the unfolded conformation.

Furthermore, in this model, stability increases only if both dihedral and angular interactions are present. This directly descends by the fact that both torsional and angular degrees of freedom have to be fixed to uniquely determine the conformation of a polymer chain. That is in agreement with the increase in denaturation resistance due to secondary structure optimization experimentally found(64).

Despite of the diverse thermodynamic behavior of the system to heating and cooling in the three cases, the agreement in SRSH profiles (Fig. 8) is in accordance with the idea that the topology of the native state plays the same role in the three regimes. Furthermore, the weak correlation with the number of contacts per amino acid indicates that the summation on the native contacts of the SCSHs does not yield trivial results. The per-amino-acid number of contacts represents, of course, an important part of the topological features of the proteins. Nevertheless, other topological features, like chain connectivity, play a determinant role.

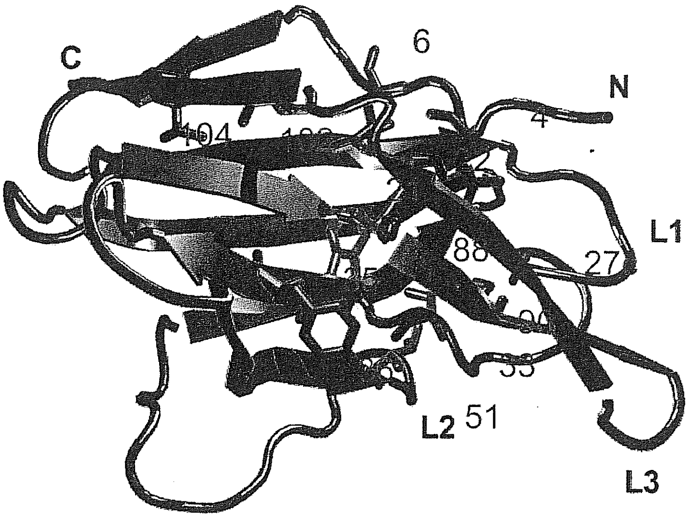


Figure 10. A cartoon representing the structure of the Ig variable domain. The residues with large SRSH are shown in red licorice and numbered according to standard Kabat numeration for VL domain. The N and C termina are indicated with the corresponding letter. Complementary determining regions (CDR) are indicated as L1, L2 and L3.

That indication is confirmed by the fact that the number of contacts per amino acid is not enough to guess the relevance of the amino acids in the folding process, as shown in the previous section. On the contrary, our topology-based estimate is far more reliable and let a non-casual discrimination of important residues.

Finally, the surprising match of the single residue specific heat profiles with the degree of variability of the amino acids on generic Ig variable domain sequences is a strong indication that the topology of the native state has a very important role in the folding process of this type of proteins, too. The non-sequence-specific nature of this evidence and the extent of its validity to slightly different topologies are confirmed by the match between the conserved sites on VL sequences and the high SRSH sites on titin. The role of topology on the degree of conservation of amino acids has been addressed by different methods and they are reviewed in (65).

The amino acids with high SRSH are supposed to play a relevant role in the folding process due to geometrical reasons; mutations occurring at those sites might affect the protein stability and/or its folding mechanism depending on the strength of the perturbation. The method used here allows identifying them in a general and sequence-independent way. The results are valid for proteins sharing the same Ig variable domain topology (without the covalent disulphide bond). Moreover the agreement of SRSH profiles of titin and Ig variable domain underscores the robustness of the results with respect to topology changes.

However the relevance of a mutation on the folding behavior of a given sequence would generally depend both on the geometric position in which it occurs and on the detailed chemical changes it produces. The approach used here only addresses the first issue and takes into account the local detailed chemistry underlying the positive and negative interactions only in an average way. Indeed, some mutations can be very effective for a particular Ig sequence and irrelevant for a different one because of the different chemical context in which they occur; thus, it is not expected that all relevant mutations lie on high SRSH sites and vice versa.

The identified sites represent good targets for the design of optimized intracellular antibodies. Both a semi-rational method and a selection-based approach (see above) would take advantage of a reduction in the number of sites on which the optimization has to be performed. In the first case the search for stabilizing interactions would be focused on these sites only. In the second case, libraries could be built up with antibody sequences in which interactions involving these sites have been optimized, letting a higher yield of intrabody domains than standard libraries do.

### Selected intrabody sequences show relevant peculiarities

A generic *in vivo* selection procedure for intracellular antibodies that function inside cells (ITT) has been recently proved in principle (48) and then developed (15). Thanks to that it has been possible to demonstrate that antibodies suitable for intracellular selection are present in natural repertoires and that these antibodies can be readily isolated with this new technology in a rapid and efficient way that avoids the need of any rational mutation strategy (66-68) or of more complex molecular evolution approaches (55, 69). In this section of the chapter we present an analysis that we performed on a set of intracellular antibody sequences selected through ITT by our collaborators (15). The analysis shows that intrabodies selected by ITT are characterized by a common signature of conserved amino acids, and by a high homology to a consensus sequence in the Kabat database, showing that ITT naturally performs in a robust, efficient and generic way a "biological consensus sequence selection", whose outcome is a set of validated intrabodies. This signature will allow designing improved antibody libraries biased towards intrabodies.

#### *Brief Description of ITT*

The Intrabody Trap Technology is a procedure that allows selecting *in vivo* intracellular antibodies with a determined specificity from a generic library of antibody fragments. A scheme of the selection mechanism is described in fig. 11. Briefly, ScFv from a library are first roughly selected as in vitro binders of the target antigen. Then, they and their antigens are cotransfected in yeast cells following the two-hybrid protocol (fig. 12) (70). This allows rescuing only cells where the intrabodies bind the antigens. Finally, intrabody genes from survived cells are isolated and cloned. They are able to preserve their binding activity in the critical intracellular environment.

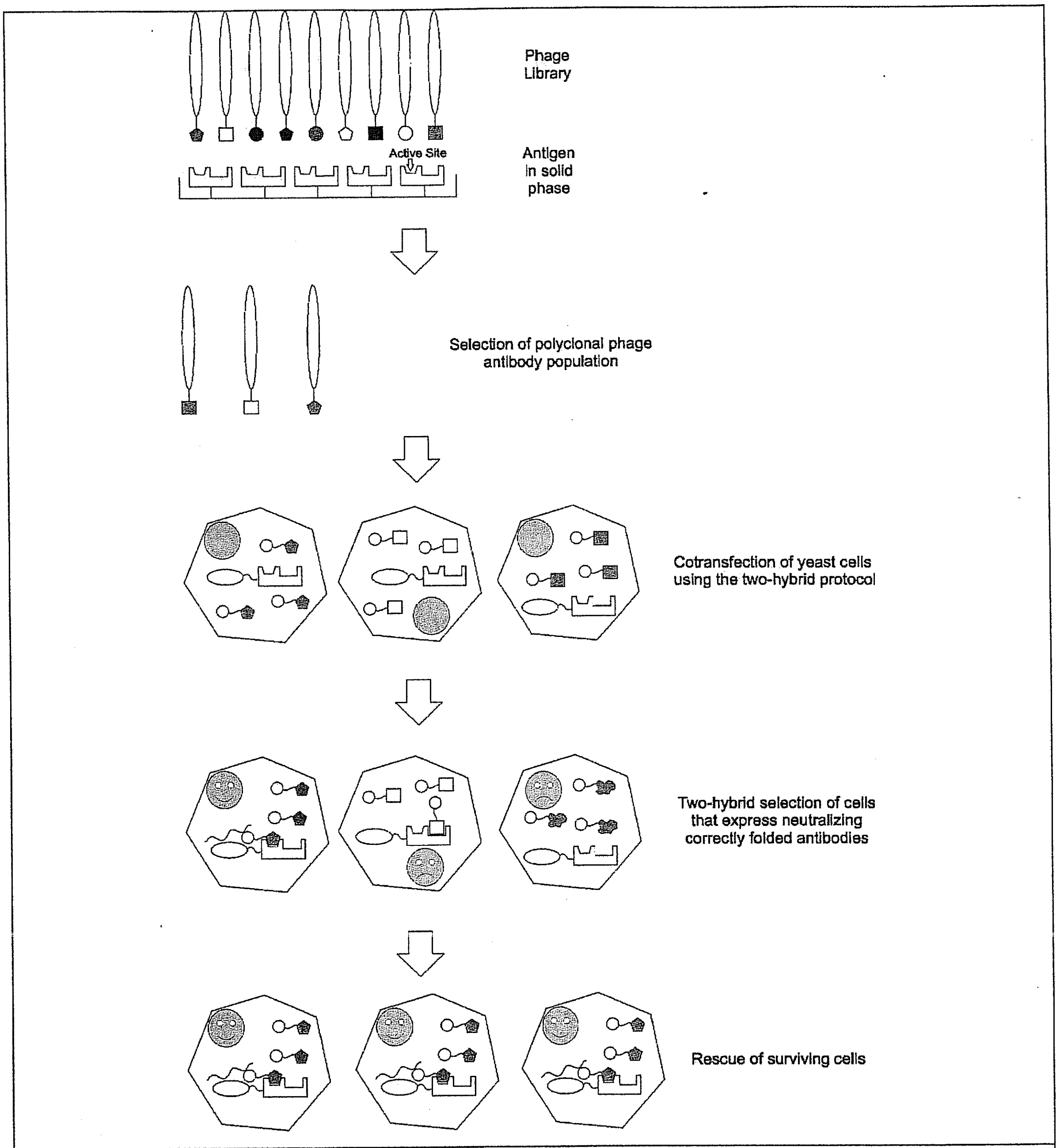

Phage
Library
Active Site
Antigen
in solid
phase
Selection of polyclonal phage
antibody population
Cotransfection of yeast cells
using the two-hybrid protocol
Two-hybrid selection of cells
that express neutralizing
correctly folded antibodies
Rescue of surviving cells

Figure 11. A schematic view of the Intrabody Trap Technology (ITT). It involves a combination of the yeast two-hybrid technology and the use of a phage-display library for the selection of antibodies. The preselection step by using phage display technology allows increasing the percentage of putative candidate scFvs in the input library. The yeast two-hybrid approach is adapted to isolate scFv-antigen interaction pairs under condition of intracellular expression. After the *in vivo* selection, positive clones are isolated and scFvs are used for *in vitro* and *in vivo* applications.

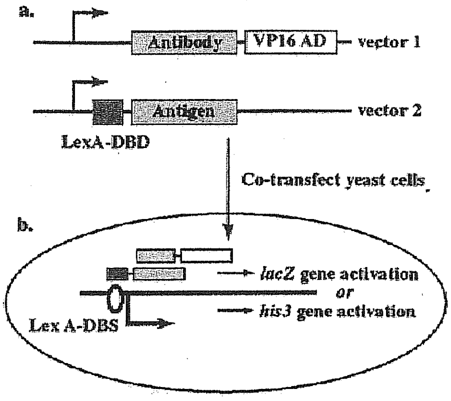


Figure 12. Diagram of antibody-antigen *in vivo* interaction assay. The two-hybrid system of Fields and Song (70) was adapted to detect antibody-antigen interaction *in vivo*.(a) Yeast expression constructs were prepared encoding either (i) an antibody fragment, in the form of scFv, linked to the VP16 transcriptional activation domain (AD) or (ii) the LexA DNA binding domain (DBD) linked to a target antigen sequence. *(*b) These constructs are cotransfected into yeast cells unable to synthesis histidine and carrying either the histidine *(his*) gene or the *lacZ* gene controlled by a minimal transcription promoter with a LexA DNA binding site (DBS). If antibody-antigen interaction occurs *in vivo*, the resulting complex can bind to the LexA DBS upstream of *his* or *lacZ* genes and transcription of these genes occurs (the VP16 activation domain is thus brought close to the DNA transcription start site and can recruit accessory factors needed for transcription). The transcriptional activation of the *his* gene facilitates growth of yeast on media lacking histidine and activation of the *lacZ* gene produces β-gal, which can be assayed with 5-bro-mo- 4-chloro-3-indolyl b-D-galactoside to yield blue yeast colonies. Neither feature of the transfected yeast will occur if the antibody fragment does not function inside cells.

## *Analysis of sequences*

In order to gain insight into the properties of the subset of antibodies that represent potential intrabodies, sequence analysis of intrabodies selected with ITT (ITT set) and of a group of scFv fragments from the input library (control set) was performed (25 and 16 sequences respectively).

All the VH domains in the ITT group are classified (71, 72) in subgroup III while VL domains are kappa I (19 sequences) or kappa IV (6 sequences). Also in the control group the large majority of VH sequences are from subgroup III (only one sequence from subgroup II was found). The VL sequences in the control group are part kappa I (10 sequences) and part lambda (6 sequences, mostly from subgroup IV).

The average homology between individual members of the control set of sequences was 69% for VH and 59% for VL domains. The average homology within the ITT set was 86% for the VH and 65% for the VL domains. Within the ITT set, 76 amino acid residues in VH (6 from CDR 1, 2 from CDR 2 none from CDR 3) and 45 in the VL domains are absolutely conserved. These conserved residues define an intrabody consensus.

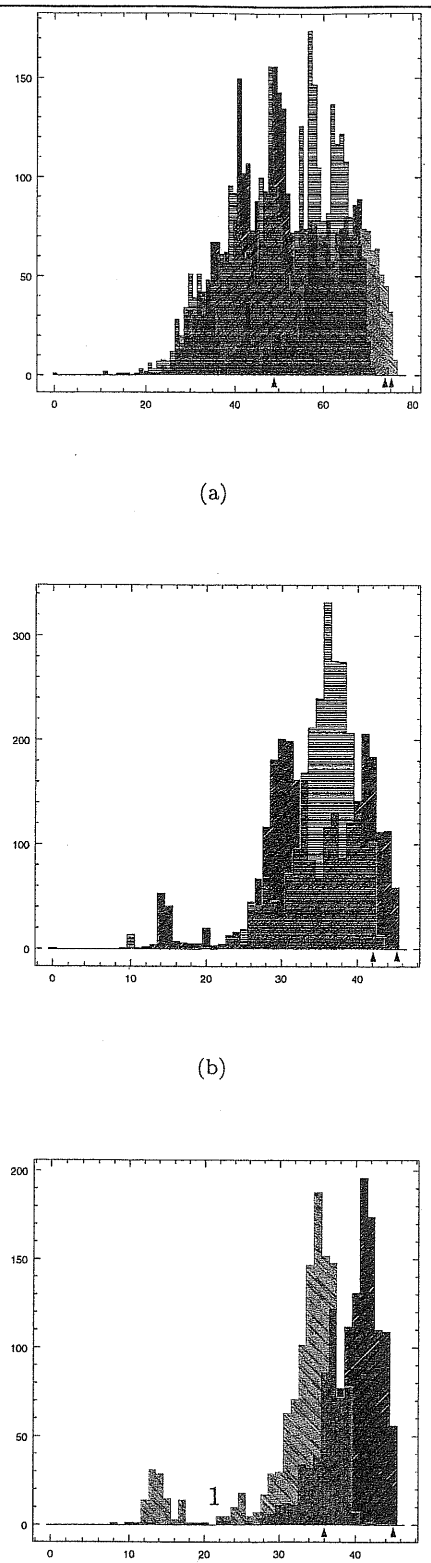
Antibody sequences in the ITT and in the control sets were compared to consensus sequences, compiled from appropriate subsets of the Kabat database. Consensus sequences are defined as the antibody sequences in which each position is occupied by the most frequent amino acid at that position in the corresponding subset of the database.

The following subsets of sequences from the Kabat database were considered: set 1, composed of the 3319 human VH sequences, set 2, composed of the 3353 murine VH sequences, set 3, 1872 sequences from the human VH3 subgroup. For the light chain, set 4 composed of the 2731 human VL sequences, set 5 composed of the 2518 murine VL sequences, set 6, composed of the 1330 human VL kappa sequences, set 7, composed of the 1265 human VL lambda sequences.

Comparison with set 1 showed that each, but two, of the conserved 76 amino acids in the ITT selected VH set is the most frequently found at this position, in the Kabat set 1. The two positions that do not follow the consensus are H67 and H73, nevertheless they are occupied by residues (Phe 67 and Asn 73) whose occurrence probability is only slightly lower than the highest one (Val in H67 and Thr in H73). In comparison with set 2, 27 amino acids out of the 76 conserved amino acids in the ITT set are not the most frequently found in the set 2 consensus. Comparison with set 3 showed that only Gln in H1 is not as frequently found as Glu at this position, while all the other residues are the most frequently found at each position. For the light chain sequences, comparison with set 4 showed that all the 45 conserved residues in the ITT selected VL are the most frequently found, with no exception.

Comparing to set 5, only 3 residues out of 45 do not follow the consensus rule, while all the 45 conserved residues match the most frequently found in set 6. Finally, 9 residues out of the 45 conserved ones do not match those most frequently found in set 7.

We then analyzed the homology distribution of sequences from the different sets, with respect to their corresponding consensus sequence in the Kabat database. The analysis



(a)

(b)

(c)

Figure 13. Distribution of the degree of homology of the sequences from different subsets of the Kabat database, with respect to the corresponding consensus sequence. The analysis was restricted to the amino acid positions corresponding to the 76 (part a) and 45 (parts b) and c) conserved residues for the heavy and the light chain respectively in the ITT set. The abscissa reports the number of amino acid residues homologous to the consensus sequence of the corresponding dataset: a) black, human VH (set 1), blue, mouse VH (set 2), red human VH3 (set 3); b) black, human VL (set 4), blue, mouse VL (set 5); c) black, human V kappa (set 6), red, human V lambda (set 7). The colored arrows below the abscissa indicate the degree of homology to the consensus sequence of the intrabody consensus sequences, for each subset of the database (correspondingly colored).

was restricted to the amino acid positions corresponding to the 76 and 45 conserved residues for the heavy and the light chain respectively in the ITT set. The histograms in Figure 13 show that in some cases the overall distribution of homologies with the consensus sequence is bimodal, while in others is unimodal. In the case of human VH sequences (black histogram in Fig. 13 a), the bimodality reflects the contribution of sequences from the VH3 family. In the case of human VL sequences, the two peaks of the distribution are contributed by VK and V λ (Fig. 13 b).

The degree of homology of the intrabody consensus sequence with the consensus sequence for each subset of the database is reported in figure 13 as the colored arrows below the abscissa. Figure 13 clearly shows that the intrabody sequences map in the tail of the corresponding distributions, in the high homology-to-consensus region, with the only exception of comparison to mouse VH and human VL lambda (see below). In this high homology region, the distributions are not densely populated, and therefore a degree of homology with the consensus sequences as high as that found for our intrabody sequences is not a common feature in the Kabat database and in its subsets. This indicates that sequences with high homology with the consensus are not very frequently found in the database while they are in the ITT selected antibodies. The comparison of the intrabody consensus sequence with the mouse VH and human lambda VL distributions is less good, consistent with the species difference (mouse versus human) and with the enrichment in human VK in the intrabody set.

The results of this study have been confirmed by the analysis of two different sets of sequences obtained with ITT by other different research groups that very kindly sent us for analysis.

The first group of six sequences was supplied by the group of Terry Rabbits at MRC. These antibody ScFv have been selected against BCR antigen. The initial phage display library was the same as before. The VH segment of these sequences has 80 conserved residues. 66 of them overlap with the set obtained before both in position and in amino acid type. The position conserved in a set and not in the other are 25, six of them lie in the CDR region and are probably related to the different antigen for the two sets. However the 66 residues conserved in both set are also the most frequently found in natural repertoire.

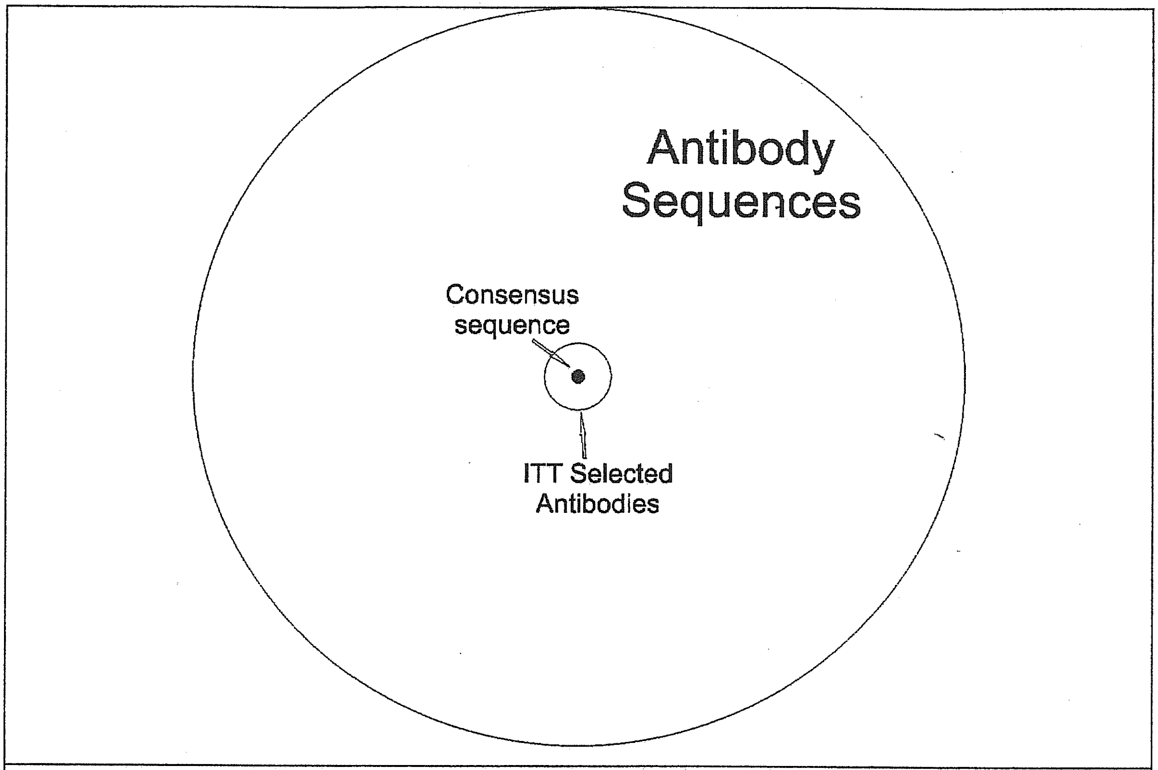


Figure 14 The Intrabody Trap Technology operates as a sort of filter on the natural repertoire allowing selecting antibodies that share characteristics of correct folding and binding to antigen. The sequences obtained from this filtration process are very close to the consensus sequence of the corresponding natural repertoire.

Other data collected at SISSA led to very similar results. In this case the initial library was completely different (mouse spleen). Nevertheless, for the VH part of the sequences there are 63 conserved residues that coincide with the most frequently found in those positions in kabat mouse antibody sequences.

This analysis allows concluding that ITT leads to the selection of a population of antibodies that appear to be present at low frequency in antibody repertoires. ITT selected antibodies have the remarkable property of having a large subset of residues much closer to the consensus sequence than the average antibody in the library or the repertoire (Fig 14).

## Conclusions

In this chapter we presented the work done on the issue of intracellular expression of antibodies and their folding ability. Two approaches have been followed. The first one concerned the exploitation of the topological characteristics of the Ig variable domain by means of molecular dynamics simulations of a topology based model protein. Comparison of the results obtained by this analysis with experimental mutations that improve intracellular expression, enlightened the importance of the topology in the folding of this kind of proteins and its use in the determination of crucial mutational sites. These findings are strengthened by the results obtained on a protein with a similar topology, besides comparison with data on amino acid conservation in the Ig variable domain family of proteins.

The second approach has been pursued thanks to the availability of ScFv sequences selected for intracellular expression. It involved the analysis of these sequences and led to the discovery of relevant shared characteristics. These comprise the presence of a large set of conserved amino acid concentrated in the VH subunit and the closeness of this set to the consensus sequence computed from the natural repertoire corresponding to the input library. These findings further validate the experimental technique used to obtain them and draw a way to the construction of an intrabody rich ScFv library that will be of great importance in fast throughput functional genomics .

The experimental outcome of the second approach can also be used to validate the second approach and set its limits. The topological analysis allowed ranking amino acids positions of the Ig variable domain structure according to their SRSH that we have previously defined and explained (see "Chapter 1"). While the first ten highest-ranking amino acids positions mostly belong to the set of conserved amino acids in the intrabody sequences, this does not happen for the amino acid positions that rank lower than those. From this data seems to emerge that intracellular intrabodies not only need to be optimized in topologically relevant positions, but, after that is achieved, their stability depends on a widespread cumulative interaction between amino acids belonging to the large conserved set identified through the second approach. This set seems not to be strongly related to the topology of the protein. The more amino acids in the set belong to the consensus sequence, the more stable is the resulting antibody. This hypothesis has also been supported by recent work in this direction (73), where a generic antibody sequence was stepwise mutated towards the consensus sequence and its stability was evaluated as being increasingly higher than wild type. As an intriguing perspective, we would like to figure out what is the selective pressure that led to the identification of the conserved set in intrabodies and what are the structural characteristics of this set.

# Chapter 3 - Prions

The present chapter is organized as follows: in the introduction, the biological and biochemical aspects of prion disease and prion protein are presented, current theories on the cause of the disease are exposed and our objectives are declared; in the second part, the folding process of prion and a related protein are investigated through minimalist models that are based on the topology of the native state. Finally results are discussed.

## *Introduction*

Prions are pathogenic agents substantially different from bacteria and viruses. They cause transmissible spongiform encephalopathies (TSE). TSE are a group of neurodegenerative diseases that affect both humans and animals. Most of them (80%) arise sporadically; some (19%) are genetically induced, and a small proportion (1%) can be transmitted between mammals by inoculation with, or dietary exposure to, infected tissues. They do not elicit any immunological response and the symptoms (progressive dementia in humans, ataxia in bovines and sheeps) arise very late after infection. There are descriptions of cases of prion diseases in humans dated 1930 (74, 75); the analogous form in sheep (scrapie disease) was already known since 18th century. Only in the late fifties the common origin of several strains of the diseases was understood (76-78), mainly thanks to microscopic inspection of brain tissue (that usually presents spongiform degeneration and astrocytic gliosis, fig. 1) and to experiments of transmission of the disease after inoculation of infected tissues. Because of the very special character of prion disorders only few years ago, thanks also to the work of the Nobel laureate S.E. Prusiner, it has been possible to partially clarify and explain some important features of them, even if different theories have been also elaborated (79).

The "proteinaceous" nature of the infectivity factor in prion diseases was clarified by experiments on radiation resistance of the scrapie agent (80) completed by Prusiner's experiments (81) on the differential resistance to proteolitic and nucleic acid alteration factors. The only procedures that were able to reduce the infectivity of the scrapie inocula consisted in hydrolization or modifications of proteins while treatments oriented to nucleic acid alterations had no effects. Later, tissues enriched in the infectivity agent were obtained exploiting its partial protease resistance (82) and a part of the sequence of the protein was determined (27), leading to the discovery of the coding gene PrP (83). The absence of differences between the gene in normal and infected animals was a clue for the hypothesis that a posttranslational modification in the protein coded by PrP characterizes the disease. Then, the idea that prion protein can exist in two different conformations has been demonstrated by several experiments. When the secondary structures of the PrP isoforms were compared by optical spectroscopy, they were found to be markedly different. Fourier-transform infra-red (FTIR) and circular dichroism (CD) studies showed that $PrP^C$ (the normal non infective *c*ellular form) contains about 40% α-helix and little β-sheet, whereas $PrP^{Sc}$ (from *sc*rapie) is composed of about 30% α-helix and 45% β-sheet (84, 85). The two proteins, however, shares the same sequence being coded by the same gene. Thus, the discovery of prions put into question the consolidated idea that each protein sequence codes for a unique biologically active conformation (1).

The actual knowledge of the problem is the following: most of the prion diseases involve the modification of the protein conformation expressed by the PrP gene universally present in mammalian and avian cells. Infective prions are composed of an isoform of the normal cellular protein $PrP^c$. This infective form is denoted $PrP^{sc}$. It is believed that normal $PrP^c$ undergoes a transition to the $PrP^{sc}$ conformer after infection by other $PrP^{sc}$

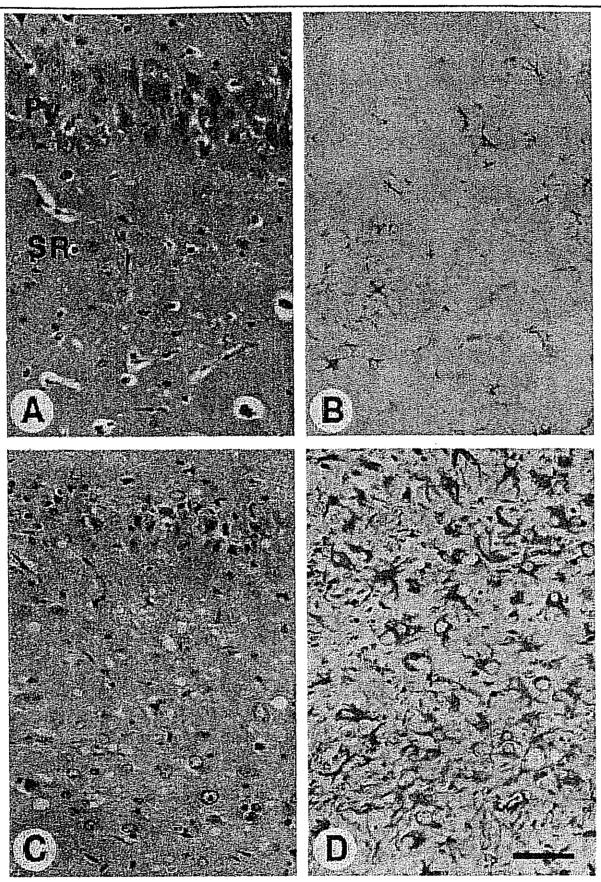


Figure 1 Normal tissue (A and B) and tissue affected by prion disease degeneration (C and D) detected with two different staining techniques (hematoxylin and eosin for A and C, immunostaining for GFAP in B and D). In C spongiform degeneration is present with 10-30µm vacuoles. In D numerous reactive astrocytes are identified. Tissues are obtained from mouse hippocampus.

molecules. This is confirmed by experiments showing that prion brain plaques in bovines injected with sheep infective scrapie inocula are composed of bovine prions and not by sheep prions.

On the other hand, up to now, 20 single point mutations in the PrP gene in Humans have been associated to the genetic forms of the prion disease. The first mutation to be identified was the P102L mutation in patients affected by the GSS (a familial form of neurodegenerative disease)(86-88). Transgenic mice expressing the same mutated PrP gene spontaneously develop a prion disease. Other familial forms of the disease have always been put in correspondence with a mutation in the PrP gene. Their genetic origin has been clearly established for some of them (89-93).

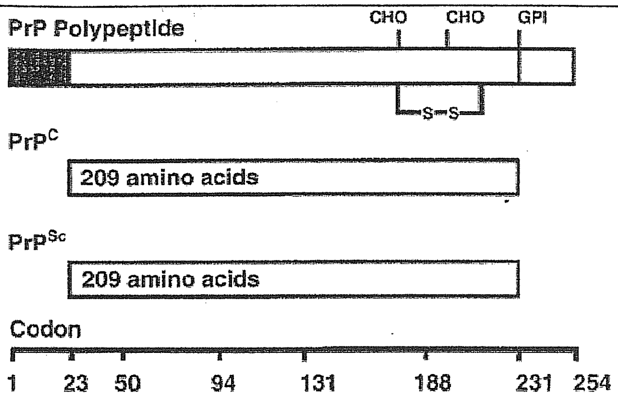


Figure 2 PrP consists of 254 amino acids (in Syrian hamster). Attached carbohydrate (CHO) and a glycosyl-phosphatidylinositol (GPI) anchor are indicated. The disulphide bridge is also indicated (S-S). After processing of the N and C termini, both PrP$^C$ and PrP$^{Sc}$ consist of 209 residues.

The normal cellular prion protein (PrP$^C$) is a glycoprotein (Prio_Human; P04156) containing 209 amino acids. It also contains two N-linked glycosylation sites and a glycosyl-phosphatidylinositol (GPI) anchor (fig. 2). It is normally attached to the cell membrane by the (GPI) anchor (94). The functions of the prion protein are not yet clear, although possible roles in copper transport, synaptic function, circadian rhythm, promoting genetic diversity, and signal transduction have been suggested (95-99).

Up to now, no x-ray structure of any form of the Prion protein has been obtained. However several groups contributed NMR structures of different variants of the PrP transcript (100-104). In all these structures only the C-terminal tail (about 110 residues) of the PrP protein and mutants results defined. Given the high helical content of these structures they should represent the protein in the cellular conformation (even if one of this protein fragment revealed to be lethal if injected (105)). There are evidences that the N-terminal tail of the protein, whose total length is about 210 residues, has not a defined structure, still in some of the forms (Syrian Hamster prion protein) an interaction with the C-terminal part was detected (106). Furthermore, Cu$^{2+}$ seems to help the formation of secondary structures in the N-teminal part of the protein.

The PrP$^c$ structure is characterized by two C-terminal helices connected by a disulphide bond between two cysteine residues and an anti-parallel beta-sheet at the N-terminus of the structured part. Between the two strands that form the beta, resides a third helix found in the protein (fig. 3).

The only known features of the PrP$^{sc}$ form are its beta-rich content and the propensity to form oligomers and amyloid aggregates (84). Unfortunately, its structure has not yet been solved.

The models that have been formulated up to now suggest that the conformational change that leads from PrP$^{c}$ to PrP$^{sc}$ should occur in the N-terminal part of the PrP$^{c}$ concentrated between residues 90-145 (perhaps 175) (107); indeed, it has been verified that the disulphide bridge present in the C-terminal part and involving the two helices is needed and untouched during PrP$^{sc}$ formation (108, 109). Studies conducted on transgenic animals seem to confirm that PrP$^{sc}$ act as a template for normal PrP$^{c}$ to refold in the PrP$^{sc}$ form. It has also been verified that the conformational change occurs after PrP$^{c}$ reaches the cell surface (110, 111) and takes place in calveolae-like domains (112-115).

Chimeric forms of prions have been studied in order to determine the binding site of PrP$^{c}$ on PrP$^{sc}$ (116, 117). Both the regions 95-170 and 180-205 have been identified as the binding determinant.

Thermodynamic and Kinetic characterization of PrP$^{c}$→PrP$^{sc}$ interconversion has been studied recently by Prusiner and coworkers (118) through folding and unfolding experiments. The folding and the unfolding are triggered by urea concentration and evaluated by far-UV CD at 220nm, that allows monitoring the secondary structure content of the protein. The effects of incubation in environments with different salt concentrations and the prion concentration dependence of the process are examined. In low salt concentration the α-rich PrP form (presumably cellular PrP$^{c}$) folds and unfolds with a hysteresis. This is interpreted as a mark of a non-two-state behavior; the protein, after being unfolded, does not go back exactly to the initial folded state; some of the molecules assume a different state. When PrP is incubated in high salt concentration, the refolding curve changes as the incubation period increases and the character of the folded species slowly becomes more and more β-like. The conversion to the β-rich conformation is also accompanied by the formation of oligomers (8-mer prevalently as revealed by size-exclusion chromatography). That is interpreted as the signature of the PrP$^{sc}$ form. The speed of the conversion increases as the initial concentration of prion protein is increased.

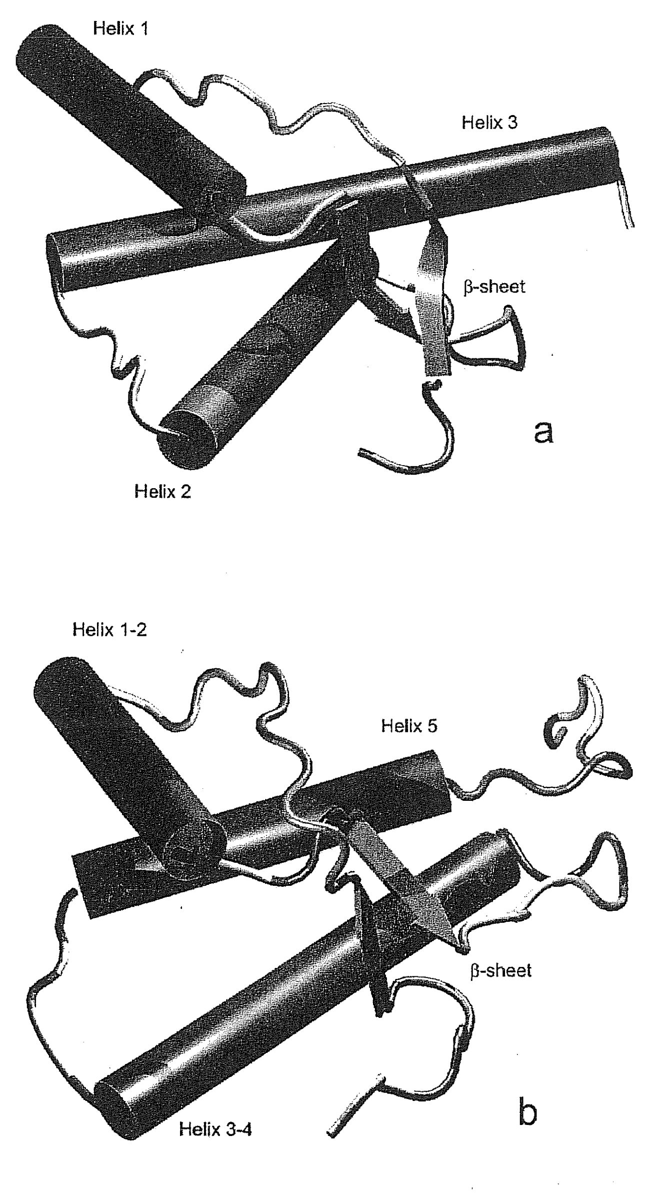


Figure 3 Proteins showing prion characteristic topology; helices are colored in magenta, beta-strand in yellow and turns in cyan. (A) Human prion protein (pdbid 1qlx). (B) Doppel protein (pdbid 1I17 model 1). Secondary structure elements are named according to the pdb file.

According to the model proposed in (118), the conversion between the α-rich and β-rich form is obtained through the complete unfolding of the protein. The low energy barrier separating the unfolded and the α-rich state accounts for the fast emergence of the α-rich conformation during refolding experiments. On the other hand the slow transition to the β-rich oligomeric state is explained assuming a higher thermodynamic stability and a higher free energy barrier to the unfolded state.

On the basis of this model they extract from the data a quantitative estimate of the relative depth of the free energy minima of the different species and of the free energy barrier

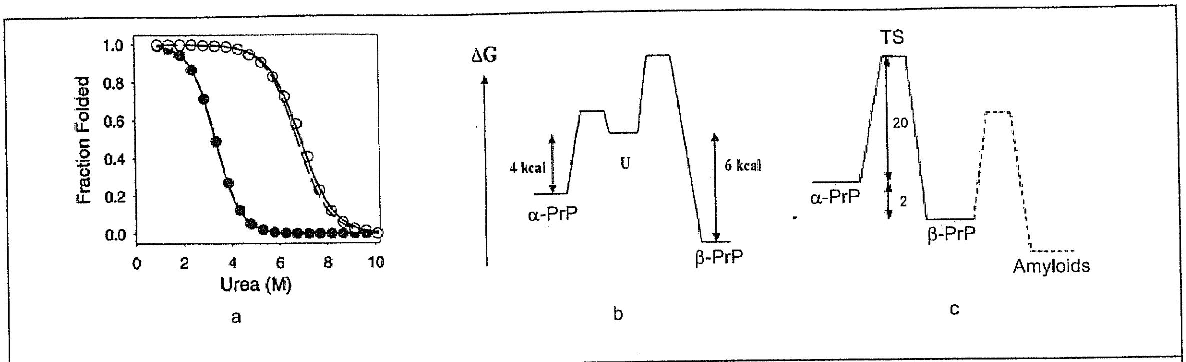


Figure 4 Thermodynamic model for the folding of PrP protein in the α-rich and β-rich form. (a) The curve of denaturation for the α-rich (unfolding branch, filled circles) and β-rich form (open circles). (b) Free energy scheme for the two competing transitions. The stability of the β-rich form is higher than the α-rich form, however a high free energy barrier prevents a fast conversion to the β-rich form. The system fold fast to the α-rich and then, it slowly converts to the more stable β-rich form passing through an unfolded conformation. (c) The height of the free energy barrier in the experimental condition has been estimated according to the conversion time and Arrhenius law. The barrier in physiological conditions should be higher (~35 Kcal/mol). The presence of very stable amyloid structures has also been detected.

between the α-rich conformer and the β-rich conformer. Extrapolating those numbers to physiological conditions leads to estimate the height of the free energy barrier as ~35-40 Kcal/mol high. This big number would explain the very slow process of incubation of the disease. According to this study, the cellular form of the prion protein represents a kinetic trap of the folding process. The height of the free energy barrier between alpha and beta conformer is influenced both by the unfoding characteristics of the alpha conformer and by the speed of the oligomerization process. The presence of mutations that favors the arousal of the disease (i.e. the interconversion between normal and scrapie prion form) should be, thus, linked either to the lowering of the free energy associated to the transition state between the two isoforms (maybe involving the stabilization of the binding between $PrP^c$ and $PrP^{sc}$ that is an obligatory step to oligomerization) or to the destabilization of the cellular prion protein.

It has been verified that transgenic mice lacking the PrP gene ($PrnP^{0/0}$) are not susceptible to infection by prion injection. The experiments on these animals were primarily meant to understand the function of $PrP^c$ normal form. However, part of the $PrPn^{0/0}$ lines did not show any phenotypical impairment. Another part showed late-onset ataxia and a neurodegeneration completely different from usual prion disease (Purkinje-cell death). Later, the up-regulation of a prion homologous gene (that was not expressed in normal subjects) was discovered in these transgenic mice. The product of this gene was named doppel protein (Dpl). This protein shares 25% sequence identity, on average, with the prion proteins and has a topology very similar to $PrP^c$ (fig. 3b): both have three helices and a beta sheet, with two helices being at the C-terminal; tertiary structure contacts involve inter C-terminal helices contacts and contacts between beta-sheet and C-teminal helices. However Dpl does not show the conversion to a scrapie conformation upon prion infection of mice and is not associated to the arousal of the prion disease. Several hypotheses have been put forward to explain the different behavior of Dpl compared to prions regarding the absence of misfolded configurations. In particular, the attention has been focused on the lack in Dpl of the palindrome sequence in the flexible N-terminal tail (residue 112-119), the presence of the charged LYS in position homologous to prion 171 that characterizes some prion polymorphisms that seem to be immune to CJD. A clearer scenario will undoubtedly emerge with the accumulation of experimental findings on chimeric prion-Dpl proteins.

Summarizing, the crucial factors that come into play to determine the prion infectivity and the progression of the disease at a molecular level seem to be: 1) the presence of alternative conformations for PrP sequence beside the normal cellular conformation; that is the lack of a unique stable native conformation; 2) the stabilization of those alternative conformations through binding to already misfolded structures; that is, $PrP^{sc}$ aggregates catalyze the conversion $PrP^c \rightarrow PrP^{sc}$.

The type of analysis carried out here specifically address the first issue regarding the stability of the cellular form of the prion protein and its folding characteristics as it can be determined by native state topology. Important clues may come from appropriate numerical studies addressing the kinetic behavior of the prion structure in the folding process.

The model we adopt builds on the importance of the native state topology in steering the folding process that is in bringing into contact pairs of amino acids that are found in interaction in the native state.

The focus of the present work is precisely the determination of the folding bottlenecks for the cellular prion and the identification of the key amino acids taking part to the corresponding crucial contacts. The goal of the present study is to examine the connection between the set of such key residues with those that are known to be associated with harmful PrP$^{Sc}$ mutations. A connection between the two sets is expected precisely from the paramount role played in the folding process by the key topological residues. As argued before, a random mutation of such sites will usually result in a disruption of the folding process. Only fine-tuned mutations can lead to a wild-type-like native state (as for viral enzymes mutating under drug attack (9)) or in another viable structure (as postulated by Prusiner for the Prion). Furthermore, we will focus our attention on how the differences between the topology of Dpl and PrP may impact on the folding process thus aiding or avoiding the formation of misfolded conformers. It is important to stress that our study builds only on the known NMR structure of PrP$^{C}$ (and Dpl). Hence, although we can confidently identify the crucial folding residues, we cannot speculate whether their mutation results in a different native state form. A primary quantity of interest that we shall calculate is the contribution of individual native contacts to the cooperativity of the folding transition. The crucial contacts will be identified as those with the largest cooperativity.

## Topological approach to prion protein folding

### The model for prion

As stated in the Chapter 1, our models are built on a simplistic representation of the three dimensional structure of the proteins. Thus, we extracted from the NMR structure of the human prion protein (deposited at PDB with id 1qlx) the 104 resolved $C_\alpha$ atom positions, which will be used to represent the amino acid positions in our simplified model. Then, we determined the contacts between amino acids by simply checking the relative distance of the $C_\alpha$'s. If the distance was less than a cut off of 6.5Å then the amino acids were considered in contact, no contact otherwise.

The contact map obtained in this way contains only 85 contact pairs (excluding pairs of amino acids that are separated by 3 residues or less along the sequence). This number of contacts is very small if compared to that obtained from other proteins of the same length (i.e. Ig domain in "Chapter 2" contains 162 contacts). On this basis we formulated a second model whose contact map is constructed considering all atoms: if two atoms (excluding hydrogen atoms) belonging to two different residues are closer than $1.244 \cdot (R_i + R_j)$ where $R_i$, $R_j$ are the van der Waals radii of the atoms (119), the corresponding C$_\alpha$'s are considered in contact. A comparison of the results obtained with the system built on C$_\alpha$ contact map (S1) and the system built on all-atom contact map (S2) would give a clearer understanding of the validity of the methods.

The Hamiltonian we adopt is analogous to the one used in the previous chapter and described in Chapter 1. We set the value of *g, h* and *k* in Eq. 1.3 to 50, 5 and 0.3 respectively, however the type of analysis that we performed has been shown to be almost independent of *g*, *h* and *k* (21).

The same procedure as in the previous chapter has been used to sample the different thermodynamic regimes of our systems and to determine the temperature dependent specific heat (fig. 5). As before, we used the peaks in the specific heat to pinpoint the folding transition temperatures. After having localized them we performed a long run of $16 \cdot 10^6$ time steps for S1 and $8 \cdot 10^6$ time steps for S2 at the transition temperature in order to have a very careful data sampling in that region. It should be noted that for S1 the specific heat presented two well-separated peaks (fig 5 A). In correspondence of the low temperature peak all the tertiary structure contacts disappear and only the β-strand contacts persist at higher temperatures. The conformational change that has been hypothesized for the prion (120) should involve a reorganization of the structure at the tertiary and secondary level. The secondary structure rearrangement should consist in an extension of the β-strand already present in the cellular form (121) and a shortening of the helices. In order to achieve this change, the β-sheet not necessarily has to unfold, while the other tertiary contacts needs necessarily to break. This is the reason why we decided to run the molecular dynamics simulation at the transition temperature where tertiary contacts are lost. Faster runs of $1 \cdot 10^6$ time steps each where made at higher temperatures in order to determine the complete unfolding transition temperature, where also the secondary structure is lost.

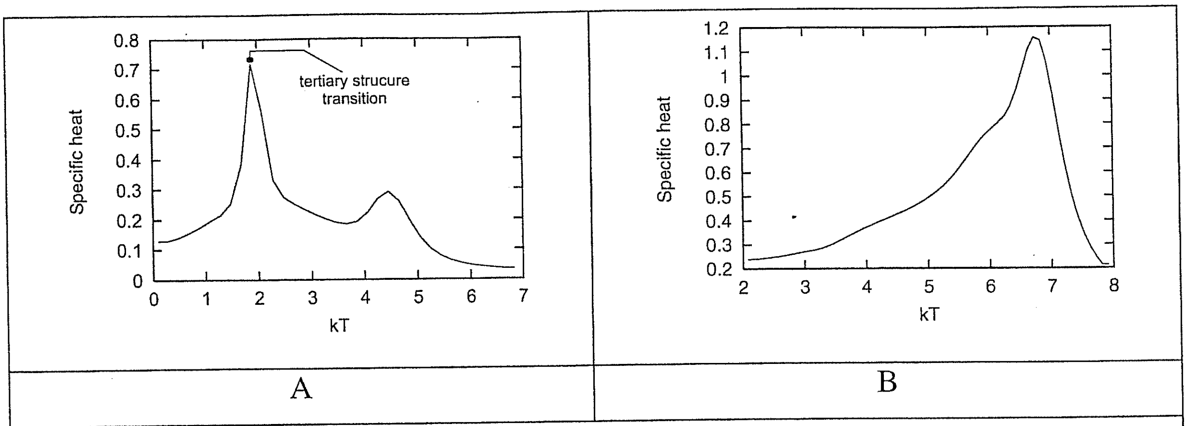


Figure 5. Temperature dependent specific heat for prion models. (A) Specific heat for model S1. The low temperature peak is identified as being related to tertiary structure formation. A long simulation has been, thus, performed at that temperature in order to obtain a very careful sampling for further analysis. The high temperature peak is related to β-sheet formation. (B) Specific heat for the model S2 presents a single peak at folding transition temperature.

A detailed analysis has been carried out at single contact level, starting from data collected at transition temperature. Eq. 1.10 has been used to measure the specific heat associated to each contact along MD trajectories at transition temperatures and Eq. 1.13 for the SRSH.

### *Determination of folding pathways*

In order to have a detailed picture of the different folding pathways that the proteins could follow, we analyzed in full detail the thermodynamic and some of the kinetic features of the models.

In the first part of this analysis we identified the different folding events that in principle could occur independently one from the others. This was naively done for prion protein as follows. On the sequence of the proteins we arbitrarily isolated 3 sets of contiguous residues: the first one (Bh) from residue 125 to residue 172 comprises all N-terminal residues (β-sheet included) before the helix 2 (defined according to the indications in the PDB file); the second set (h1) from residue 173 to 194 comprises residues of helix 2; and finally the third set (h2) from residue 200 to 228 is composed of residues from helix 3. Then, we identified 4 possible events.

- The formation of contacts between residues within Bh;

- the formation of contacts between Bh and h1;
- the formation of contacts between Bh and h2;
- the formation of contacts between h1 and h2.

This classification almost comprises all the contacts with a contact order of 5 or more. α–helical contacts are almost totally excluded from these sets, but we will see that their contribution to the total specific heat at folding transition is negligible. Furthermore, we restricted our attention to the contacts that in each group have the highest contact specific heat. Indeed, in runs at non-zero temperature, but below the transition temperature, the system always present a certain degree of mobility; some of the native contacts will be floppy, continuously switching on and off. The stronger correlation of the contacts with high specific heat with the overall behavior of the protein allows obtaining clearer results. Indeed, the contacts with low specific heat do not correlate with the behavior of the whole system, i.e. they form and break randomly, and considering them in the calculations would substantially increase noise.

We computed the percentage of contacts within each of the 4 groups formed along the trajectory ($Q_i$, with i = Bh, Bh-h1, Bh-h2, h1-h2). Then, we binned the long folding transition trajectories according to the value of this 4 reaction coordinates. This allows getting a detailed picture of the free energy landscape of our system at the folding transition.

In the case of S1, since the global specific heat showed 2 peaks (see fig. 5 A) a simulation done at the low temperature peak would not show folding events that usually occur in correspondence of the high temperature peak. Thus, we decided to carry out 100 folding runs. We started the simulations from completely unfolded conformations, obtained from high temperature runs. The simulations were performed at a temperature below the low temperature transition peak seen in the specific heat (fig. 5 A) and were stopped after 200,000 time steps that allowed all of them reaching the native state. Then, the $Q_i$ along the 100 runs were monitored and were averaged over the runs as a function of the elapsed time since the beginning of the run.

### Doppel protein

We carried out a parallel analysis for the doppel protein (Dpl, PDB id 1I17). Dpl protein structure has been obtained through NMR (122). For this protein no minimized average

NMR model has been released, thus, we decided to use NMR model 1 that conventionally presents less unsatisfied NMR constraints than the others. For this protein, we built both the contact map using $C_\alpha$ (DS1) and the contact map considering all-atom distances (DS2). For both DS1 and DS2 we run MD simulations on a range of temperatures that allowed pinpointing the transition events from the folded to the completely unfolded conformation through an analysis of the specific heat (fig.6). For DS1 there is more than one peak in the specific heat, corresponding to the disruption of parts of the contacts present in the native state. We anticipate that the temperature at which these events occur also dictate the kinetic order of the events in the folding process; i.e. the group of contacts that are formed at the highest temperatures corresponds to the group of contacts that are more readily formed in a folding run performed at low temperature and starting from an extended, completely unfolded conformation.

The analysis of the contact specific heats has been carried on, as in the previous case, after having identified the interesting transition temperatures for both DS1 and DS2. Contact specific heats have been computed by applying Eq. 1.10 on long trajectories ($8 \cdot 10^7$ time steps) performed at transition temperatures.

For comparison with prion a detailed analysis of the folding pathways has been carried

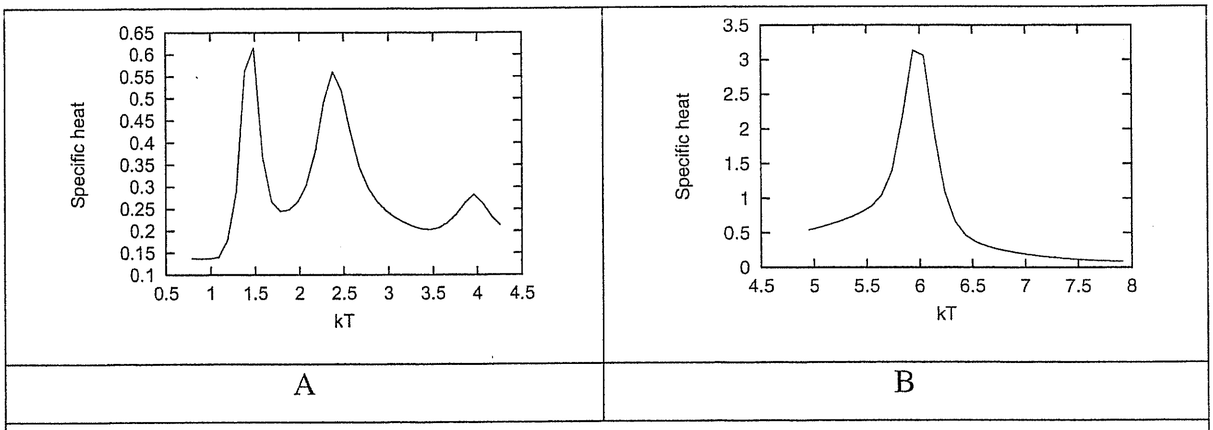


Figure 6. Temperature dependent specific heat for Dpl protein models. (A) Model DS1. The three peaks are related to main folding events: the low temperature peak is related to the formation of contacts between the two terminal helices (set h1-h2, see below). The intermediate temperature peak is related to the formation of contacts between β-sheet and the central helix (Helix 3, 4 in the PDB file). The high temperature peak is related to the formation of the β-sheet. (B) Model DS2. The cooperative behavior in this model is higher than in model DS1, as showed by the presence of a single peak in the specific heat.

out, too. We identified on Dpl protein the sets of amino acids structurally homologous of the sets of amino acids identified for prion: residues from 1 to 50 before helix 3 (Bh), residues from 51 to 76 comprising helix 3 and 4 (h1) and residues from 78 to 96 comprising helix 5 (h2). The numeration of the helices is in agreement with the secondary structure indications present in the PDB file. The same viable folding events were also identified.

For model DS1 folding runs were performed using the same protocol as that used for prion protein. The $Q_i$ were monitored along the 100 folding trajectories and the average behavior during the folding process has been extracted.

For model DS2, the $Q_i$ from a long trajectory ($8 \cdot 10^7$ time steps) at the unique folding transition have been extracted and the free energy profiles have been computed, assuming a good sampling of the conformational space.

## Results

### Contact specific heats

Sites whose mutation is associated to the emergence of $PrP^{sc}$ are summarised in table 1. The top amino acid ranked according to the SRSH computed with both model S1 and model S2 contains a high fraction of those 14 sites of mutations. In the top 10 amino acids according to model S1 we found 6 important mutations and 4 according to model S2. In the top 30 amino acids important mutations both according to S1 and S2 are 9. The statistical relevance of this match can be obtained from Eq. 2.3 that gives the probability to have the same or a better match by chance. For the top 10 amino acids the probability

| | |
|---|---|
| Prion known sites of mutations and experimentally relevant residues | (129), **(161)**, (171), 178, **179**, **180**, **183**, 198, 200, 208, **210**, **214**, 217, (219) |

Table 1 Sites of mutations (in the NMR solved part of Human PrP) causing the inherited prion disease or being fundamental for correct folding of $PrP^C$; the sites in parentheses are polymorphisms, ie some of them are known to influence the onset as well as the phenotype of the disease. Sites 214 and 179 represent cysteine residues that form a disulfide bond needed to observe the $PrP^C$ $\alpha$-form of prion (123). Residues with specific heat ranking within the top 10 are in boldface (model S1).

to have better results by chance is 0.03% and 2.7% for model S1 and S2 respectively. For the top 30 amino acids that probability is 0.3% for both models. Thus, the contact specific heats confidently allow identifying amino acids important in the folding process of the prion protein. In fig. 7 we present both the contact maps for model S1 and S2 colored according to the contact specific heat. Crucial contacts connect secondary structure elements to form the tertiary structure of prion. Mainly they link the β-sheet-helix-1 region to the helix-2-helix-3 region (Bh–h1 and Bh-h2 sets) and helix-2 to helix-3 (h1-h2 set).
The parts of the protein that highly contribute to the folding specific heat of our topological models are located on the C-terminal strand of the β-sheet, and on the two C-terminal helices, with high density on helix-3 (fig. 8).

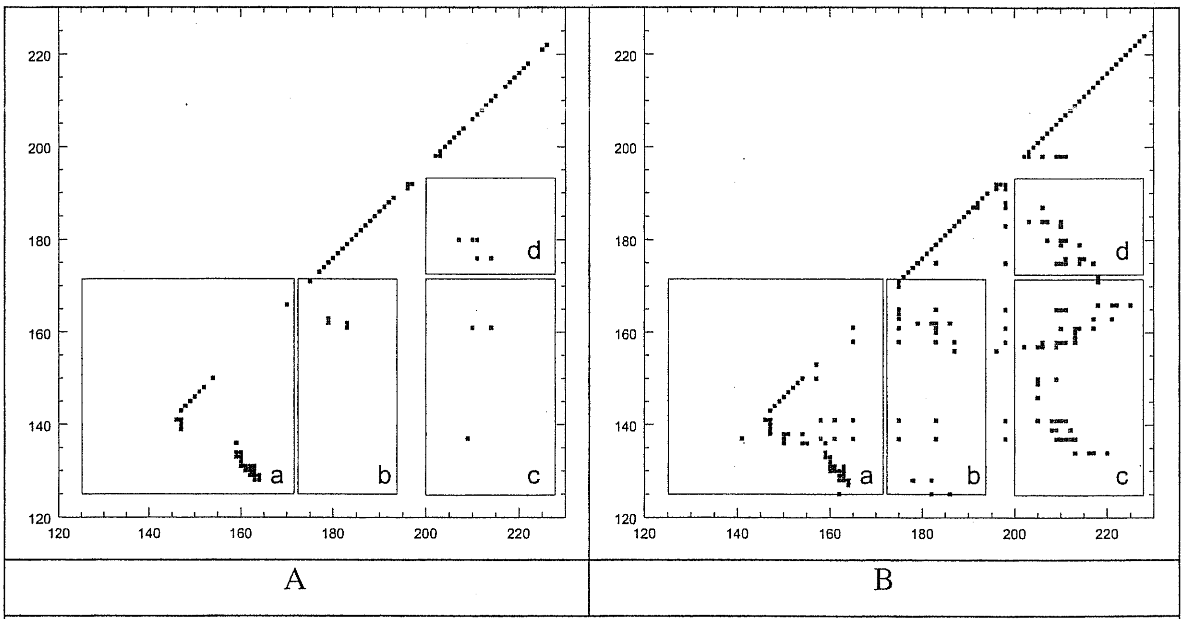


Figure 7. Contact specific heats of prion protein (PDB 1qlx) according to model S1 (A) and model S2 (B). Red contacts show high specific heat, while blue contacts show low specific heat. In both maps the relevant sets of contacts (as defined in the text) are squared; (a) set Bh, (b) set Bh-h1, (c) set Bh-h2, (d) set h1-h2. Red spots are prevalent in set Bh-h1, Bh-h2 and h1-h2.

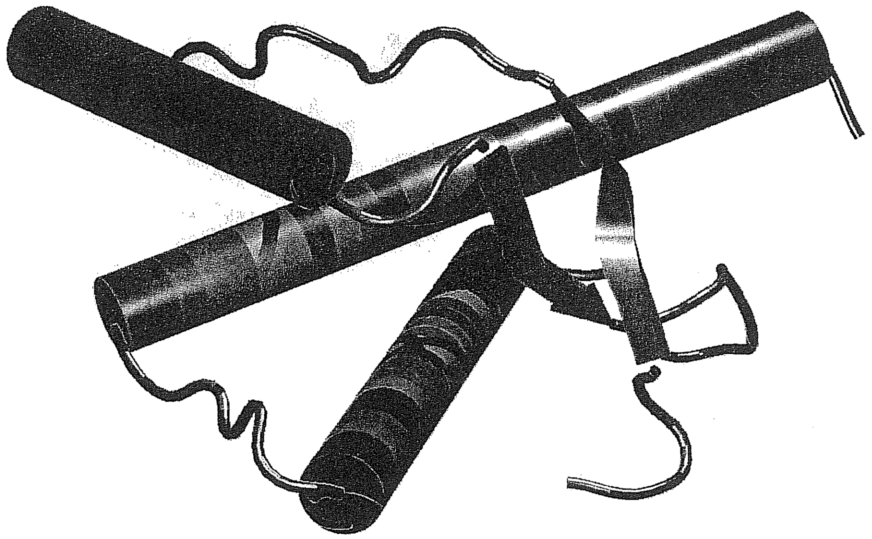


Figure 8. The three-dimensional structure of prion protein where amino acids have been colored according to the value of their highest contact specific heat (model S2). Red indicates high values, blue low values and green intermediate values. Mainly, red spots are located on the C-terminal strand of the β-sheet, on helix-2 and helix-3.

### *Folding pathways*

From the long run performed at the lower temperature peak of the specific heat of model S1, we previously computed the contacts with the highest contribution to the total specific heat. Fig 7 shows that they are all located in the set Bh–h1, Bh–h2 and h1-h2, according to the classification previously defined (contacts in Bh break apart at higher temperatures, thus they do not present high specific heats at the folding temperature). The formation of the two contacts in Bh-h2 only occurs when both the other two sets are formed. On the other hand Bh-h1 and h1-h2 can independently form. Obviously, the folding of the protein occurs when both sets are formed, but this can happen in several ways. The system S1 experienced 35 folding and unfolding events, during the transition temperature run. The analysis of the effective free energy landscape, computed as the

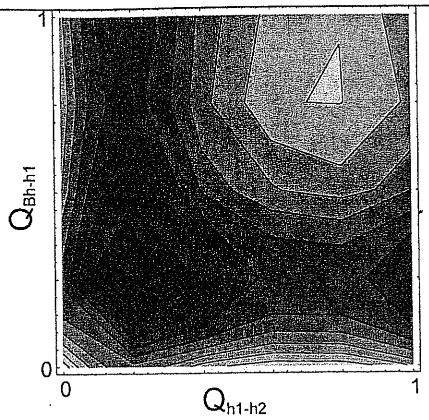


Figure 9. Effective free energy surface of prion model S1 at transition temperature as a function of the formation of contacts in sets $Q_{Bh-h1}$ and $Q_{h1-h2}$ corresponding to the contacts N-terminal - helix 2 and helix 2 – helix 3 respectively. The data represents $-\log(P(Q_{Bh-h1},Q_{h1-h2}))$ where $P(Q_{Bh-h1},Q_{h1-h2})$ is the frequency of the state $(Q_{Bh-h1},Q_{h1-h2})$ along the trajectory obtained from the run at transition temperature (lower peak in the specific heat). White spots represent low values, black spots represent high values. The minimum corresponding to $(Q_{Bh-h1}=0,Q_{h1-h2}=0)$ represents the free energy of the unfolded conformations. The minimum corresponding to $(Q_{Bh-h1}\neq 0,Q_{h1-h2}\neq 0)$ represents the free energy of the folded conformation. Minima corresponding to $(Q_{Bh-h1}\neq 0,Q_{h1-h2}=0)$ and $(Q_{Bh-h1}=0,Q_{h1-h2}\neq 0)$ represents partially structured metastable conformations and obligatory steps for routes from the folded to the unfolded conformations and vice versa.

$-\log(P(Q_{Bh-h1},Q_{h1-h2}))$ in fig. 9, shows clearly the presence of 4 local minima corresponding the different pattern of formation of the two sets of contacts (none formed, first formed and second not formed, the vice versa, both formed). The pathways from the completely unfolded conformation (none formed) and the completely formed one (both formed) can in principle follow two ways, the energetic barriers being comparable.

Thus, at the folding transition, two alternatives routes to the native $PrP^C$ emerged. The first one corresponded to the establishment of contacts between the β-sheet and the central helix (followed by the formation of all the other contacts). In the second case the contacts between the two C-teminal helices are formed followed by formation of all the other contacts. Trajectories following the first pathway from the completely unfolded to the completely folded conformation (all native contacts formed) are less populated (22%) than the second pathway (75%). Only 1 trajectory out of 35 showed a mixed behavior

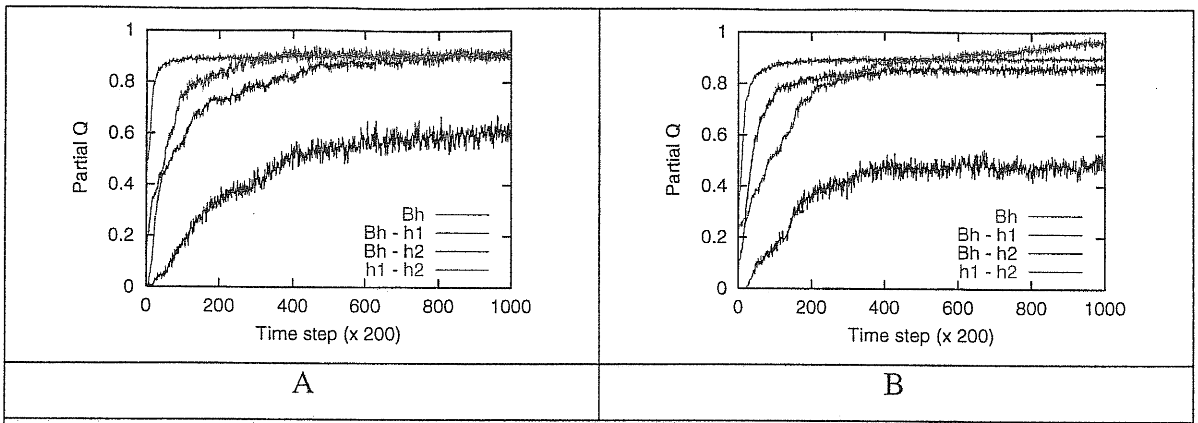


Figure 10. Average time dependent behavior of prion model S1 (A) and Dpl model DS1 (B) computed on 100 folding runs at low temperature started from completely unfolded conformations. Time is measured from the beginning of the run. Each curve has been fitted with an exponential function and the characteristic time τ has been measured (A) In prion model, the emergence of the contacts in set Bh is the earliest event (τ=1800, red line). The second event consists of the formation of contacts in set h1-h2 (τ=13000, magenta line). Third event is the formation of contacts in set Bh-h1 (τ=18000, green line). Set Bh-h2 is only composed of two contacts in model S1; only one of the two is formed in the folded conformation at this temperature; thus the relative overlap with native contacts only reaches 0.5 (τ=50000, blue line). (B) In Dpl protein model DS1, Bh contacts are the first to form, as for the prion (τ=2700, red line). However the emergence of contacts in set h1-h2 (τ=23000, magenta line) follows the formation of contacts in set Bh-h1 (τ=9000, green line). In set Bh-h2 (τ=34000, blue line), there are only two contacts as for the prion; thus, the same explanation for its time behavior has to be applied (see above).

jumping from pathway 1 to pathway 2 before reaching the folded state. The length of the two types of trajectories is quite similar (25000 time steps on average with a large spread (standard deviation 20000)). The study of unfolding trajectories led to quite similar results. The inverse of the first pathway (i.e. loss of contacts between C-terminal helices followed by complete unfolding) is populated in 22% of cases; the inverse of the second pathway is populated in 75% of the cases. Only one exception is seen in which the system goes to the unfolded conformation without passing through pathway one or two. Unfolding trajectories seems to be a little bit shorter than folding trajectories, but not significantly.

The analysis of the 100 folding runs allowed obtaining the average behavior of the system in terms of the temporal formation of the different sets of contacts from the beginning of the run. Fig 10 clearly shows that Bh set is the first to be formed. Very soon, h1-

h2 follows and, after that, Bh-h1. These results state a precise order of events in the folding process of the prion determined according to its geometrical architecture; they will be related to a similar analysis on Dpl and compared to experimental facts. Furthermore, they justify the focus that we adopted to study the transition that occurs at the lower temperature peak where no clear order of the folding events was present.

Model S2 presents a single peak in the specific heat. The analysis of the long simulation performed at that temperature has led to the construction of an effective free energy landscape. As before, it is based on the $-\log(P(Q_i,Q_j))$ and in this case all the combinations of *i* and *j* have been considered. Fig. 11 presents these results. Most of the contour plots present two energy minima, as expected because of the presence of a single folded and an unfolded conformation. However most of the plots show the presence of multiple pathways connecting the two minima. In particular the plot corresponding to the Bh-h1 set versus the h1-h2 set present two possible successions of events: either the Bh-h1 set is formed followed by formation of h1-h2 or the vice versa.

The monitoring of the behavior of Dpl with model DS1 led to the following picture. The specific heat of model DS1 shows three different peaks (fig. 6). As usual, we have found the correspondence of the peaks with well-defined events. The highest temperature peak corresponds, as in the case of prion model S1, to the disruption of the beta-sheet contacts (Bh set); the intermediate temperature peak corresponds to the breaking of the contacts in set Bh-h1; finally the low temperature peak correspond to the rupture of the contacts in set h1-h2. This temperature succession of events has a kinetic counterpart.

The analysis of the 100 folding runs for DS1, performed by monitoring the time dependence of the $Q_i$, led to kinetically characterize the folding process. Fig. 10 B shows that the first event that takes place from the beginning of the folding simulation is the formation of the Bh contacts. This is followed by the formation of the Bh-h1 contacts. The formation of contacts in h1-h2 comes as the third event. Thus, the second and the third events are exchanged with respect to what happens in prion model S1.

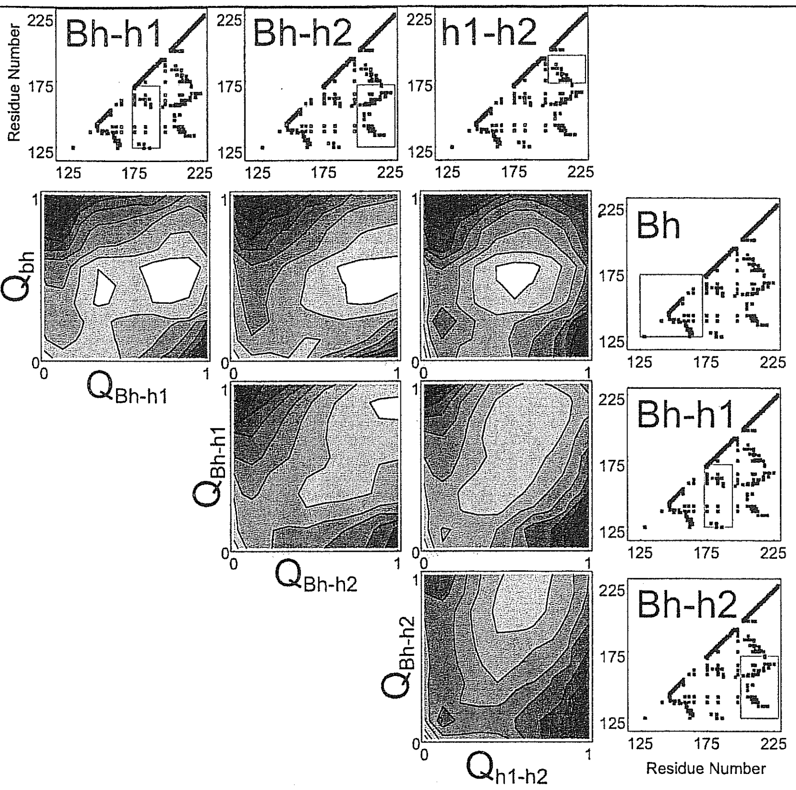


Figure 11. Effective free energy landscape for prion model S2. Upper and right squares represent the contact map of the model in which representative contacts of each set are squared. Contour plots represent $-\log(P(Q_i,Q_j))$ for each pair of set of contacts computed along the trajectory from the transition temperature simulation. White spots represent low values, black spots represent high values. The minima close to the $(Q_i=0,Q_j=0)$ represent the value for the unfolded conformation while the minima close to $(Q_i=1,Q_j=1)$ represent the value for the folded conformation. In most of the contour plots it is not possible to draw a favorable path from the folded to the unfolded conformation.

Model DS2 shows a very cooperative behavior (fig. 6 B). Thus, the MD simulation performed at transition temperature contains information on all the folding events that lead the model from the folded to the unfolded conformation and vice versa. This information is exposed in fig. 12 as the effective free energy landscape according to the $Q_i$ coordinates. In this case, too, the contour plots shows quite clearly two separated minima. They also show that in most of plots there is just one pathway of minimal energy that connects the two conformations. The pathway follows successive steps; as a first step, the formation of the contacts in set Bh clearly precedes the formation of Bh-h2 and h1-h2

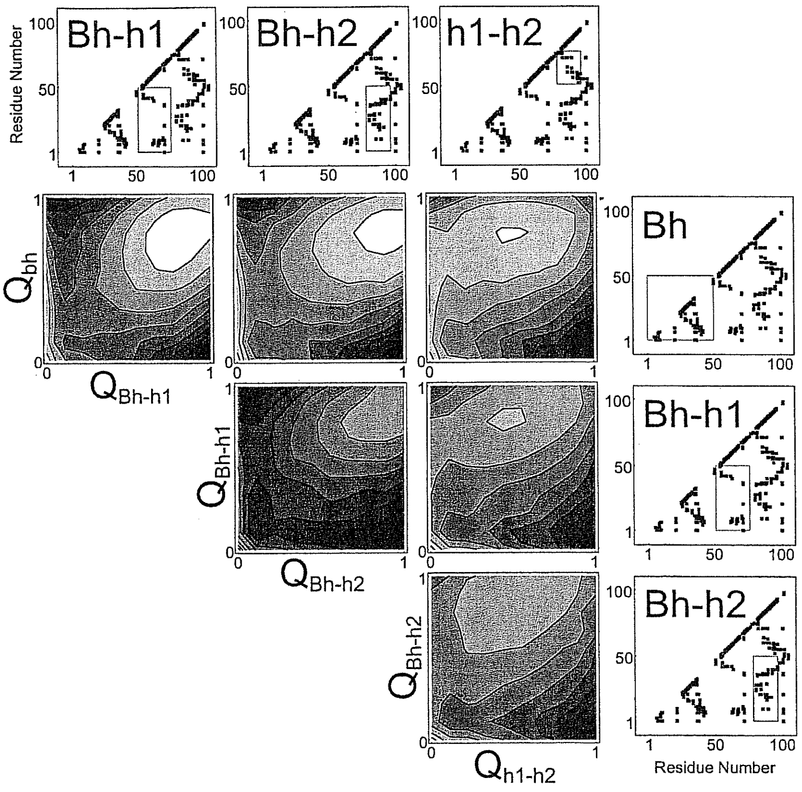


Figure 12. Effective free energy landscape for Dpl model DS2. Upper and right squares represent the contact map of the model in which contacts of each set are squared. Contour plots represent $-\log(P(Q_i, Q_j))$ for each pair of set of contacts computed along the trajectory from the transition temperature simulation. White spots represent low values, black spots represent high values. The minima close to the $(Q_i = 0, Q_j = 0)$ represent the value for the unfolded conformation while the minima close to $(Q_i = 1, Q_j = 1)$ represent the value for the folded conformation. In most of the contour plots there exists a unique pathway of minimal energy from the folded to the unfolded conformation. Following these pathways allow consecutively ordering the events. Thus, formation of Bh contacts, on average, precedes the formation of the other sets of contacts. Bh-h1 contact formation precedes Bh-h2 and h1-h2 contact formation. And emergence of contacts in h1-h2 follows the creation of contacts in Bh-h2.

sets of contacts and slightly precedes the formation of set Bh-h1. Then, the formation of contacts in set Bh-h1 occurs followed by the formation of contacts in set Bh-h2. The last step is the formation of contacts in h1-h2.

### *Analysis of sequence conservation*

An accurate analysis of sequence conservation along the prion family of protein is less meaningful than in the case of antibody sequences for which thousands of sequences are known (see "Chapter 2"). Indeed in the case of prion only few tenths of sequences are known and they do not present a large variability. However, in this case too, amino acids with high specific heat present low sequence entropy, as defined in Eq 2.2 (data not shown).

## *Discussion*

In this study we have found the amino acids that should play a relevant role in the folding process of the prion because of their geometric position within the three-dimensional structure of the protein. In principle, their effective function in the real system could be widely affected by the chemical interactions that take place in between the amino acids of the protein and the elements of the surrounding environment. However, a good fraction of the mutations known to participate in the folding bottlenecks of the prion protein results to be located on topologically relevant amino acid positions, thus indicating that evolution has promoted amino acid substitutions such that their interactions try to stabilize regions that, for mere geometrical reasons, result particularly relevant for promoting an efficient folding.

If the importance of some of the influential mutations seems to be explainable through a very simple geometrical model, based on the characteristic of optimal stability and kinetic accessibility of the native-state of the $PrP^C$, the mutations that do not follow this paradigm will maybe affect different aspects of the process that leads the cellular form of the prion to the scrapie form. As it has been indicated before, the other feature of such a route is the template binding of $PrP^C$ to $PrP^{Sc}$ involving the interconversion of the first species into the second. Thus, the present results not only indicate a viable explanation for the relevance of some of the mutations in the arousal of the prion diseases but also assign a probable role to mutations that escape the geometric explanation, provided that the folding stability of $PrP^C$ and its binding affinity to $PrP^{Sc}$ oligomers play a crucial role in the folding process of prions. Indeed, the amino acids that do not present high specific heats within the present models and that are known to be involved in the development of the disease (specifically MET 129, GLU 200 and GLU 219), are mostly charged residues with a large exposed surface area.

The detailed dissection of the folding process carried out in the present study has been dictated by the characteristics of low cooperativity experimentally established for prion (118, 123).

Despite the substantial topological similarities of Dpl and $PrP^C$, the detailed studies of the folding trajectories of Dpl, have revealed significant differences of the folding pathways from $PrP^C$. Remarkably, for Dpl only one main pathway is observed, and the series of folding steps is more distinctly marked than for $PrP^C$. Such noticeably different behavior can be ascribed to the different set of contacts stabilizing the assembly of the α-helices and β–sheets. The long-range (in sequence separation) nature of such contacts and the fair degree of burial of the sites involved in their formation ultimately lead to their crucial folding role within our topological approach. Although, the present picture would certainly be modified by the introduction of suitable amino acid specific pair-wise interactions, it is appealing to connect the different folding routes to the appearance of misfolded conformers. According to this view, the latter would result from following the folding route to $PrP^C$ where the inter C-terminal helices contacts are formed while the contacts with the β-sheet are not yet formed (fig. 10 A and 11). Indeed, in this pathway, which is alternative to that found for Dpl (fig. 10 B and 12), the N-terminal part of the protein is free to rearrange its structure before reaching the native-state. This scenario is consistent with several recent models that identify the N-terminal part of the protein as the one that undergoes the conformational change that leads the prion from its cellular to its scrapie (misfolded) conformation (98, 121, 124).

# Chapter 4 – NGF/TrkA Complex

In this chapter a structural analysis of the Nerve Growth Factor (NGF) in complex with domain 5 of Tyrosine Kinase A (TrkA) receptor has been performed. The analysis is based on Molecular Dynamics (MD) simulations of an all-atom model of the complex immersed in explicit water molecules. The organization of the chapter is the following. In the first part a short introduction on the experimental aspects of the problem and on the motivation of this work has been presented. Then, a description of the methodologies used for the preparation of the system, the molecular dynamics simulations and the analysis is given. Finally the results are exposed and discussed.

### *Introduction*

NGF is currently under investigation as therapeutic agent for the treatment of neurodegenerative disorders (125). It serves as a survival and a differentiation factor for the nervous system. Neurons respond to NGF based on NGF interactions with two cell surface receptors, p75 and p140 receptor tyrosine kinase A (TrkA) (126).

The recent determination of the X-ray structures of the NGF/TrkA complex (127) (fig. 1) has provided a structural basis to understand the effect of a large wealth of chemical and genetic modification experiments performed on NGF and TrkA. In particular it has confirmed the crucial role played by specific domains on NGF, such as the N- and C- termini (128-132), the hairpin loops and residues 91-97 (129, 133, 134). Analogously, it has allowed rationalizing mutagenesis experiments on TrkA (135) that have shown that the EF loop as well as other residues determines the affinity for the NGF, the latter being fundamental for specificity.

Furthermore, a conformational change of the N terminus of NGF upon binding has been assessed by the determination of its structure in the NGF-TrkA complex that was not resolved in the crystallographic structure of the free NGF(136). Thus, the N-terminus of NGF probably undergoes a transition from an unstructured conformation to a structured conformation upon binding to TrkA.

However, in spite of the structural information, the rationalization of few mutagenesis experiments (135, 137) appears to be challenging. Indeed, few mutations at the NGF/TrkA surface that involve charge neutralization (such as E324A, E331A, E334A) do not cause loss of bioactivity (Table 1). In contrast, replacement of a *polar* residue with a neutral one, which is expected to cause only a small decrease in the interaction energy (T352A, H353A, N356A) affect dramatically the binding.

Clearly, a deep understanding of the key factors governing the molecular recognition between the TrkA receptor and its ligand would be of outmost importance to design novel mutants and, more importantly, to design new and powerful peptide NGF mimics (138).

In order to investigate the role of thermal fluctuations on ligand/receptor binding, we have performed molecular dynamics simulations complemented by an electrostatic analysis of the TrkA/NGF complex. The calculations suggest that the intrinsic flexibility of the ligand/receptors regions of NGF, mediated by residues in the region 352-356, are very important for the formation of a productive NGF-TrkA complex. Thus, the complex uses principles other than structure and energetics for the molecular recognition.

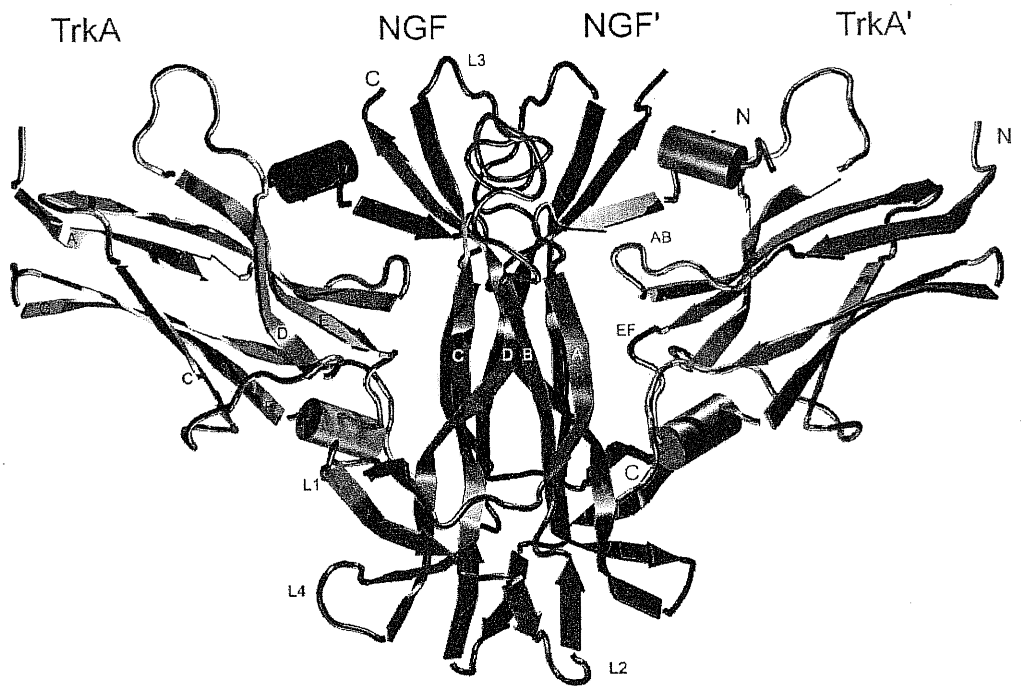


Figure 1. Cartoon of three-dimensional structure of the complex formed by NGF homodimer and two TrkA domain 5 subunits where secondary structural elements are represented. The NGF chains are in red and in blue respectively. The TrkA subunits are in orange. The encoding of the secondary structural elements is the same as used by crystallographers (127). N and C termini of the chains are indicated. The regions of interactions between TrkA and NGF have been identified by crystallographers. The first is located along the four central beta-strands and L1 Beta-hairpin of NGF and the AB, C'D, EF loops and C-terminus of TrkA (*common patch*). It comprises residues well conserved along both the neurotrophin family (R103) and the Trk receptor family (F327, N349); it has been supposed that similar interactions take place in the other neurotrophin-receptor complexes (NT-3 with TrkC and BDNF with TrkB). The second region is specific for NGF-TrkA complex. It involves residues from the N terminus of NGF and from ABED strand of TrkA (*specific patch*).

| | TrkA Mutant | $IC_{50}$(mutant)/$IC_{50}$(TrkA) |
|---|---|---|
| Mutations that largely decrease binding | T352A | > 100 |
| | H353A | > 100 |
| | N356A | > 100 |
| | P302E | > 100 |
| | H343A | 78 ± 21 |
| | H343E | 68 ± 26 |
| Mutations that do not largely affect binding | E324A | 2.0 ± 0.3 |
| | E331A | 1.3 ± 0.1 |
| | E334A | 3.1 ± 0.1 |
| | E339A | 6.6 ± 0.5 |
| | R342A | 1.4 ± 0.1 |
| | R347A | 1.7 ± 0.1 |
| Mutations that slightly improve binding | E295A | 0.6 ± 0.1 |
| | H298A | 0.8 ± 0.1 |
| | S326A | 0.7 ± 0.0 |

Table 1. Effect of selected point mutations on $IC_{50}$ (135). Mutations largely affecting electrostatics (third row) seem to play a minor role with respect to mutations that only slightly affect it (T352A, H353A, N356A).

## Methods

### Structural model

Our starting model is based on the NGF-TrkA-d5 complex from *Homo Sapiens*, whose X-ray structure has recently been solved at 2.2 Å resolution (127) (PDB entry 1www)(fig. 1). The same numeration as in the pdb entry for the residues of the complex has been used. Residues belonging to NGF identical subunits 'V' and 'W' are numbered from 2 to 115, and 2' to 115', respectively. Residues belonging to X and Y identical subunits of TrkA are numbered from 282 to 382 and from 282' to 382'. Residues P61(61'), N62(62'), P63(63'), V64(64'), D65(65'), S66(66') are not present in the X-ray structure. They were added using the following procedure. First, the residues D61(61'), P62(62'), P63(63'),V64(64'), D65(65'), D66(66') were extracted from the NGF-serine proteases complex from mouse (136)(PDB entry: 1sgf). Then, D61(61'),P62(62'),D66(66') side chains, were replaced by P,P,N,N,D,D, respectively, using the InsightII program. All histidine sidechains have been considered as protonated in $N_\varepsilon$ according to their H-bond pattern. Acetyl groups and N-methyl were added at the N termini and the C-termini of each chain. The overall charge of the complex was –2. Electroneutrality was insured by adding 2 $Na^+$ ions. The first (the second) $Na^+$ ion is close both to D60(D60') charged oxygen and to N62(N62') carboxyl backbone oxygen of NGF. The system was immersed in a periodic box of 13030 water molecules that allowed 16 Å minimum distance between images in neighboring cells.

### Molecular Dynamics Simulations

The all-atom AMBER5(139) force field (parm96) was used for the protein and $Na^+$. The TIP3P model was used for water (140). Periodic boundary conditions were applied. A residue-based cutoff of 10 Å was used for the Van der Waals interactions. Electrostatic interactions were computed using the Ewald particle mesh method (141). The dielectric constant was set to 1.0. Bonds involving hydrogen atoms were constrained with the SHAKE algorithm (142). The time step was set to 1.5fs. Constant room temperature and pressure simulations were achieved by coupling the systems with Berendsen thermostat and barostat with 0.2ps coupling time constant (28).

The protocol adopted for the simulation was the following: (i) Minimization of protein hydrogen atoms, sodium counterions and waters; (ii) 3ps MD at room temperature of the same atoms; (iii) minimization of the same atoms plus residues 61-66 and 61'-66'; (iv) 24 ps MD at room temperature of the same atoms; (v) 15ps MD from 0 K to 300 K at 1 Atm pressure of the entire system; (vi) 2.6 ns of room pressure/temperature MD molecular dynamics. The last 1.9 ns were collected for analysis. The final MD structure is available at http://www.sissa.it/cm/bc.

### *Calculated properties*

(i)RMSD's and geometrical properties have been measured through the *carnal* module of the AMBER package.

(ii) The water molecules whose oxygen is located within 3.5 Å from both receptor and protein ligand, were identified as *interface waters. Ordered waters* were defined as follows: a 1 Å cubic grid was defined within the simulation box. Grids at different time instants were shifted so as to keep the protein conformation fitted to the initial structure. The population of a single grid cell was then calculated as number of times a water oxygen was present during the dynamics divided by the number of MD steps. Only the grid cells whose oxygen was located within 3.5 Å from both receptor and protein ligand region were considered. The averaged value turned out to be $0.013(0.009)\,\text{Å}^{-3}$. The grid cells occupied more than $(0.013+4\times 0.009)\,\text{Å}^{-3}$ were selected. The water molecules that occupied the grid cells with MD-average population larger than $(0.013+4\times 0.009)\,\text{Å}^{-3}$ were defined as *ordered waters.*

(iii) A principal component analysis has been performed on the C$_\alpha$'s trajectories (we excluded from the analysis the two residues of each chain terminus). The diagonalization of the covariance matrix led to the identification of the eigenvectors with the largest eigenvalues. The covariance matrix has been obtained from C$_\alpha$'s trajectories through:

$$C_{ij} = \langle (x_i - \langle x_i \rangle) \cdot (x_i - \langle x_i \rangle) \rangle$$

Where the averages are computed along the trajectory and the $x_i$'s are the coordinates of the C$_\alpha$ atoms (index *i=3n* represents the *x* coordinate of the *n*–th C$_\alpha$, index *i=3n+1* the *y* coordinate, index *i=3n+2* the *z* coordinate). Diagonalization of the covariance matrix has been obtained through the *SSYEV* routine from the lapack package (143);

(iv) The program *Hingefind* (144) has been used to obtain the principal hinge axes of the complex on the basis of the principal eigenvectors described above. This program processes two protein structures that are generated from each eigenvector. The structures are extracted from the trajectory in correspondence of the minimal and maximal projection of the trajectory itself along the eigenvector. The program is based on an algorithm that divides the amino acids in subsets, simultaneously minimizing the RMSD inside each set. They were partitioned into domains with the *tolerance* (see (144) for the definition) set below the rmsd for the pair of structures. The hinges of rotation of the domains have been identified.

(vi) Atomic fluctuations as obtained from the simulation are compared with crystallographic B-factors through the formula (145):

$$B = 8\pi^2 (RMSF)^2 / 3$$

The RMSF is the root mean square fluctuation of the atom around its average position during the molecular dynamics. B is the crystallographic B-factor of the atom.

(v) Electrostatic force field based calculations were obtained with module *anal* from the AMBER suit of programs for MD simulations. Program DELPHI (146) has been used for continuum electrostatic calculations. This program solves the Poisson-Boltzmann (PB) equation for a set of atomic charges immersed in a continuum dielectric. The use of a continuum electrostatic model assumes an averaging process on the solvent degrees of freedom; this is the reason why energies computed in this way are commonly called electrostatic free energies.

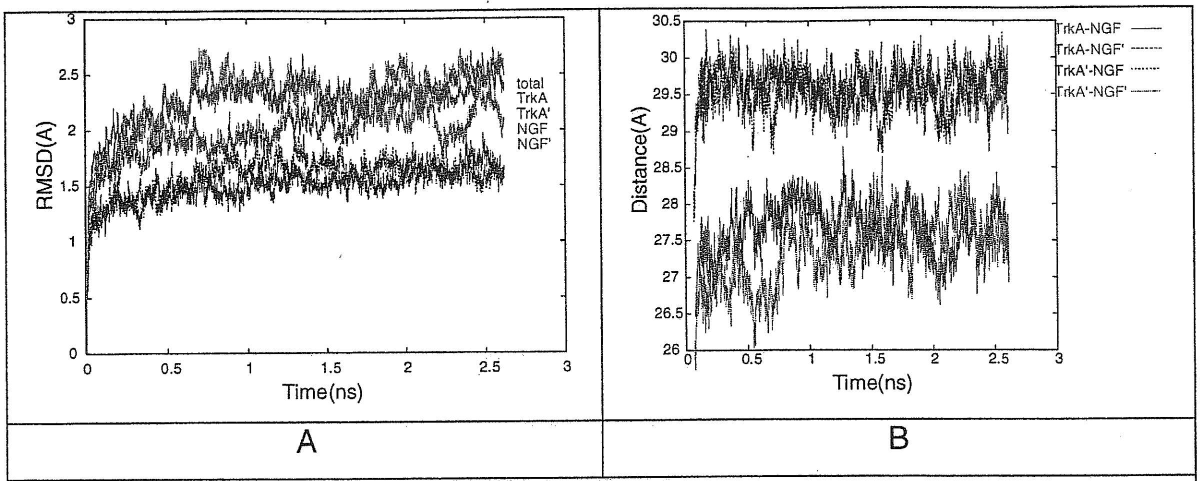


Figure 2. (A) RMSD from the initial conformation computed for the structures sampled along the trajectory. A plateau is reached after 0.7ns. TrkAs subunits show an asymptotic RMSD of 1.5 while NGF monomers reach an asymptotic RMSD of more than 2. This underlines the higher flexibility of the NGF monomers with respect to the TrkAs domains. (B) The distances between the centers of mass of the four chains. Consistently with (A), a plateau is reached after 0.7ns.

## *Results*

### *Structural Properties*

The overall structure of the complex and of the single subunits appear to be equilibrated after ≈ 0.7 ns, as deciphered by a plot of the rms deviations from the energy-minimized structures of the single chains and of the whole complex (fig. 2 A). Consistently, the distances between NGF and TrkA centers of mass oscillate around an equilibrium conformation after 0.7 ns (fig. 2 B). Indeed the crystallographic B-factors of the atoms do not strictly correlate with the analogous measure obtained from the RMS fluctuations computed along the MD trajectory (data not shown). The TrkAs are very mobile in the simulation while they are much more static in the crystal. The discrepancies between theory and experiment may be due to the presence of packing forces in the X-ray structure, where TrkAs are closely packed with their symmetry related analogous and the top loops of NGF, their movement being, thus, inhibited.

The secondary structure elements are well conserved during the dynamics, as showed by a comparison between X-ray and MD-averaged structures (Fig, 3). The structures of selected MD snapshots exhibited good Ramachandran plots (147), except for very few resi-

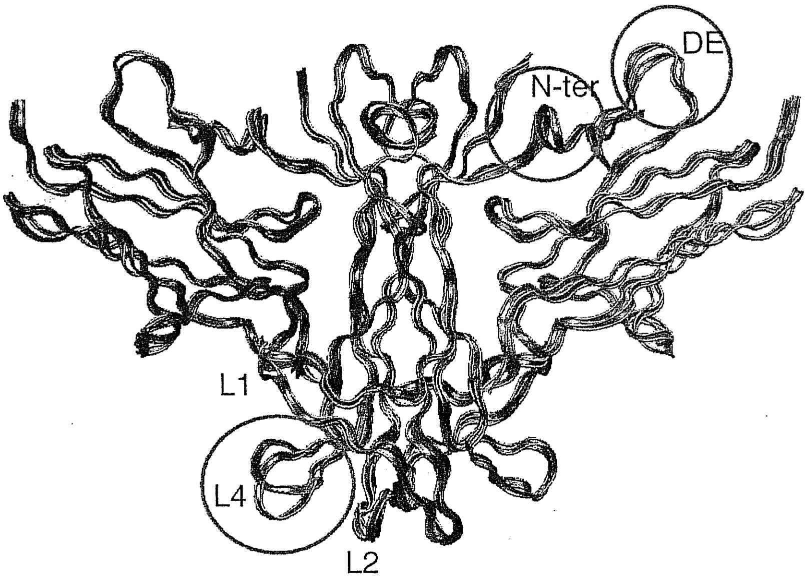


Figure 3. Comparison between X-ray structure (in blue TrkA, in red NGF, in grey NGF' and in orange TrkA') and the average structure after 2.6ns MD simulation (cyan). Secondary structural elements are fully maintained. Differences (green circles) are located in the N-terminus of NGF, in DE loop of TrkA' and in loop L4 of NGF.

dues belonging to loops regions of NGF far from the NGF/TrkA interface (Asp66, Asp66', Asn46, Asn46'), which exhibited φ and ψ angles in relatively high-energy regions of the plot.

The conformational properties of the two symmetry-related NGF and TrkA subunits differ. The structural differences in L3 and L4 loops of NGF pointed out by the crystallographers (127), are maintained. The rmsd between the CA positions of the symmetry related halves of the average structure of the complex is 0.9; the difference is mainly located on the L4 loop of NGF. Other differences (half in magnitude with respect to L4) are found in the N-terminus of NGF and in the DE loop of Trk. Instead, almost negligible structural differences are found for L2 and L1 loops. The discrepancies at the N-terminus are mainly due to the different ARG9 side chain interactions (fig 4 A). On one interface this residue mainly interacts with GLU334 charged side chain; on the other interface its charged side chain interacts both with GLU334 side chain and with GLU334 backbone carboxyl oxygen.

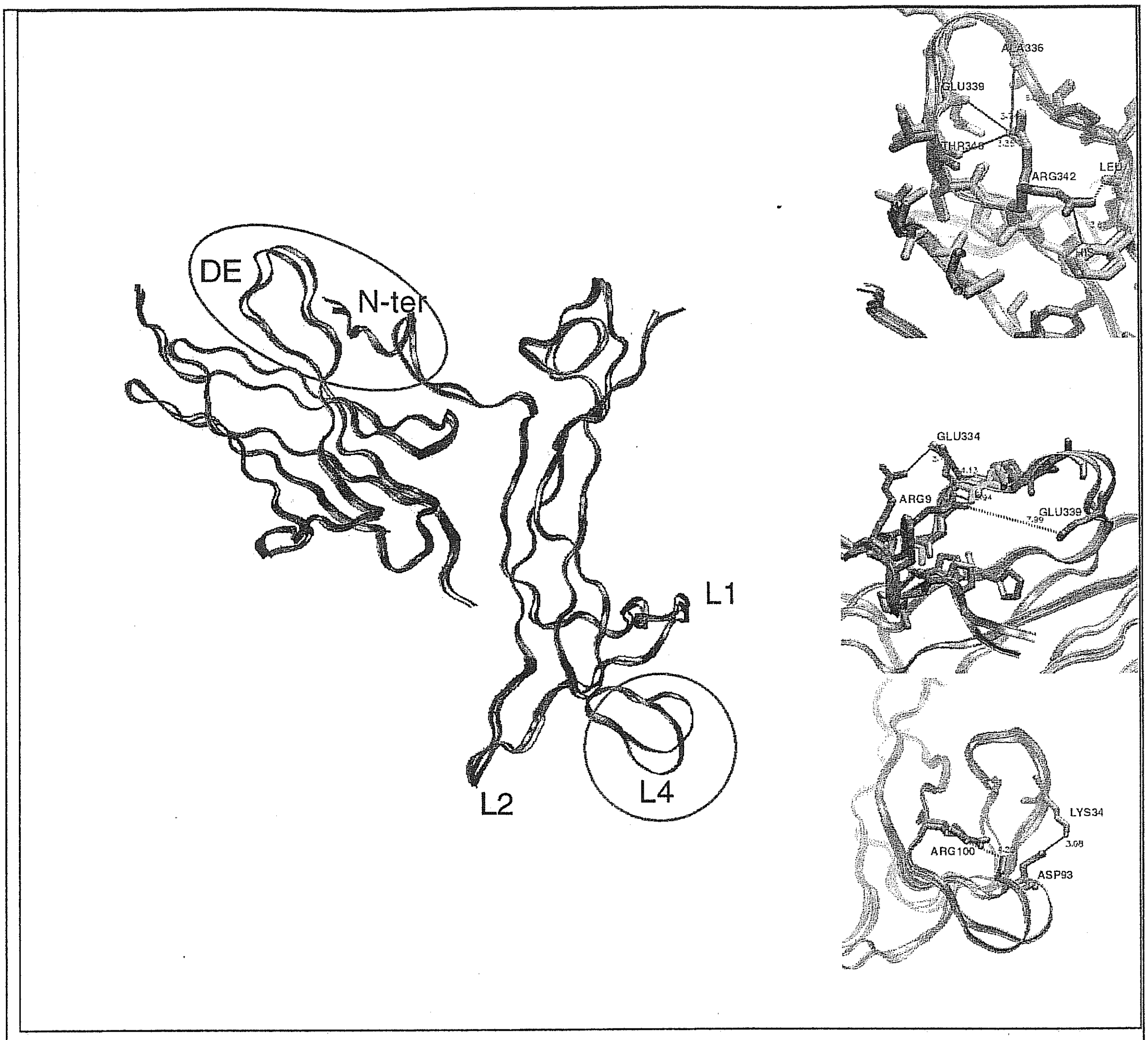


Figure 4. Differences between the two average halves of the complex. On the left side the whole structures are represented (red for the TrkA' and NGF and cyan for TrkA and NGF'). Circled in orange, the locations of the major differences. On the right hand a detailed view of the differing parts (green and dark green for TrkA' and NGF respectively, and yellow and dark yellow for TrkA and NGF' respectively). In the different cases the difference is related to a side chain tilt in a charged residue.

The differences in the DE loop are due to the different ARG342 interactions (fig. 4 B). On one interface this amino acid directs its charged side chain towards ALA 336, GLU 339 and THR 348 carboxyl oxygen; on the second interface it preferentially interacts with a histidine side chain and a leucine backbone.

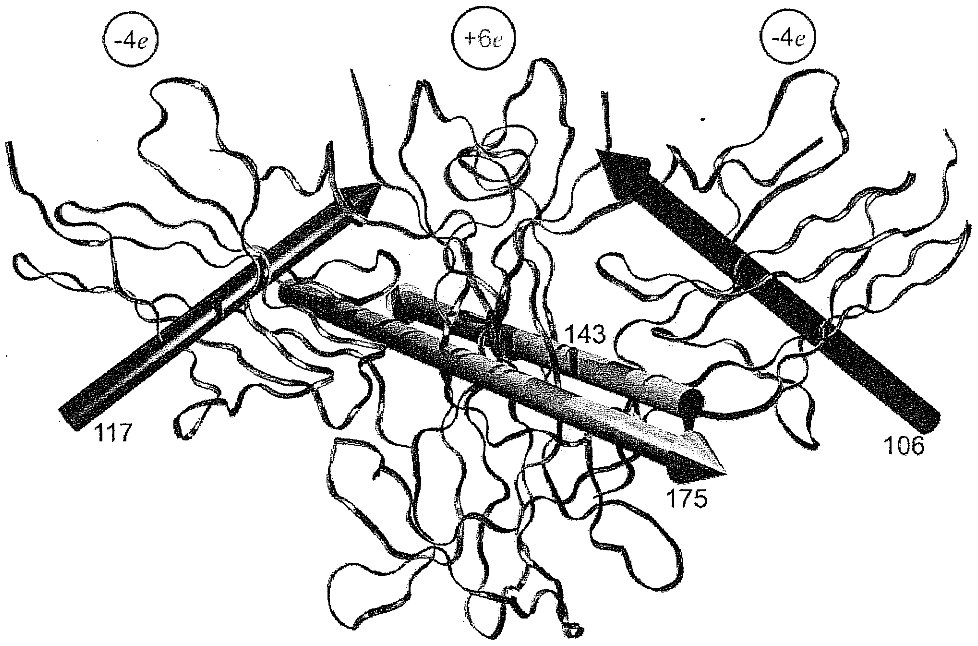


Figure 5. Long-range electrostatic interactions in the NGF-TrkA complex. Beside a huge non-specific attractive interaction due to the opposite charges on the two species (in blue circle for NGF homodimer, in red circle for TrkA subunits), a favorable dipolar interaction is also present (green cylindrical arrows for TrkAs and yellow cylindrical arrows for NGF monomers; Debye units are used). The dipolar interaction seems to be particularly favorable for stability of the NGF homodimer.

The differences at the L4 loop are related to the different ASP 93 interactions (fig. 4 C). On the first interface it interacts with ARG 100; on the second interface it interacts with LYS 34.

### *Ligand/protein Interactions*

Long electrostatic interactions highly stabilize the adduct as TrkA subunit is negatively charged and NGF homodimer is positively charged. Ligand and receptor are further stabilized by a strong dipolar coupling (Fig. 5). Notably the dipolar interaction stabilizes also the two NGF monomers, which, having the same charge, would repel each other.

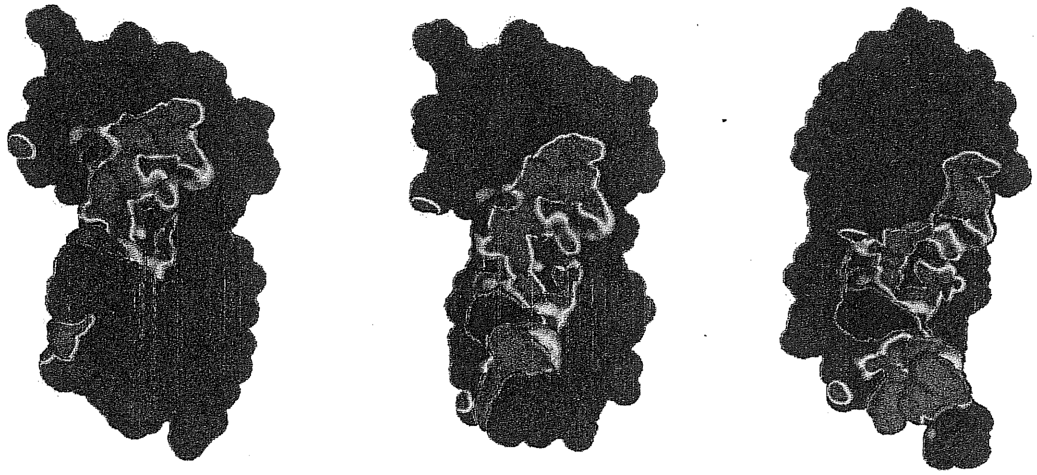


Figure 6. Electrostatic potential generated from free NGF on its surface (blue negative, red positive). For clarity, the potential is plotted only in correspondence of the contacts with TrkA. Only a relevant negatively charged spot is seen in correspondence of residue GLU11. Three different views are shown that differ for a slight rotation. The structure of NGF has been extracted from the minimized structure of the complex.

These global dipolar interactions are fully maintained during the dynamics; the complex is conformationally very stable as shown by the RMSD in fig. 2.

Electrostatic calculations based on the solution of the PB equation (see "Calculated properties") reveal that there is an evident electrostatic complementarity between the two contacting surfaces of ligand and receptor. The calculations on the free components have been obtained by simply extracting their structure from the minimized structure of the complex. Fig. 6 shows that in solvated free NGF the surface of contact with TrkA is mostly positively charged, with a negative potential region generated by GLU11. On the other hand, in solvated free TrkA domain, the surface of contact with NGF is mostly negatively charged, but for a positive potential region located on ARG347 (fig. 7). The two spots results to be complementary; indeed, ARG347(ARG347') and GLU11(GLU11') form a salt bridge that is present in the crystal structure and is conserved along the simulation. Nevertheless mutagenesis experiments do not indicate these two residues as being fundamental for binding (tab. 1); this may be related to the pres-

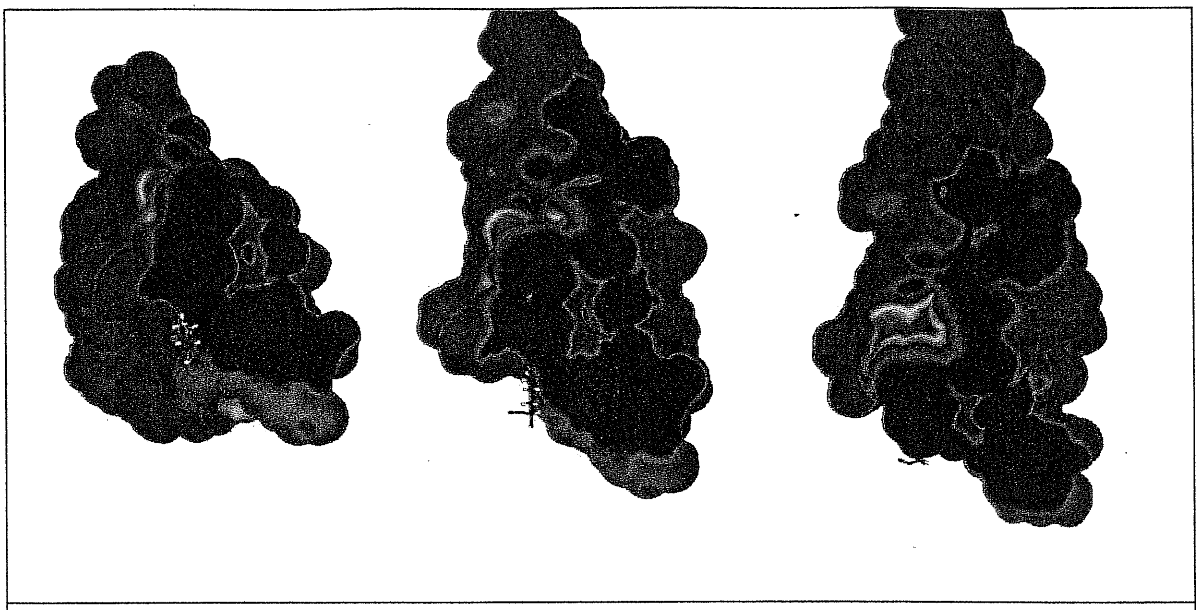


Figure 7. Electrostatic potential generated by free TrkA in water. For clarity, the potential has been plotted only in the region of contact with NGF. Only a relevant positive spot (red) is seen in correspondence of residue ARG 347. The positive spot is the counter part of the negative spot seen on NGF surface. ARG 103 of NGF is reported as sticks to better understand the interface orientation. The structure of TrkA has been extracted from the minimized structure of the complex.

ence of the other electrostatic interactions at surface that may be sufficient for the binding.

The electrostatic free energy difference in water between the complex (minimized structure at t=0) and its free components (from the minimized structure at t=0) is ~-100Kcal/mol. We computed the contributions to this value related to each residue. The residues that mostly favor the electrostatic binding are listed in tab. 2.

We also listed the residues that have the highest unfavorable effect on the electrostatic binding. In this list mainly charged and polar amino acids appear. GLU 11, which from fig. 6 and 7 seem to be important for the complementarity of the contact regions, seems to have unfavorable interactions in the minimized structure. Nevertheless GLU 11' on the opposite interface results to have favorable interactions ($\Delta G_{el} = -0.7 Kcal/mol$). The favorable electrostatic interactions are concentrated on the N-terminus and beginning of strand A of NGF(NGF') and on the AB and EF loops of TrkA(TrkA'); these regions are close one to each other as can be seen in fig. 1. On the other hand, unfavorable interactions are spread along the sequences of the protein chains; however some of them, THR

| Residue | $\Delta G_{el}$ $(Kcal/mol)$ |
|---|---|
| GLU 295 | -7.1 |
| GLU 295' | -6.9 |
| SER 19' | -4.9 |
| ARG 9' | -4.6 |
| GLN 350' | -4.0 |
| HIS 297' | -3.6 |
| HIS 297 | -3.4 |
| SER 17 | -3.1 |
| SER 13 | -3.1 |
| GLU 350 | -2.8 |
| ASN 349' | 1.9 |
| ARG 59 | 1.9 |
| ASN 349 | 1.8 |
| ARG 59' | 1.5 |
| PHE 303 | 1.5 |
| PHE 303' | 1.3 |
| THR 83 | 0.86 |
| THR 83' | 0.79 |
| GLU 11 | 0.77 |
| GLU 334 | 0.57 |

Table 2. The main amino acid contributions to the electrostatic free energy of binding of the NGF-TrkA complex. The most favorable (upper box) and the most unfavorable (bottom box) contributions are shown. Data have been collected using DELPHI to solve the Poisson-Boltzmann equation for the complex and its free components.

83(83'), ASN 349(349') and GLU 11, are located in the same region of the favorable interactions. PHE 303 and 303' are the only aromatic residues with large unfavorable electrostatic interactions. This occurs mainly due to the backbone oxygen atom that interacts with the N-terminus of NGF(NGF').

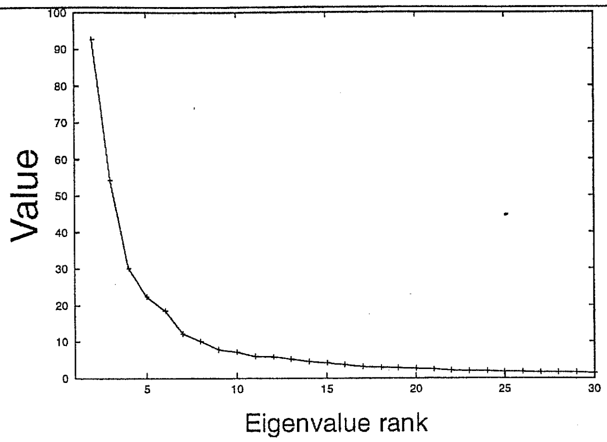


Figure 8. Spectrum of the eigenvalues of the covariance matrix. The first four eigenvalues account for more than half of the fluctuations of the system during the MD simulation. This allows representing a large part of the complex motions of the system just through few coordinates, i.e. the projections of the trajectory along the first eigenvectors.

A similar electrostatic analysis has been carried out with explicit solvent along the trajectory using the program *anal.* According to this analysis, water-mediated interactions accounted for about 50% of the total binding. Water molecules in fixed positions show higher per-molecule energy (-12~-13Kcal/mol) than the others (-8.6~-9.9Kcal/mol), indicating that those positions are energetically favored. Interface water molecules has been 55 ± 4, 56 ± 5 respectively at the two interfaces. The ordered water molecules (see *Calculated Properties* for the definition) were on average 20 ± 3 at the first interface and 17 ± 3 at the second.

### *Large scale motions*

The long-time fluctuations can be probed by the essential modes (eigenvectors) associated to the covariance matrix (see section *Calculated Properties*). Fig. 8 shows that the largest ten eigenvectors account for more than two thirds of all the fluctuations of the system; the four largest eigenvector account for more than half of the total fluctuations.

We displayed the eigenvectors as a vector field on the structure of the complex. The vector on each $C_\alpha$ atom represents the component of the motion of the eigenvector on that residue (fig. 9). This allows obtaining a visual rationalization of the meaning of the eigenvector. The eigenvectors of the covariance matrix with the largest eigenvalue (principal eigenvectors) represent collective motions of the residues. The amino acids cooperatively move along defined directions or according to overall rotations. In particular, we found that according to the first 5 eigenvectors, amino acids belonging to the same peptide chain tend to group together in concerted motions (fig. 9, big arrows). We detailed this analysis by exploiting the program *Hingefind* (144) to correctly identify the domains of motion and related screw axes according to each eigenvector (fig. 11 and par. "*Calculated Properties*"). We also identified the amino acids presenting the largest movements along the sequence of each chain (fig. 10). The more flexible parts, as expected, are located in correspondence of loops or minimally structured regions.

The first eigenvector represents a sort of wings' motion of the complex perpendicular to its main plane (the plane on which all the 4 chains lie). The relative motions of the different chains are directed perpendicular to the plane; the two NGF subunit move in the opposite direction with respect to the TrkA subunits. This mode is large on the TrkA and TrkA' domains and on the N and C termini of NGF that strictly follow the motion of TrkAs (fig. 9 A, 10 I, 11 A). Looking at the projection of the trajectory along this eigenvector we found that the complex prefer two particular conformations (fig. 12 I).

The second eigenvector represents a collective wing like motion but, in this case, it occurs along the same plane of the complex (fig. 9 B, 11 B). The bimodality of the behavior is less strong than in the previous case (fig. 12 II). The eigenvectors with lower rank present more complex motions where new domains of motion are present (fig. 11 C D E).

The binning of the projection of the trajectory along the first eigenvector (fig. 12 I) enlightens the presence of a sort of two-state behavior. Interestingly the system seems to sample more frequently some regions of the space, sometimes jumping from one region to the other. These regions could be related to free energy minima of the system. They correspond to two different angular positions of the TrkA's subunits with respect to the NGF homodimer (the angle is defined around the hinge axes of fig. 11 A). The angular difference is 10deg for both TrkA subunits. To identify the amino acids mostly involved in the motion, we have selected those whose $C_\alpha$ falls closer than 5 Å to the hinge axes (fig. 13). Interestingly most of them have been found important for binding in previous experiments.

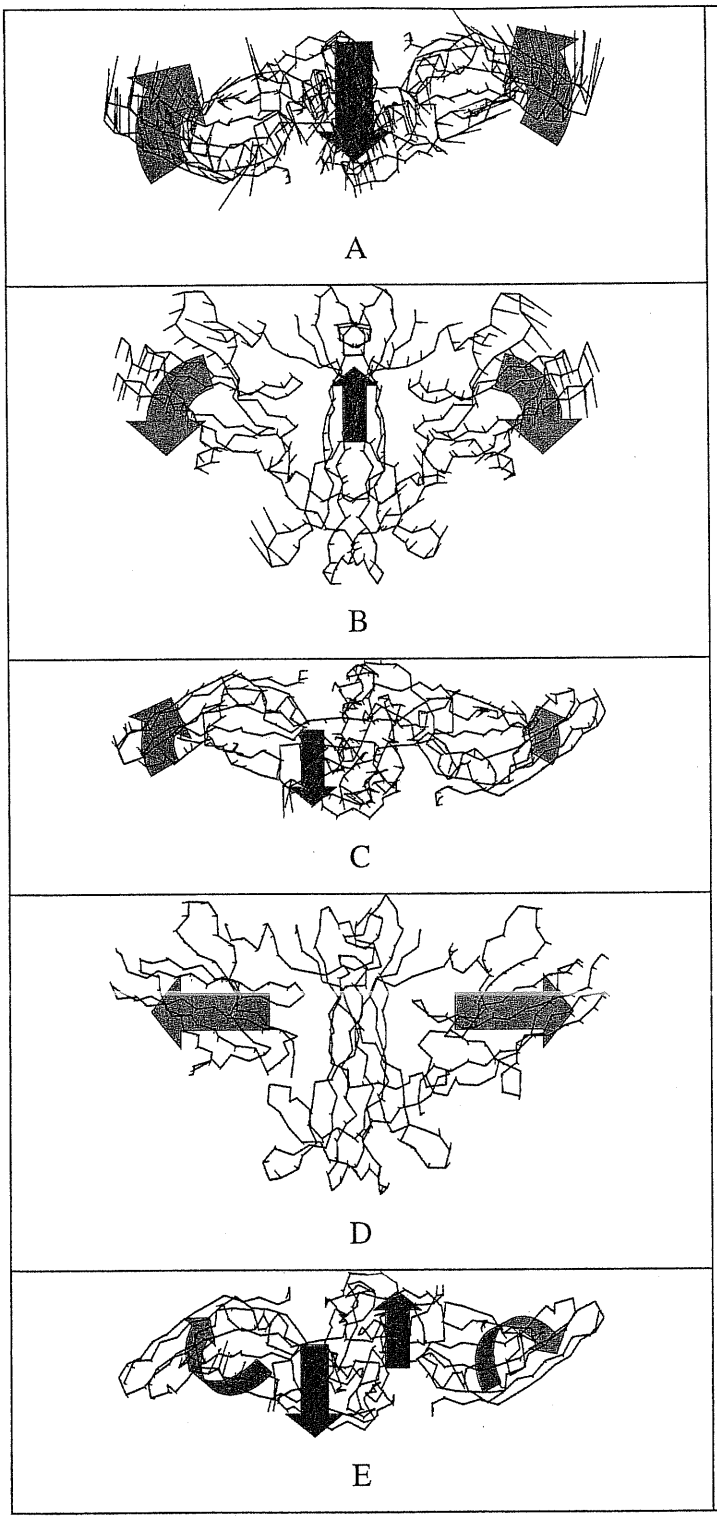


Figure 9. The first 5 eigenvectors of the covariance matrix. They are represented as vector fields on the $C_\alpha$ atoms of the average structure of the complex. The red-ended stick departing from every $C_\alpha$ atom is proportional (both in length and direction) to the component of the eigenvector along that atom. (A), (B), (C), (D) and (E) represent the 1st, 2nd, 3rd, 4th and 5th eigenvector respectively. Big arrows under the structures pictorially represent the sum of the motion of the amino acids of each domain. Blue arrows are referred to NGF and green arrows to the TrkA domains.

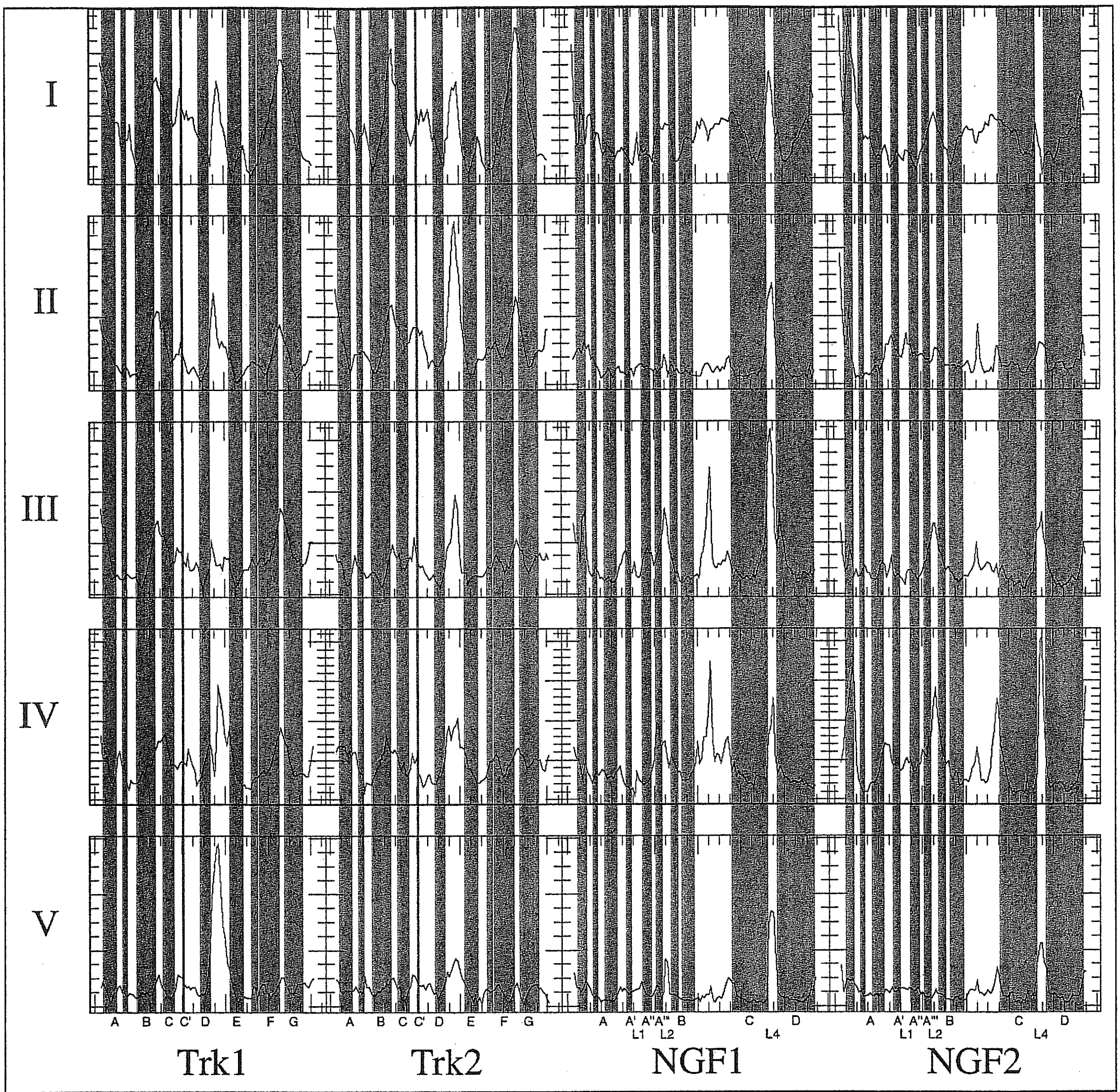


Figure 10. The components of the principal eigenvectors along the amino acid sequences of each peptide chain. I, II, III, IV and V refer to the 1st, 2nd, 3rd, 4th the 5th eigenvector respectively. Blue stripes highlight amino acids in β-strand conformation; green stripes highlight amino acids in helix conformation. On the bottom of the graph the letters indicate the different secondary structure elements according to the nomenclature used in the crystallographic structure (127). Loop regions on NGF are also indicated. Is evident that the largest components of motions are located in regions where no secondary structure is present.

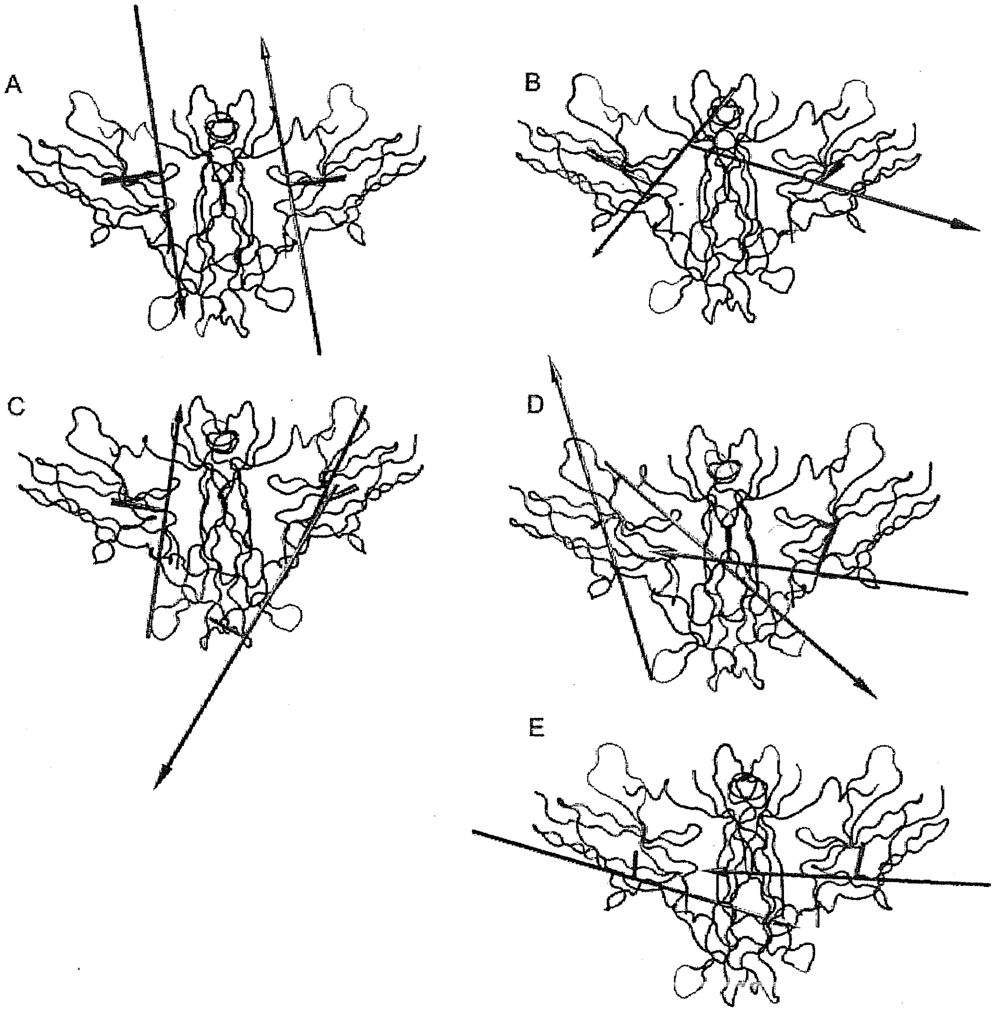


Figure 11. Domains of motion and relative hinge axes according the principal eigenvectors. These computations have been obtained through the program *Hingefind* (144). (A), (B), (C), (D) and (E) are related to the 1st, 2nd, 3rd, 4th and 5th eigenvector respectively. Different colors represent different domain of motions. Colored arrows represent the screw axes that characterize the motion of the domain of the same color with respect to the central domain (in blue). For the first two eigenvectors the domains of motion are simply related to the different protein chains. For the others eigenvectors a more complex description is needed and other domains appear.

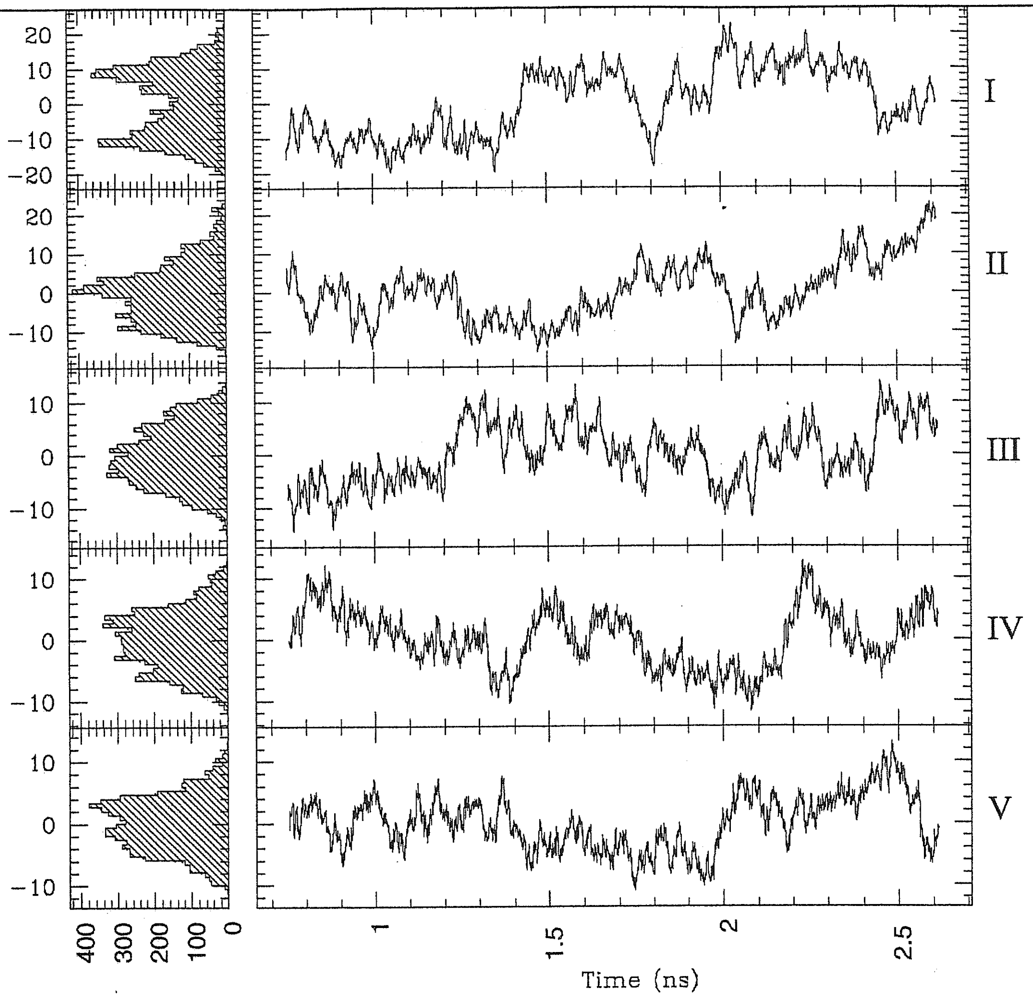


Figure 12. Projections of the trajectory along the principal eigenvectors. I, II, III, IV and V refer to the 1st, 2nd, 3rd, 4th the 5th eigenvector respectively. On the right, the projections are presented as a function of time, on the left they are binned in histograms. All the time dependent graphs show a non-monotonic behavior that indicates a good sampling of the conformational space and the absence of slow drifts in the simulation. Interestingly, the binning of the projection along the first eigenvector shows a double peak, maybe related to two well-separated free energy minima.

## ***Discussion***

Structural properties of the complex revealed by X-ray crystallography (127) are maintained during the MD simulations. The differences between the two interfaces are also maintained. Those differences involve mainly charged residues but do not elicit dramatic changes in the dynamic behavior of the system.

The binding of NGF homodimer to TrkA receptor presents intriguing characteristics. A huge non-specific electrostatic (charge-charge and charge-dipole) interaction is present (fig. 5). On the other hand, many mutations that affect binding are concentrated in correspondence of the *common* or the *specific patch* (fig. 13). The electrostatic analysis can explain the relevance of some important residues undergoing mutations (HIS 4 on NGF and MET 296, HIS 297 and GLN 350 on TrkA) and in particular seems to prove that electrostatic interactions are important for the contacts in the *specific patch* and in the first part of strand A of NGF, where residues stabilizing electrostatics are concentrated (fig. 13). On the other hand, the data on water mediated interaction, the non-specific electrostatic interactions, along with the overall low correlation between residues stabilizing the electrostatics and residues important for binding, may support the idea that several non-specific interactions may be strong enough to neutralize the effects of relatively large electrostatic changes introduced in mutagenesis experiments.

Our calculations show that the amino acids most involved in the collective fluctuations of the system partly correspond to those that affect binding upon mutation. The set of residues close to the hinge axes partially overlap the residues that are in contact at interface. Thus, flexibility of the structure in the contact region may concur to the importance of its residues for the binding. Furthermore, other important mutations are unrelated to the contacts between the two subunits while they seem to be related to the flexibility of the structure because they are close to the hinge axes. Thus, the mobility of the complex, too, may play a role in its functionality.

Summarizing our results, we have clarified the structural and functional role of amino acids playing a crucial role for NGF/TrkA binding. We have also offered an explanation for the influence of some mutations on binding, which does not emerge from visual inspection of the X-ray structure. In particular several of this mutations involve amino acids that critically participate in the collective motions of the complex. Thus, we have provided evidence that the conformational fluctuations, which strictly depend on the native structure, affect the molecular recognition of this complex.

Furthermore, this work may provide the basis for the design of new peptides that mimic NGF in TrkA binding function. These should exploit the favorable electrostatic interactions present at the *specific patch* and at the beginning of NGF strand A. This approach

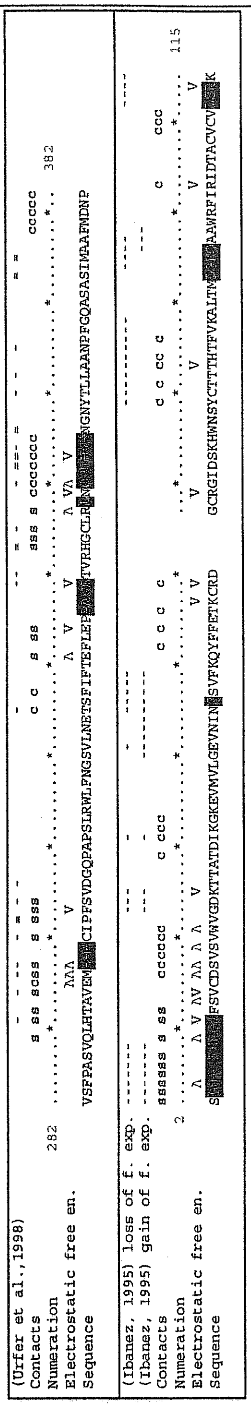


Figure 13. The sequence of TrkA (box A) and of NGF (box B) where residues have been highlighted according to several factors. (A) In the first row the sites that affect binding upon mutation have been marked with a "-". Sites that largely affect binding ($IC_{50}$ >100) have been marked with a "=". On the central row, sites belonging to the *specific patch* have been marked with an "s", while sites belonging to the *common patch* have been marked with a "c". On the 4th row amino acid with very favorable (Λ) and very unfavorable (V) electrostatic interactions are marked. On the row containing the sequence of the protein TrkA, sites close to the hinge axes have been marked with black background. (B) Same as (A) for NGF ligand. In the 1st and 2nd row relevant mutations found with gain of function and loss of function experiments respectively are marked.

would specifically address the NGF-TrkA interactions, without affecting the function of other neurotrophin receptors (TrkB, TrkC). The drawback of this approach is in the fact that other factors seem to overcome detailed electrostatics in the determination of the binding between NGF and TrkA except for the *specific patch*.

# Conclusions

The work carried out in this thesis has been mainly oriented towards the analysis of the role that the geometrical features of the proteins exert in protein folding and dynamics. This analysis has been directed with particular emphasis to specific biological issues whose phenomenology is of considerable interest. The results have been supported by parallel investigations that have supplied both a validation to, and the range of applicability of, the results themselves.

In Chapter 2 the role played by the geometric position of each amino acid in the folding process of the immunoglobulin variable domain (Ig) has been identified and measured through Molecular Dynamics (MD) simulations of models based on the topology of its native state. This measure allowed identifying the parts of the protein that, for geometrical reasons, when mutated, would result in relevant protein stability changes. Simulations have been performed without considering the covalent disulphide bond present in most of the Ig domains. The results are in good agreement with site directed mutagenesis experiments on the folding of intracellular antibodies in which the disulphide bond does not form. We have also found agreement with data on amino acid conservation in the immunoglobulin variable domain sequences. Our analysis indicates a new way for a rational approach to the design of intra-cellular antibodies more resistant to the suppression of the disulfide bond that occurs in the cytoplasm.

The study has been then extended to effective intracellular antibody sequences selected by our experimental collaborators through the Intracellular Trap Technology (ITT) (48). We have identified a remarkable characteristic shared by these proteins. Actually, a large subset of amino acids is conserved in the sequences selected by ITT. The conserved amino acids coincide with the amino acids most frequently present in the same positions (consensus sequence) of generic Ig sequences collected from standard databases. We have demonstrated that the closeness of the intrabody sequences to the Ig consensus sequence is a remarkable feature only shared by very few other sequences in standard databases. Thanks to this finding, we have been able to characterize the selection mechanism operated by ITT, which act as a sort of filter in the space of Ig sequences. The rationalization of this mechanism is of fundamental importance for the building up of antibody libraries optimized for intracellular expression and, as a consequence, for the development of effective probing procedures in functional genomics .

In Chapter 3 we have analyzed the topological characteristics of the putative cellular form of prion protein ($PrP^C$), that is the protein directly involved in the emergence of prion disease. We have used a model very close to that used in Chapter 1 to measure the contribution of each residue to the folding transition exclusively associated to the position that it has in the native conformation. We have verified that this measure is related to the propensities of the amino acids to elicit, when mutated, the $PrP^C$ degeneration into the disease-associated form ($PrP^{Sc}$). The intriguing folding characteristics of the prion protein led us to carry out a more detailed analysis of the folding pathways of $PrP^C$. In this kind of analysis we have compared the folding routes favored by $PrP^C$ topology to those preferential for doppel protein (Dpl) topology. Dpl shares with prion a very similar structure, 25% sequence identity, but does not show a prion-like degeneration. We have found that, notwithstanding a similar topology, the folding itineraries chosen by Dpl and $PrP^C$, and only dictated by their topology, differ. While for Dpl a precise sequence of folding events is observed, $PrP^C$ seems to follow several alternatives. The alternative routes can justify a scenario where, after the partial formation of the C-terminal part of $PrP^C$, comprising Helix 2 and 3, the N-terminal portion, still unstructured, is exposed to a possible rearrangement. This route is not seen in Dpl. Indeed, experimental studies indicate in the N-terminal part (residues 90-145) one of the principal determinants of the $PrP^C$ conversion to $PrP^{Sc}$ (124). Furthermore, our picture is supported by experiments showing that the conversion between $PrP^C$ and $PrP^{Sc}$ is favored by an unfolding-refolding cycle of $PrP^C$ (118). The scenario that we propose is only based on the topological characteristics of the systems and might change with the introduction of detailed chemical interactions. However, an experimental confirmation of this hypothesis would not be surprising because, as verified from different points of view in this thesis, in (9) and in recent works on the role played by topology in determining folding rates (8) and folding transition state ensembles (19), chemical interactions in proteins seem optimized in order to avoid frustration. Instead, a test of our hypothesis, along with analogous studies performed on different systems, would help to understand the limits of the topological approach to protein folding developed in this work.

Finally in Chapter 4 we have conducted an analysis at atomic detail of the dynamical characteristics of the complex between the Nerve Growth Factor (NGF) and the Tyrosine Kinase A receptor (TrkA), whose binding reaction is involved in neuronal cell growth and differentiation. The analysis of the dynamical characteristics of this system, performed through a 2.6ns MD simulation in explicit water, has been complemented by a study of the electrostatic interactions in the binding between the two components. We have found that the electrostatic features of the complex can only partly explain its binding determinants. Indeed, amino acids important for the binding because of the network of their electrostatic interactions seem to be concentrated in one part of the contact surface (SP) between NGF homodimer and TrkA. SP is mainly specific of this ligand-receptor binding reaction, while the remaining part has many features in common with the binding reaction of other neurotrophins to their tyrosine kinase receptors. We have also noticed that the contribution to the electrostatic binding from non-specific sources, like overall charge and dipole interactions and water-mediated interactions, is very large and could overcome or dump the effects of specific electrostatic interactions. On the other hand, the analysis of the collective motions that characterize the complex, has led to the identification of distinct domain of motions of the amino acids and of the hinge axes around which the motions occur. We have found a relationship between amino acids present at the hinge points and part of the mutations that mainly affect binding. Thus, our analysis has allowed identifying a possible strategy for the design of NGF-mimic peptides, where electrostatic interactions at SP are optimized. Furthermore, we have proposed a possible functional role in the motions of the complex, mainly dependent on the native conformations of its constituent proteins, and, as a consequence, a possible way to its inhibition.

# Acknowledgments

I would like to thank my supervisor Prof. Amos Maritan for having given me the opportunity to work on these fascinating topics and Prof. Paolo Carloni with whom I have collaborated in the last year.

I would like to thank Prof. Antonino Cattaneo and Dr. Michela Visintin for their fundamental collaboration in the work on intracellular antibodies and for having introduced me to the experimental aspects of protein expression and analysis.

I would like to thank Dr. Hoang X. Trinh for his essential collaboration and Dr. Critstian Micheletti for help in the analysis of the results in the work on prions.

# Reference List

1. Anfinsen, C. B. (1973) *Science* **181**, 223-230.

2. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4**, 10-19.

3. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (2000) *Reviews of Modern Physics* **72**, 259-314.

4. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Fold. Des* **1** , 441-450.

5. Maritan, A., Micheletti, C. & Banavar, J. R. (2000) *Physical Review Letters* **84**, 3009-3012.

6. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., I & Baker, D. (1999) *Nat. Struct. Biol.* **6**, 1016-1024.

7. Martinez, J. C. & Serrano, L. (1999) *Nat. Struct. Biol.* **6**, 1010-1016.

8. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277**, 985-994.

9. Cecconi, F., Micheletti, C., Carloni, P. & Maritan, A. (2001) *Proteins* **43**, 365-372.

10. Settanni, G. The role of topology in immunoglobulin variable domain. Protein Folding and Evolution. (course CXLV). 2000. Varènna, International School of Physics "Enrico Fermi".

11. Mirny, L. A. & Shakhnovich, E. I. (1999) *J. Mol. Biol.* **291**, 177-196.

12. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224**, 771-782.

13. Wriggers, W. & Schulten, K. (1998) *Biophys. J.* **75**, 646-661.

14. Piana, S. Ab initio molecular dynamics on HIV-1 protease. 2000. SISSA.
Ref Type: Thesis/Dissertation

15. Visintin, M., Settanni, G., Maritan, A., Graziosi, S., Marks, J. D. & Cattaneo, A. (2001) *submitted to J. Mol. Biol.*

16. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. U. S. A* **84**, 7524-7528.

17. Go, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 183-210.

18. Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. (1999) *Physical Review Letters* **82**, 3372-3375.

19. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) *J. Mol. Biol.* **298**, 937-953.

20. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. U. S. A* **89**, 8721-8725.

21. Settanni, G., Cattaneo, A. & Maritan, A. Role of Native-State Topology in the Stabilization of Intracellular Antibodies. Biophysical Journal . 2001.
Ref Type: In Press

22. Vendruscolo, M., Najmanovich, R. & Domany, E. (2000) *Proteins* **38**, 134-148.

23. Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995) *Computer Physics Communications* **91**, 1-41.

24. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *Journal of Computational Chemistry* **4**, 187-217.

25. Leach, A. R. (1996) *Molecular modelling. Principles and applications* (Addison Wesley, Singapore).

26. Allen, M. P. & Tildesley, D. J. (1989) *Computer simulation of liquids* (Clarendon Press, Oxford England).

27. Prusiner, S. B., Groth, D. F., Bolton, D. C., Kent, S. B. & Hood, L. E. (1984) *Cell* **38**, 127-134.

28. Berendsen, H. J., Postma, J. P. M., van Gunsteren, W. F., Di Nola, A. & Haak, J. R. (1984) *J. Chem. Phys.* **81**, 3684-3690.

29. Ferrenberg, A. M. & Swendsen, R. H. (1989) *Physical Review Letters* **63**, 1195-1198.

30. Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S. M., Lee, T., Pope, S. H., Riordan, G. S. & Whitlow, M. (1988) *Science* **242**, 423-426.

31. Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M. S., Novotny, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E. & Crea, R. (1988) *Proc. Natl. Acad. Sci. U. S. A* **85**, 5879-5883.

32. Biocca, S., Neuberger, M. S. & Cattaneo, A. (1990) *EMBO J.* **9**, 101-108.

33. Biocca, S., Ruberti, F., Tafani, M., Pierandrei-Amaldi, P. & Cattaneo, A. (1995) *Biotechnology (N. Y. )* **13**, 1110-1115.

34. Richardson, J. H. & Marasco, W. A. (1995) *Trends Biotechnol.* **13**, 306-310.

35. Biocca, S., Pierandrei-Amaldi, P. & Cattaneo, A. (1993) *Biochem. Biophys. Res. Commun.* **197**, 422-427.

36. Cardinale, A., Lener, M., Messina, S., Cattaneo, A. & Biocca, S. (1998) *FEBS Lett.* **439**, 197-202.

37. Tavladoraki, P., Benvenuto, E., Trinca, S., De Martinis, D., Cattaneo, A. & Galeffi, P. (1993) *Nature* **366**, 469-472.

38. Duan, L., Zhang, H., Oakes, J. W., Bagasra, O. & Pomerantz, R. J. (1994) *Hum. Gene Ther.* **5**, 1315-1324.

39. Mhashilkar, A. M., Bagley, J., Chen, S. Y., Szilvay, A. M., Helland, D. G. & Marasco, W. A. (1995) *EMBO J.* **14**, 1542-1551.

40. Gargano, N. & Cattaneo, A. (1997) *J. Gen. Virol.* **78 ( Pt 10)**, 2591-2599.

41. Marasco, W. A. (1995) *Immunotechnology.* **1**, 1-19.

42. Cattaneo, A. & Biocca, S. (1997) *Intracellular Antibodies: Development and Applications* (Springer, New York).

43. Gilbert, H. F. (1990) *Adv. Enzymol. Relat. Areas Mol. Biol.* **63**, 69-172.

44. Hwang, C., Sinskey, A. J. & Lodish, H. F. (1992) *Science* **257**, 1496-1502.

45. Johnson, G. & Wu, T. T. (2000) *Nucleic Acids Res.* **28**, 214-218.

46. Williams, A. F. & Barclay, A. N. (1988) *Annu. Rev. Immunol.* **6**, 381-405.

47. Martineau, P., Jones, P. & Winter, G. (1998) *J. Mol. Biol.* **280**, 117-127.

48. Visintin, M., Tse, E., Axelson, H., Rabbitts, T. H. & Cattaneo, A. (1999) *Proc. Natl. Acad. Sci. U. S. A* **96**, 11723-11728.

49. Goto, Y. & Hamaguchi, K. (1979) *J. Biochem. (Tokyo)* **86**, 1433-1441.

50. Frisch, C., Kolmar, H. & Fritz, H. J. (1994) *Biological Chemistry Hoppe-Seyler* **375**, 353-356.

51. Cattaneo, A. & Biocca, S. (1999) *Trends Biotechnol.* **17**, 115-121.

52. Jung, S. & Pluckthun, A. (1997) *Protein Eng* **10**, 959-966.

53. Rudikoff, S. & Pumphrey, J. G. (1986) *Proc. Natl. Acad. Sci. U. S. A* **83**, 7875-7878.

54. Proba, K., Honegger, A. & Pluckthun, A. (1997) *J. Mol. Biol.* **265**, 161-172.

55. Proba, K., Worn, A., Honegger, A. & Pluckthun, A. (1998) *J. Mol. Biol.* **275**, 245-253.

56. Worn, A. & Pluckthun, A. (1998) *Biochemistry* **37**, 13120-13127.

57. Worn, A. & Pluckthun, A. (1999) *Biochemistry* **38**, 8739-8750.

58. Visintin, M. The development of the Intrabody Trap Technology (ITT) for Functional Genomics. 2000. International School for Advanced Studies.
Ref Type: Thesis/Dissertation

59. Nieba, L., Honegger, A., Krebber, C. & Pluckthun, A. (1997) *Protein Eng* **10**, 435-444.

60. Freire, E., Murphy, K. P., Sanchez-Ruiz, J. M., Galisteo, M. L. & Privalov, P. L. (1992) *Biochemistry* **31**, 250-256.

61. Kolmar, H., Frisch, C., Gotze, K. & Fritz, H. J. (1995) *J. Mol. Biol.* **251**, 471-476.

62. Langedijk, A. C., Honegger, A., Maat, J., Planta, R. J., van Schaik, R. C. & Pluckthun, A. (1998) *J. Mol. Biol.* **283**, 95-110.

63. Improta, S., Politou, A. S. & Pastore, A. (1996) *Structure.* **4**, 323-337.

64. Niggemann, M. & Steipe, B. (2000) *J. Mol. Biol.* **296**, 181-195.

65. Gunasekaran, K., Eyles, S. J., Hagler, A. T. & Gierasch, L. M. (2001) *Curr. Opin. Struct. Biol.* **11**, 83-93.

66. Jung, S., Honegger, A. & Pluckthun, A. (1999) *J. Mol. Biol.* **294**, 163-180.

67. Nieba, L., Honegger, A., Krebber, C. & Pluckthun, A. (1997) *Protein Eng* **10**, 435-444.

68. Pluckthun, A. (1990) *Nature* **347**, 497-498.

69. Spada, S., Honegger, A. & Pluckthun, A. (1998) *J. Mol. Biol.* **283**, 395-407.

70. Fields, S. & Song, O. (1989) *Nature* **340**, 245-246.

71. Martin, A. C. (1996) *Proteins* **25**, 130-133.

72. Deret, S., Maissiat, C., Aucouturier, P. & Chomilier, J. (1995) *Comput. Appl. Biosci.* **11**, 435-439.

73. Wirtz, P. & Steipe, B. (1999) *Protein Sci.* **8**, 2245-2250.

74. Meggendorfer, F. (1930) *Z. Gesamte Neurol. Psychiatr.* **128**, 337-341.

75. Stender, A. (1930) *Z. Gesamte Neurol. Psychiatr.* **128**, 528-543.

76. Hadlow, W. J. (1959) *Lancet* **ii**, 289-290.

77. Klatzo, I., Gajdusek, D. C. & Zigas, V. (1959) *Lab. Invest.* **8**, 799-847.

78. Zlotnik, I. & Stamp, J. L. (1961) *World Neurology* **2**, 895-907.

79. Farquhar, C. F., Somerville, R. A. & Bruce, M. E. (1998) *Nature* **391**, 345-346.

80. Alper, T., Cramp, W. A., Haig, D. A. & Clarke, M. C. (1967) *Nature* **214**, 764-766.

81. Prusiner, S. B. (1982) *Science* **216**, 136-144.

82. Bolton, D. C., McKinley, M. P. & Prusiner, S. B. (1982) *Science* **218**, 1309-1311.

83. Basler, K., Oesch, B., Scott, M., Westaway, D., Walchli, M., Groth, D. F., McKinley, M. P., Prusiner, S. B. & Weissmann, C. (1986) *Cell* **46**, 417-428.

84. Pan, K. M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A., Groth, D., Mehlhorn, I., Huang, Z. W., Fletterick, R. J., Cohen, F. E. *et al.* (1993) *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10962-10966.

85. Pergami, P., Jaffe, H. & Safar, J. (1996) *Analytical Biochemistry* **236**, 63-73.

86. Doh-ura, K., Tateishi, J., Sasaki, H., Kitamoto, T. & Sakaki, Y. (1989) *Biochem. Biophys. Res. Commun.* **163**, 974-979.

87. Goldgaber, D., Goldfarb, L. G., Brown, P., Asher, D. M., Brown, W. T., Lin, S., Teener, J. W., Feinstone, S. M., Rubenstein, R., Kascsak, R. J. *et al.* (1989) *Exp. Neurol.* **106**, 204-206.

88. Kretzschmar, H. A., Honold, G., Seitelberger, F., Feucht, M., Wessely, P., Mehraein, P. & Budka, H. (1991) *Lancet* **337**, 1160.

89. Hsiao, K., Baker, H. F., Crow, T. J., Poulter, M., Owen, F., Terwilliger, J. D., Westaway, D., Ott, J. & Prusiner, S. B. (1989) *Nature* **338**, 342-345.

90. Gabizon, R., Rosenmann, H., Meiner, Z., Kahana, I., Kahana, E., Shugart, Y., Ott, J. & Prusiner, S. B. (1993) *Am. J. Hum. Genet.* **53**, 828-835.

91. Dlouhy, S. R., Hsiao, K., Farlow, M. R., Foroud, T., Conneally, P. M., Johnson, P., Prusiner, S. B., Hodes, M. E. & Ghetti, B. (1992) *Nat. Genet.* **1**, 64-67.

92. Petersen, R. B., Tabaton, M., Berg, L., Schrank, B., Torack, R. M., Leal, S., Julien, J., Vital, C., Deleplanque, B., Pendlebury, W. W. *et al.* (1992) *Neurology* **42**, 1859-1863.

93. Poulter, M., Baker, H. F., Frith, C. D., Leach, M., Lofthouse, R., Ridley, R. M., Shah, T., Owen, F., Collinge, J., Brown, J. *et al.* (1992) *Brain* **115 ( Pt 3)**, 675-685.

94. Stahl, N., Borchelt, D. R., Hsiao, K. & Prusiner, S. B. (1987) *Cell* **51**, 229-240.

95. Collinge, J., Whittington, M. A., Sidle, K. C. L., Smith, C. J., Palmer, M. S., Clarke, A. R. & Jefferys, J. G. R. (1994) *Nature* **370**, 295-297.

96. Montrasio, F., Frigg, R., Glatzel, M., Klein, M. A., Mackay, F., Aguzzi, A. & Weissmann, C. (2000) *Science* **288**, 1257-1259.

97. True, H. L. & Lindquist, S. L. (2000) *Nature* **407**, 477-483.

98. Viles, J. H., Cohen, F. E., Prusiner, S. B., Goodin, D. B., Wright, P. E. & Dyson, H. J. (1999) *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2042-2047.

99. Tobler, I., Gaus, S. E., Deboer, T., Achermann, P., Fischer, M., Rulicke, T., Moser, M., Oesch, B., McBride, P. A. & Manson, J. C. (1996) *Nature* **380**, 639-642.

100. Calzolai, L., Lysek, D. A., Guntert, P., von Schroetter, C., Riek, R., Zahn, R. & Wuthrich, K. (2000) *Proc. Natl. Acad. Sci. U. S. A* **97**, 8340-8345.

101. Lopez, G. F., Zahn, R., Riek, R. & Wuthrich, K. (2000) *Proc. Natl. Acad. Sci. U. S. A* **97**, 8334-8339.

102. Liu, A., Riek, R., Wider, G., von Schroetter, C., Zahn, R. & Wuthrich, K. (2000) *J. Biomol. NMR* **16**, 127-138.

103. Zahn, R., Liu, A., Luhrs, T., Riek, R., von Schroetter, C., Lopez, G. F., Billeter, M., Calzolai, L., Wider, G. & Wuthrich, K. (2000) *Proc. Natl. Acad. Sci. U. S. A* **97**, 145-150.

104. James, T. L., Liu, H., Ulyanov, N. B., Farr-Jones, S., Zhang, H., Donne, D. G., Kaneko, K., Groth, D., Mehlhorn, I., Prusiner, S. B. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A* **94**, 10086-10091.

105. Shmerling, D., Hegyi, I., Fischer, M., Blattler, T., Brändner, S., Gotz, J., Rulicke, T., Flechsig, E., Cozzio, A., von Mering, C. *et al.* (1998) *Cell* **93**, 203-214.

106. Donne, D. G., Viles, J. H., Groth, D., Mehlhorn, I., James, T. L., Cohen, F. E., Prusiner, S. B., Wright, P. E. & Dyson, H. J. (1997) *Proceedings of the National Academy of Sciences of the United States of America* **94**, 13452-13457.

107. Peretz, D., Williamson, R. A., Matsunaga, Y., Serban, H., Pinilla, C., Bastidas, R. B., Rozenshteyn, R., James, T. L., Houghten, R. A., Cohen, F. E. *et al.* (1997) *J. Mol. Biol.* **273**, 614-622.

108. Turk, E., Teplow, D. B., Hood, L. E. & Prusiner, S. B. (1988) *Eur. J. Biochem.* **176**, 21-30.

109. Muramoto, T., Scott, M., Cohen, F. E. & Prusiner, S. B. (1996) *Proceedings of the National Academy of Sciences of the United States of America* **93**, 15457-15462.

110. Caughey, B. & Raymond, G. J. (1991) *Journal of Biological Chemistry* **266**, 18217-18223.

111. Borchelt, D. R., Taraboulos, A. & Prusiner, S. B. (1992) *Journal of Biological Chemistry* **267**, 16188-16199.

112. Taraboulos, A., Scott, M., Semenov, A., Avraham, D., Laszlo, L. & Prusiner, S. B. (1995) *Journal of Cell Biology* **129**, 121-132.

113. Vey, M., Pilkuhn, S., Wille, H., Nixon, R., Dearmond, S. J., Smart, E. J., Anderson, R. G. W., Taraboulos, A. & Prusiner, S. B. (1996) *Proceedings of the National Academy of Sciences of the United States of America* **93**, 14945-14949.

114. Kaneko, K., Vey, M., Scott, M., Pilkuhn, S., Cohen, F. E. & Prusiner, S. B. (1997) *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2333-2338.

115. Naslavsky, N., Stein, R., Yanai, A., Friedlander, G. & Taraboulos, A. (1997) *Journal of Biological Chemistry* **272**, 6324-6331.

116. Scott, M., Groth, D., Foster, D., Torchia, M., Yang, S. L., Dearmond, S. J. & Prusiner, S. B. (1993) *Cell* **73**, 979-988.

117. Telling, G. C., Scott, M., Hsiao, K. K., Foster, D., Yang, S. L., Torchia, M., Sidle, K. C. L., Collinge, J., Dearmond, S. J. & Prusiner, S. B. (1994) *Proceedings of the National Academy of Sciences of the United States of America* **91**, 9936-9940.

118. Baskakov, I. V., Legname, G., Prusiner, S. B. & Cohen, F. E. (2001) *J. Biol. Chem.* **276**, 19687-19690.

119. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999) *J. Mol. Biol.* **290**, 253-266.

120. Prusiner, S. B. (1998) *Proc. Natl. Acad. Sci. U. S. A* **95**, 13363-13383.

121. Alonso, D. O., Dearmond, S. J., Cohen, F. E. & Daggett, V. (2001) *Proc. Natl. Acad. Sci. U. S. A* **98**, 2985-2989.

122. Mo, H., Moore, R. C., Cohen, F. E., Westaway, D., Prusiner, S. B., Wright, P. E. & Dyson, H. J. (2001) *Proc. Natl. Acad. Sci. U. S. A* **98**, 2352-2357.

123. Jackson, G. S., Hill, A. F., Joseph, C., Hosszu, L., Power, A., Waltho, J. P., Clarke, A. R. & Collinge, J. (1999) *Biochim. Biophys. Acta* **1431**, 1-13.

124. Zhang, H., Kaneko, K., Nguyen, J. T., Livshits, T. L., Baldwin, M. A., Cohen, F. E., James, T. L. & Prusiner, S. B. (1995) *J. Mol. Biol.* **250**, 514-526.

125. Yuen, E. C. & Mobley, W. C. (1995) *Molecular Medicine Today* **1**, 278-286.

126. Bothwell, M. (1995) *Annual Review of Neuroscience* **18**, 223-253.

127. Wiesmann, C., Ultsch, M. H., Bass, S. H. & de Vos, A. M. (1999) *Nature* **401**, 184-188.

128. Ibanez, C. F., Ebendal, T., Barbany, G., Murrayrust, J., Blundell, T. L. & Persson, H. (1992) *Cell* **69**, 329-341.

129. Ibanez, C. F., Ilag, L. L., Murrayrust, J. & Persson, H. (1993) *Embo Journal* **12**, 2281-2293.

130. Urfer, R., Tsoulfas, P., Soppet, D., Escandon, E., Parada, L. F. & Presta, L. G. (1994) *Embo Journal* **13**, 5896-5909.

131. Kahle, P., Burton, L. E., Schmelzer, C. H. & Hertel, C. (1992) *Journal of Biological Chemistry* **267**, 22707-22710.

132. Drinkwater, C. C., Barker, P. A., Suter, U. & Shooter, E. M. (1993) *Journal of Biological Chemistry* **268**, 23202-23207.

133. Ilag, L. L., Lonnerberg, P., Persson, H. & Ibanez, C. F. (1994) *Journal of Biological Chemistry* **269**, 19941-19946.

134. Kullander, K. & Ebendal, T. (1994) *Journal of Neuroscience Research* **39**, 195-210.

135. Urfer, R., Tsoulfas, P., O'Connell, L., Hongo, J. A., Zhao, W. & Presta, L. G. (1998) *J. Biol. Chem.* **273**, 5829-5840.

136. Bax, B., Blundell, T. L., Murray-Rust, J. & McDonald, N. Q. (1997) *Structure.* **5**, 1275-1285.

137. Ibanez, C. F. (1995) *Trends Biotechnol.* **13**, 217-227.

138. LeSauteur, L., Cheung, N. K. V., Lisbona, R. & Saragovi, H. U. (1996) *Nature Biotechnology* **14**, 1120-1122.