Self-consistent field approach to protein structure and stability

Promotoren: Dr. N.C.M. Laane
hoogleraar in de Biochemie

Dr. R.R. Crichton
professeur de Biochimie
a l'Université de Louvain-la-Neuve, België

Co-promotor: Dr. ir. J.J.M. Vervoort
wetenschappelijk medewerker bij het departement Biomoleculaire Wetenschappen, laboratorium voor Biochemie

Self-consistent field approach to protein structure and stability

Roumen Atanasov Dimitrov

Proefschrift

ter verkrijging van de graad van doctor,
op gezag van de rector magnificus
van de Landbouwuniversiteit Wageningen,
dr. C. M. Karssen,
in het openbaar te verdedigen
op dinsdag 12 januari 1999
des namiddags te half twee in de Aula.

Dimitrov, Roumen Atanasov

Self-consistent field approach to protein structure and stability

Thesis Wageningen. - With ref. - With summary in Bulgarian, English and Dutch

ISBN:90-5485-986-5

Keywords: molecular field theory/ free energy minimization/ electrostatic interactions/ protein folding/ transition state/ Φ-values/ nucleus/ topology/ protein structure prediction

Published by: Grafisch Service Centrum, Nude 40, Wageningen, The Netherlands

Propositions

1. The view that the usage of lattice models is inadequate to yield insight into real folding problems does not help to get a climpse of the Holy Grail.

   Herman J. Berendsen 1998. A Glimpse of the Holy Grail? Science, **282**: 642-643.
   This thesis

2. In contrast to the "new view" of the Bryngelson et al. about protein folding, this dissertation demonstrates that the sequence, topological and symmetrical restrictions completely determine the folding pathway of two-state small protein molecules.

   Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins Struct. Funct. Genet.* **21**:167-195.
   This thesis

3. Folding patterns and their favorable topologies are the key for the transfer of the information coded in the sequence to the information coded in the residue interactions in the 3D structure of the protein.

   This thesis

4. The ease with which one can obtain secondary structure prediction from gene-sequences is inversely related to the knowledge needed to interpret these data.

5. The game of chess is not a good analogy for protein sequence, but is a good analogy for protein folding.

   Wlodek Mandecki 1998. The game of chess and searches in protein sequence space. TIBTECH **16**: 200-202.; **Stanislaw Wronski** 1998. The game of chess is not a good analogy for protein sequences. TIBTECH **16**: 407

6. Teaching based on Computer Informatics Technology will lead to social isolation.

7. The whole point of this proposition is to express the whole point of this proposition.

8. It's not that life's too short, you're just dead for so long.

9. The geographical distribution of Nobel price winners in science is a reflection of scientific political forces.

Propositions belonging to the thesis entitled " **Self-consistent field approach to protein structure and stability**".

Roumen Atanasov Dimitrov
Wageningen, 12 January 1999.

to my parents, my wife Dilnora and my daughter Yasmina

CONTENTS

1 Introduction

Proteins on Earth today are the consequences of a complex process of biological evolution. A typical protein contains about a hundred amino acids linked in a well-defined sequence. Evolution has created many variants, some of which perform similar functions in different organisms. The common feature of these biologically active biomolecules is that under proper conditions they all fold into close-packed, typically globular-like structures. X-Ray diffraction and nuclear magnetic resonance spectroscopy methods demonstrate that the 3D structure of proteins (over 7700 coordinate sets are available now in the Brookhaven Protein Databank) are thermodynamically well-defined. Since many proteins have been found to fold *in vitro*, and the native state reached *in vitro* is the same as that reached *in vivo*, it follows that the one-dimensional information contained in the sequence of amino acids is sufficient for a protein molecule to organize itself into its folded conformation. Thus, a conclusion was made that the native structure of the protein molecule corresponds to its thermodynamically most stable state i.e. to the minimum of its free energy.[1] Christian Anfinsen in his 1972 Nobel prize acceptance lecture described the thermodynamic hypothesis of protein folding as follows:

"This hypothesis states that the 3D structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the amino acid sequence, in a given environment."

Soon after Anfinsen's thermodynamic hypothesis was stated it was shown by Cyrus Levinthal that this hypothesis implied a paradoxical result: a typical protein molecule has far too many conformations to permit a thorough search for the global minimum.[2] To overcome the Levintal paradox researchers attempted to preserve the concept of the global minimum by arguing that compactness and the secondary structure organization of the protein molecules reduce the effective size of the conformational space available to them. However, the compactness itself only lessens, but does not eliminate the conformational problem. Hence, the Levintal paradox is real, and the only way to overcome the Levintal paradox are the guiding forces engineered by molecular evolution. The first functionally important action of any protein after biosynthesis is to fold. In other words, only a tiny fraction of the total possible

conformations available to a polypeptide chain is sampled during folding and this subset of conformations may be viewed as a kinetic pathway. A unique folding pathway, if it exists, can be elucidated by experiments. But experiments do not confirm this point of view, rather that in the earlier stages of folding a protein possesses a large ensemble of structures. The problem is then not to find a single route but to characterize the dynamics of the ensemble through a statistical description of the topography of the energy landscape.[3-6] The most detailed description of the energy landscape will be obtained by specifying the free energy average over the solvent coordinates as a function of the coordinates of every atom in the protein. The free energy of the solution at a particular fixed protein conformation will be named further as an energy not to be mistaken with the free energy over different conformational states of the protein molecule. At this fine level of description, the energy surface of the protein is riddled with many local minima. Some of these minima correspond to large conformational differences and the interconversion between them can be quite slow. Not all of the conformations between the unfolded and folded states of a protein molecule are equally probable. In fact, conformations with lower energy are more likely than those with higher energy. For a protein to be *kinetically* foldable, there must be a sufficient overall slope of the energy landscape. It is because of this overall slope of the funnel, that folding occurs as a progressive organization of an ensemble of partly folded conformations along numerous parallel pathways. An important question is how these parallel pathways are connected to each other. It is clear that the pathways, being numerous at the top of the funnel, have to converge as they approach the unique conformation of the native protein. In the funnel the spontaneous tendency to fold is opposed by random thermal motions as measured by its entropy. While energy and entropy are in balance at the top and bottom of the funnel, they are not for intermediate positions. At these intermediate positions initiation of favorable energetic contacts requires the protein to first pay an entropic and energy price before the downhill tendency of the energy landscape can be manifested. The set of structures which fulfill these requirements represent the transition state (TS) for protein folding. The projection of the funnel landscape on an appropriate chosen reaction coordinate results in an energy profile where the denatured and the folded states sit at the bottom of energy wells, whereas the transition state is at a maximum. The denatured state is populated by a vast number of distinct conformations with a broad spatial distribution, while the folded state is dominated by a single conformation.

Experimental and theoretical studies strongly support the fact that at the TS level the folding toward the native conformation is based on the nucleation growth mechanism.[7,8] Therefore, the folding problem of two-state small monomeric proteins can be reduced to the question of how the folding nucleus at the transition state is formed from the ensemble of partly structured conformations in the denatured state. The aim of this thesis is to show how the sequential thermodynamic approach may be of relevance in solving the protein folding and fold recognition problems. It is shown that in the denatured state the folding is energetically favored by certain highly fluctuating nucleation regions (α-helices and(or) β-hairpins), which in experiments based on site directed mutagenesis are revealed by their high Φ-values. In the TS the folding is favored by the packing of these nucleation regions together with other portions of the polypeptide chain thus leading to a broad distribution of the Φ-values. The packing process results in a nucleus with native like-topology, approximately correctly formed secondary structures and loop regions with different degree of order.

The fold recognition strategy is a consequence of the fact that the native-like nucleus is separated from all other folding alternatives by a high free energy barrier. The calculations of the free energy of the folding nucleus are based on: 1) statistical mechanics of a linear cooperative system, and 2) a self-consistent molecular mean field theory previously developed for electrostatic interactions (see chapter 3). Molecular field theory is used to describe the long-range residue-residue interactions, while one-dimensional statistical mechanics is used to find out the pathway of the protein chain in the molecular field. The basic characteristics of the molecular field are determined by rough geometric characteristics of the native structure-packing patterns. A predicting strategy can be formulated as follows: 1) A set of thermodynamically most favorable packing patterns is constructed which has to be consistent with the length of the tested protein chains. These patterns are represented by the combinatorial set of the thermodynamically most favorable mutual positions of α- and (or) β-regular regions with definite lengths and spatial orientations; 2) A set of thermodynamically most favorable tertiary folds is defined for each packing pattern; 3) Calculations are carried out to determine the free energy of the protein chain over all available sets of tertiary folds; and 4) The tertiary fold with the minimal free energy is expected to be the same or very similar to the native one. As a final result one obtains the distribution and corresponding fluctuations of the secondary regions along the sequence and their contacts in space. The free energy minimum is obtained by a minimization procedure in which proteins "fold" from a random state by

collapsing and reconfiguration. Reconfiguration follows a general drift from higher to lower energy conformations, and reconfiguration occurs between conformations that are geometrically similar. In other words, the free energy minimum is represented as a "folding" funnel with a collection of geometrically similar collapsed structures, one of which is thermodynamically stable with respect to the rest (but not necessary with respect to the whole conformational space).

The organization of the thesis is as follows: chapter 2 presents a review of the basic physical principles that govern protein structure and focuses on the thermodynamics as well as kinetics of protein folding and unfolding. Then chapter 3 starts with a discussion on the basic elementary interactions which contribute to protein structure and stability, with emphasis on the electrostatic interactions. Electrostatic interactions are described on the basis of a novel approach which uses the idea of a self-consistent field adapted from statistical mechanics. Properties such as titration curves, protein stability and pK_a shifts are discussed. In particular it is shown that the contribution of electrostatic interactions to the stability of proteins is close to zero. Chapter 4 concerns the application of the theory of electrostatic interactions to the calculation of the pKa's of the 98 residue β-elicitin protein, cryptogein. Unusual in this protein is the existence of four ionized groups buried in the hydrophobic core. The NMR structure of the 98 residue β-elicitin, cryptogein was determined using ^{15}N and $^{13}C/^{15}N$ labelled protein samples. Calculation of theoretical pKa's show general agreement with the experimentally determined values and are similar for both the crystal and solution structures. In chapter 5 the topological requirement for nucleus formation of a two-state folding reaction is considered. The self-consistent field approach is used to calculate the free energy of the folding nucleus and to approximate the description of the elementary long-range interactions such as hydrogen bonding and hydrophobic interactions. The local interactions between residues, which are close in sequence - as in the α-, β- or loop regions - are accounted for in an explicit form based on experimental parameters. Finally, in chapter 6 the problem of protein fold recognition of small monomeric proteins with less than 80 residues is discussed. The thesis is concluded with a summary in English, Dutch and Bulgarian.

References

1 Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **47**: 1309-1314, 1961.

2 Levintal,C. How to fold graciously. In: "Mossbauer Spectroscopy in Biological Systems." Proceedings of a meeting held at Allerton house, Monticello, Illinois. De-Brunner, P., Tsibris, J., Munck, E. (Eds.). Urbana, IL: University of Illinois Press, 22-24, 1969.

3 Harrison, S. C., and Durbin, R. Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. U.S.A.* **82**: 4028-4030, 1985.

4 Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Function, Structure and Genetics* **21**: 167-195, 1995.

5 Dill, K. A., and Chan, H. S. Principles of protein folding: a perspective from simple exact models. *Protein Sci.* **4**: 561-602, 1995.

6 Sali, A., Shakhnovich, E., and Karplus, M. Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636. 1994.

7 Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

8 Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

2 Background

Protein Architecture

At the lowest level of describtion, a protein molecule is made up of amino acid residues-NHCHR$_i$CO- linked together by peptide bonds in a definite sequence. The amino acids differ only in their side groups R_i. There are 20 different side chains specified by the genetic code. Many others occur as the products of enzymatic modification of proteins after translation. The sequence of side groups determines all that is unique about a particular protein, including its biological function and its specific 3D structure.

Depending on the chemical nature of the side chain, the amino acids can be classified into a few distinct categories: **1)** Ionized groups-Asp (D), Glu (E), Arg (R), Lys (K), His (H) and Tyr (Y). At neutral pH, Asp and Glu are negatively charged, while Lys and Arg are positively charged. Tyr has an -OH group but it dissociates only at high pH. **2)** Polar but uncharged -Cys (C), Ser (S), Thr (T), Asn (N) and Gln (Q). Cys plays a special role in proteins because its -SH group allows it to dimerize through an -S-S- bond to a second Cys, thus covalently linking regions of the polypeptide. Ser and Thr have an -OH group which is able to form hydrogen bonds. Asn and Gln have polar amide groups with even more extensive hydrogen-bonding capacities. **3)** Hydrophobic groups- Ala (A), Ile (I), Leu (L), Met (M), Phe (F), Trp (W), Val (V) and Pro (P). These groups consist only of hydrocarbons, except for the sulfur atom in Met, and the nitrogen atom in Trp. The side chains of these nonpolar amino acids are only slightly soluble in water. Pro has stronger stereochemical constraints than any other residue, with only one instead of two variable backbone angles, and it lacks the normal backbone NH for hydrogen bonding. Pro can create a kink in a polypeptide chain. Thus, Pro is, in spite of its quite strong hydrophobicity, usually found at the edge of the protein. Tyr is also a hydrophobic amino acid because of its benzene ring, but its hydroxyl group allows it to interact with water. Gly, like Cys is a special amino acid. It has a hydrogen atom as its R group; thus it is the smallest amino acid and has no special hydrophobic or hydrophilic character. The structure of all amino acids, except Gly, are asymmetrically arranged around the C_α carbon atom, because this atom is bonded to four different atoms or groups of atoms (-NH_2, -COOH, -H, and -R).

Thus all amino acids, except Gly, can have one of two stereoisomeric forms. By convention, these mirror-image structures are called the D and L forms. They cannot be interconverted without breaking a chemical bond. Only the L forms of amino acids are found in proteins.

Amino acids in a protein molecule are connected by peptide bonds[1] (fig.1). In terms of the accuracy of protein structure determination, all of the bond lengths are essentially invariant. The dihedral angle ω, which characterizes the rotation around the peptide bond, is very close to 180° producing a trans - planar peptide bond with the neighboring α-carbons and the N, H, C, and O between them all lying in one plane. However, there is evidence that ω can also vary slightly.[2] Cis peptides, with $\omega = 0°$, can occur in Pro-containing peptides but essentially never for any other residue. Since the peptide units are effectively rigid groups which are linked into a chain by covalent bonds at the C_α atoms, the only degrees of freedom they have are rotations around two bonds: the $C_\alpha - C'$ and the $N - C_\alpha$ bonds. A convention has been adopted to call the angle of rotation around the $N - C_\alpha$-bond, **phi**(Φ), and the angle around the $C_\alpha - C'$ bond from the same C_α-atom, **psi**(Ψ). In this way, each amino acid residue is associated with two conformational angles Φ and Ψ.

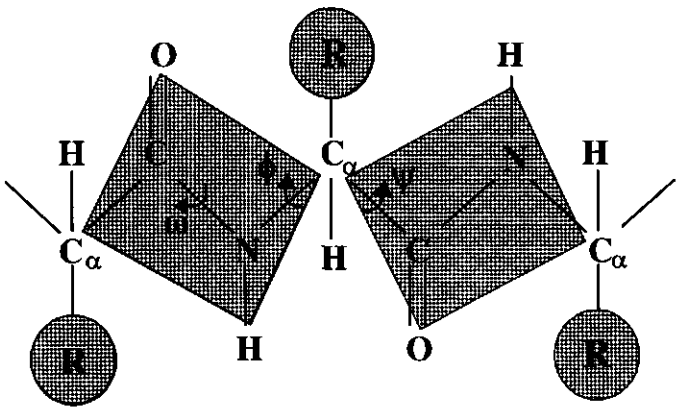


Figure-1. The peptide bond joining adjacent amino acids in a protein.

The angle pairs Φ and Ψ are usually plotted against each other in a diagram called a **Ramachandran plot**.[3] Fig.2 shows the results of such plot for Φ and Ψ taken from a large number of accurately determined protein structures. Most combinations of Φ and Ψ angles

are not allowed because of steric collisions between the side chain atoms and main chain atoms. From fig.2 it is seen that the observed values are clustered in the sterically allowed regions for the β-sheet conformations (Φ=-180/-44; Ψ=60/180), the right-handed α helical conformations (Φ=-134/-50; Ψ=-70/40) and a small region of left-handed α helical conformations (Φ=44/115; Ψ=-30/60).

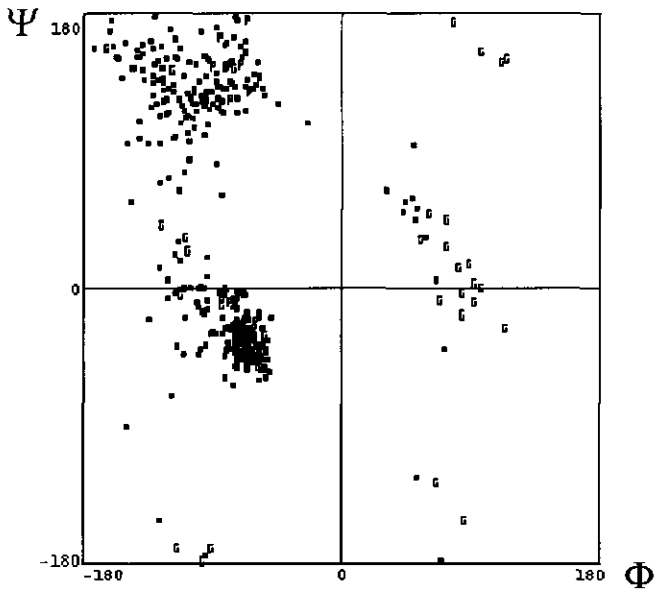


Figure-2. Example of a Ramachandran plot. The main chain dihedral angles Φ and Ψ are plotted for the phosphorylated isocitrate dehydrogenase (entry 4ICD from Protein Data Base).

In contrast to the other amino acids, Gly plays a special role because of its single hydrogen atom as a side chain. As a consequence Gly can adopt a much wider range of conformations allowing unusual main chain conformations in proteins. In proteins Gly is often required in the case where main chain atoms must approach each other very closely or in cases when pieces of backbone need to move or hinge. This is one of the main reasons why a high proportion of Gly residues is conserved among homologous protein sequences.

The α-region in the Ramachandran plot is represented by an α-helix[1] (fig.3). α-Helices in proteins are found when a stretch of consecutive residues all have the Φ, Ψ angle pair approximately around -60° and -50°, corresponding to the allowed region in the lower left quadrant of the Ramachandran plot (fig.2). A typical α-helix has 3.6 residues per turn with hydrogen bonds between $C=O$ of residue n and NH of residue $n+4$. Thus all NH and

CO groups are joined with hydrogen bonds except the first *NH* groups and the last *CO* groups at both ends of the α-helix. As a consequence, the ends of α helices are polar and are almost always at the surface of protein molecules. In proteins the α-helix is almost always right-handed. Short regions (3-5 residues) of left-handed α-helices occur occasionally.

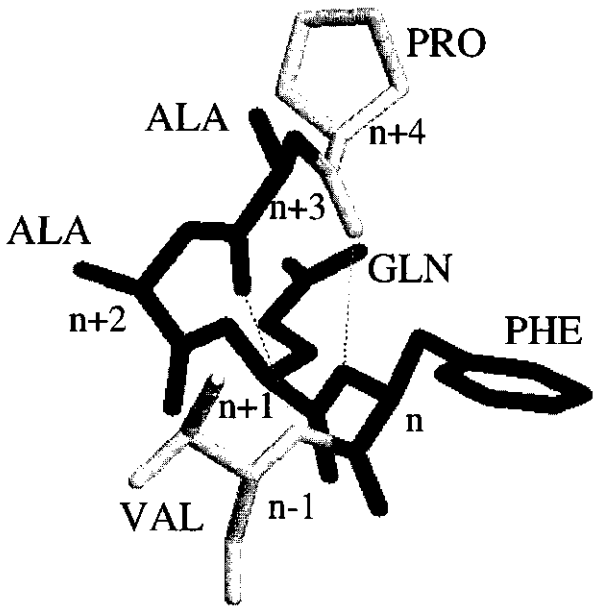


Figure-3. Schematic view of an α-helix. Only the right-handed (1-4) form of the helix is shown. H-bonds are shown as dashed lines. VAL-ALA and PHE-PRO are used as an example to show that H-bonds occur between each 4 successive amino acids.

The second major structural element found in globular proteins is the β-sheet[4] (fig.4). This structure is built up from a combination of several regions of the protein chain. These regions, β-strands, are usually from 5 to 10 residues long and have Φ, Ψ angles within the broad structurally allowed region in the upper left quadrant of the Ramachandran plot (fig.3). β-Strands are aligned adjacent to each other such that hydrogen bonds can form between *CO* groups of one β-strand and *NH* groups on an adjacent β-strand and vice versa. The β-sheets which can be formed from several such β-strands are pleated with C_α atoms successively a little above and below the plane of the β-sheet. The side chains follow this pattern such that within a β-strand they also point alternatively above and below the β-sheet. β-Strands can all run in the same direction in which case the sheet is described as parallel. In the case when the

successive strands have alternating directions, the sheet is called antiparallel. Almost all β-sheets, parallel, antiparallel, and mixed, as they occur in known protein structures have their strands twisted. This twist always has the same right-handedness.

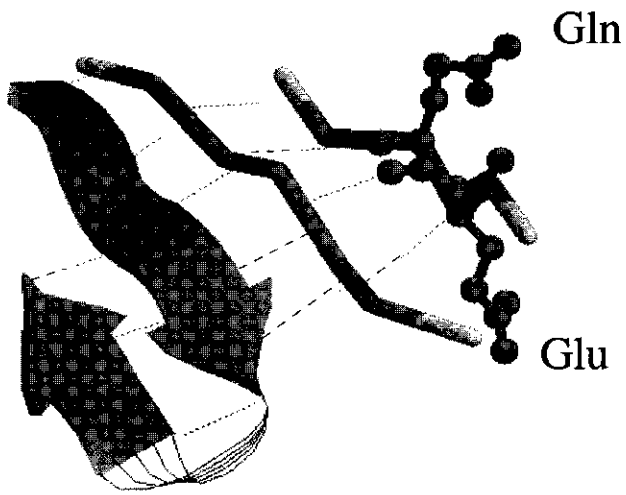


Figure-4. Example of a 4 stranded β-pleated sheet. Hydrogen bonds are indicated by dashed lines. Two adjacent β-strands are represented by arrows. Two other β-strands are represented by their backbones. The directions of amino acid side chains which are placed above and below the β-sheet, are represented by Gln and Glu as an example.

In proteins secondary structures, α -helices and β-strands are connected by loop regions of various lengths and irregular shape.[5] The loop regions are at the surface of the molecule and are rich in charged and polar hydrophilic residues. The main chain *CO* and *NH* groups of these loop regions, which in general do not form hydrogen bonds with each other, are exposed to the solvent and can form hydrogen bonds to water molecules.

Although at first sight the folded structures of proteins look complicated and even random, molecules with widely different sequences show structurally many common features. Firstly, when all atoms are included in a structural representation of a protein molecule it is clearly observed that there is very little empty space left. In other words, the extremely varied sizes and shapes of the side chains of the amino acids must all be fitted together in a very efficient packing which is comparable to that of small organic crystals[6,7] fig.5.

Figure-5. 3D-representation of a protein. The histidine-containing phosphocarrier protein from E. Coli (1poh entry from the Protein Data Base) is used an example of all atom packing restrictions on protein core formation by the side chains of the secondary structures. Only the parts of the secondary structures which contribute to the core formation are represented in atomic detail.

Secondly, hydrophobic groups tend to be buried inside the proteins forming a non-polar core, while hydrophilic side chains are nearly always exposed.[6-8] Nearly all the buried polar groups, such as the main-chain -NH and -CO groups form hydrogen bonds. Thirdly, the requirement of a structure to be compact and hydrogen bonds to be formed by buried polar groups necessitates the formation of α-helices or β-strands by a large proportion of the polypeptide chain. Major parts of α-helices and β-strands are buried, so they run across the molecule, while the interconnecting loops form few intramolecular hydrogen bonds. As a consequence these loops are positioned almost always at the surface[11] (fig.5). The requirement of close packing restricts the observed packing angles between the α-helices or/and β-strands to a limited amount of possibilities. Yet there remains sufficient room for various three-dimensional arrangements. Fourthly, amino acids are "handed" (except for glycine), and naturally occurring proteins contain only L-amino acids. As a consequence: 1) α-helices are right-handed, 2) β-strands have a right handed twist of the peptide units along their axes and a left-handed rotation between adjacent strands in the β-sheets (fig.4), 3) The connection between two parallel strands in the same sheet is almost certainly right-handed. 4) Cylindrical

sheets (that is β-barrels) are right-handed.[12-19] Lastly, most protein structures fall into five rather distinct classes,[20] which are defined according to their secondary structure content: α-proteins have only α-helices; β-proteins have mainly β-sheets; α+β proteins have α-helices and β-strands that do not mix but tend to segregate along the polypeptide chain; α/β proteins have mixed or approximately alternating segments of α-helices and β-strands; and "coil" proteins contain almost no regular secondary structures.

In the last few years one of the most exciting developments has been the realization that, although there are many protein structures, they can be classified into fold families. These fold families correspond roughly to the different patterns of super secondary structure association and their topological connections.[21-23] Comparison of the tertiary structures of homologous proteins shows that their 3D structures are conserved in evolution more than their primary structures. The amino acid sequence changes that occur over long periods of evolution lead to small structural variations: -α-helices and β-sheets shift relative to each other.[24,25] Radical insertions and deletions which lead to more extensive conformational changes tend to occur mainly in loop regions. Insertions are allowed in some β-strands to give rise to a so-called β-bulge.[2] The sequence and structural modifications are acceptable, if they maintain the stability of the protein and do not adversely affect its function. The same is true for proteins which have low sequence identities, but their structural details and, in many cases, functional features suggest that they have a common evolutionary origin. The analysis now available of the structures in the Protein Databank shows that the number of different protein folds is about 400. Of these folds, approximately 25% belong to the all α-class, 20% belong to the all β-class, 30% belong to the α/β-class, and 25% belong to the $\alpha+\beta$-class. The reason for the small number of folds is most likely to be historical. Apparently, early in evolution a wide range of general functional and catalytic properties could be realized by a relatively small number of proteins and was it easier to produce new proteins with more specific properties by the duplication, divergence and recombination of old proteins than by *ab initio* evolution.

Large and even moderately large proteins, can often be subdivided into domains which are contiguous in primary sequence and have a compact three dimensional structure. Such domains satisfy most of the architectural constraints which are valid for smaller proteins. Often they can fold independently and can be created as separate folded entities in the test tube.[26]

Finally, functioning protein molecules are often multiprotein complexes such as for example hemoglobin. Hemoglobin is not a single protein chain, but is composed of four domains which

are held together only by van der Waals interactions. This level of organization is referred to as quaternary structure.[26]

Thermodynamics and Kinetics of Folding and Unfolding of Protein Molecules

Our ideas on protein folding pathways and the properties of the native state of proteins have changed dramatically in the past few years, because of the development of new experimental and theoretical methods. Recent advances in protein engineering and NMR procedures have enabled the description of protein-folding pathways at almost atomic resolution. This allowed theory and experiments to converge, yielding the basic principles of protein structure organization and knowledge about the mechanisms of protein folding.[27-37]

The main protein properties can be classified as kinetic and thermodynamic. Thermodynamic properties of the native state include high compactness, a low energy necessary to stabilize the protein structure against the highly entropic unfolded states; low energy degeneracy, leading to a unique native structure which is important for its physiological function; a hydrophobic core in aqueous media; and abundant secondary structure. On the other hand, the kinetical properties involve a fast folding time (for the formation of the native structure under physiological conditions) and a slow unfolding time (which is important for the stability of the native state).

Studies of protein folding are usually carried out by changing environmental conditions. For example a protein molecule can be unfolded by raising or lowering of the temperature.[38-42] The latter is connected with the fact that part of the binding forces involve the entropy of the surrounding solvent. In addition, folding can be induced by changing the pH of the solution, or by adding compounds such as urea and guanidium hydrochloride. Thermodynamic characteristics such as the free energy difference between the folded and unfolded forms of the protein molecule can be studied by measuring their relative concentrations. The concentrations are determined by using probes which can distinguish between the folded and unfolded forms of the protein molecule. Thus, viscosity measurements can be used to probe whether the molecules are compact or extended. Low X-ray scattering are used to determine compactness. Probes such as visible light or fluorescence spectroscopy are sensitive to the local environment of certain amino acid residues, thus indicating the presence of secondary structures. The nuclear Overhauser effect and other NMR phenomena give information about the relative

location of two groups in a protein molecule and thus give information about the tertiary structure.[43-44] Calorimetric studies give the total energy content of a protein solution which is connected to the equilibrium constant through the Van't Hoff's law. Biological and biochemical procedures can be used for the binding of antibodies to the protein in order to determine what fraction of the molecules exhibits antigenicity. In addition, it is possible to use assays of the activity of an enzyme because the activity depends on the three dimensional environment of the active site.

The first step in understanding how proteins fold include knowledge of the rates for the elementary structural processes involved. With the availability of several new fast kinetic methods (for example nanosecond laser temperature jumps for the initiation of the folding process[45-47]), it is now possible to measure rates on any timescale of interest. Data are still rather limited, but the rates at which a number of elementary processes occur, such as the formation of the helices and loops for helical proteins or of β-hairpins and loops for β-sheet proteins, have been measured. Thus experiments based on time-resolved infrared spectroscopy for the average helix content suggest that helices may generally be expected to form in ~ 100 ns.[46,47] On the other hand, helix formation can be used to estimate the rate of turn formation. Thus modeling of helix-coil kinetic data from temperature jump experiments suggest that the rate of turn formation is ~ 500 ns. Loop-formation rates can be determined from intramolecular ligand binding studies. Using photochemical triggering and nanosecond-resolved optical spectroscopy, the time constant of the heme binding to a Met residue located about 50 residues distant from the heme along the polypeptide chain was found to be 40 μs.[48,49] Theoretical estimates[50,51] suggest that the rate of shorter loops (6-10 residues long) is approximately 10-25 faster than 50-residue loops, that is, ~ 1μs. As a result the above experimental data suggest that loop formation would be rate-limiting at least for the fastest folding proteins, i.e. approximately 1μs seems a reasonable estimate for the shortest time in which a small protein can fold.

Thermodynamical experiments showed that in most cases protein folding and unfolding involve only two thermodynamically distinct states or phases (fig.6).[52] The folded state consists of a spatially narrow distribution of structures, whereas the unfolded state contains a much larger number of distinct configurations that have a much broader spatial distribution. The state at the maximum in the free-energy profile is called the 'transition state' for protein folding. The two-state picture is also consistent with the fact that different probes, such as

viscosity or spectroscopy, indicate that the transition occurs at the same set of conditions as measured separately by each of the probes. This makes it possible to compare calorimetrically determined equilibrium constants directly with spectroscopical determined ones, and they agree.[38] One of the main experimental observations is that the free energy difference between folded and unfolded forms is often quite small, of the order of a few k_BT. The two-state thermodynamics is however not universal. There are several examples in which different probes show non-coincident transitions, such as multiple domain proteins and the molten globule state.

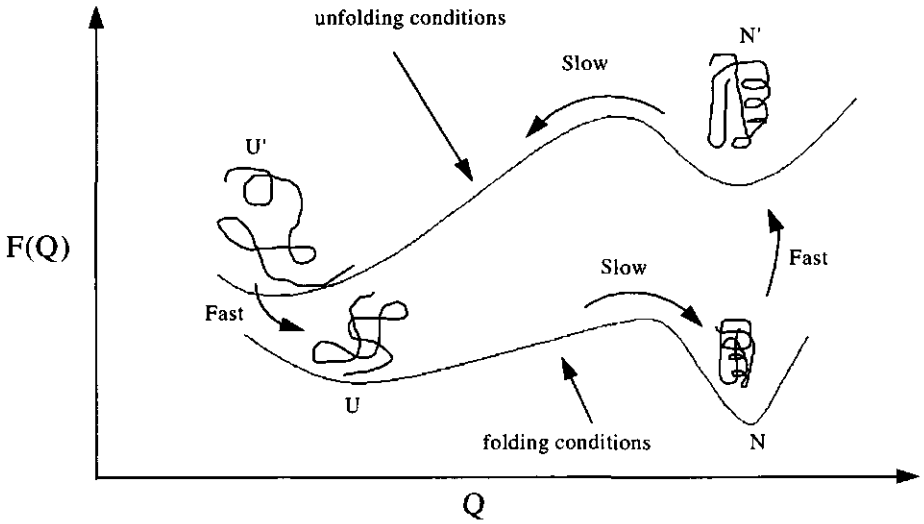


Figure-6. Free energy profiles for protein folding and unfolding for two-state proteins. Q is the reaction coordinate. The process U'→U is the response of the unfolded state to the change from unfolding to folding conditions. N→N' is the response of the native state to the change from folding to unfolding conditions. The barrier between the folded and unfolded states is not sharp, i.e. there is an ensemble of transition structures of which the probability of proceeding forward to the native state is significantly higher than back to the unfolded state.

In addition to the equilibrium measurements, kinetic consideration of the free energy profiles in fig.6 indicate that fast initial processes can occur even in two-state proteins. In experiments, in which the solvent is changed from strongly unfolding to strongly folding conditions, collapse to more compact structures occur. In the free energy profile this is a barrierless process and therefore very fast. The formation of a more compact denatured state, as the denaturant concentration decreases, is a general property. An interesting issue, that is addressed by fast mixing experiments, is whether the compact denatured state under strongly

folding conditions retains a random organization of the polypeptide chain, or whether it assumes a native-like fold, as found in the molten-globule state.[53,54] Fast processes occur also in unfolding experiments prior to the formation of the unfolded state (fig.6). Changing the solvent conditions influence the distribution of conformations in the native state and make partly unfolded structures rapidly accessible. Conformational changes in native proteins is known to occur in approximately picoseconds. Therefore, fast structural changes may occur prior to crossing the main thermodynamic barrier to the unfolded state (a process that usually takes milliseconds or longer). In two-state kinetics both folding and unfolding are single exponential processes, and the ratio of the folding and unfolding rate constants is equal to the population ratio of folded and unfolded states obtained in equilibrium experiments.[54]

Protein molecules on the one hand are finite systems and at the same time they are still very large in atomic terms. These features of protein molecules give the possibility to consider protein folding from two points of view, namely as a complex chemical reaction[28] and as a phase transition.[55] There are a few major questions around which the theoretical studies are focused: **1)** What are the sequence requirements for proteins to fold rapidly and be stable in their native conformations? **2)** What are the thermodynamic mechanism(s) of protein stabilization and the kinetic mechanism(s) of folding? **3)** Are there special native structures (packing patterns) that are more likely to correspond to the native structures of foldable proteins? **4)** What is the best approximation for protein-folding energetics? At present, because of the time scales involved in protein folding (from several microseconds to minutes), simplified lattice protein models whose dynamics are defined by Monte Carlo methods appear to be the only candidates for computational studies. Simplified models[55,56] provide a coarse-grained description (1 ms and $\geq 10\,\mathring{A}$, respectively), i.e. they cannot depict all the details of protein structures such as location and size distribution of hydrogen-bonded secondary structures, side chain conformational degrees of freedom important for packing, etc. Each of these different details has its own energy scale and it is not obvious how they can be taken into account when making the connection between simplified models and experiments on real proteins. The main requirement for simplified models is their ability to reproduce the most essential effects of protein folding: a unique native structure, a large number of conformations (the 'Levintal paradox'), fast folding to native state and cooperative folding (first-order like) transition.

From the point of view of the complex chemical[29,56] reactions it is postulated that: firstly, in its non-native interactions, a protein resembles a random heteropolymer. In its extreme form this suggests that the energies of globally distinct states can be taken as random variables which are uncorrelated (random energy model). Secondly, when a part of the protein molecule is in its correct secondary structure, the energy contributions are expected to be stabilizing. In addition, when a correct contact is made, the energy may go up occasionally, but when averaged over all possible contacts, the energy will go down. Thirdly, the conformational states of protein chains with random sequences are characterized by a rough energy landscape. The source of the roughness are the frustration arising from conflicting interactions and geometric constraints, which present a large barrier for the reconfiguration between different conformational states. Thus protein folding is regarded as the motion of a protein chain on a partially rugged, funnel-shaped energy landscape as it searches through an enormous number of possible configurations on its way to the unique native structure. The landscape of the protein funnel is characterized by three parameters: the mean square interaction energy fluctuations, ΔE^2, measuring ruggedness; a gradient toward the folded state, δE_s; and an effective configurational entropy, S_L, describing the search problem size ('Levintal paradox'). The main feature of the funnel description is the concerted change in both the energy and the entropy as one moves along the reaction coordinate (similarity of the configuration to the native structure-degree of collapse, helicity, fraction of correct contacts and dihedral angles etc.). As the entropy decreases so does the energy. The gradient of free energy determines the average drift up or down the funnel. Superimposed on this drift is a stochastic motion whose statistics depends on the jumps between local minima. In a first approximation this process can be described as diffusion. There are different kinetic scenarios depending on the variation of entropy, mean energy and ruggedness as the protein ensemble descends in the funnel. If the energy loss always exceeds the entropy, when the ensemble of intermediate structures becomes progressively more native-like, there is a 'downhill' scenario. This occurs for folding funnels with very large δE_s. The rate of folding does not follow the first-order chemical reaction, rather it depends mainly on the lifetimes of the individual microstates. If a glass transition occurs before the native state is reached, the overall kinetics become multiexponential. Different protein molecules are stucked in a few different microstates with different rate of folding. When the entropy decreases more rapidly than the energy, a free energy maximum results. At the macroscopic level the folding is described by a single-step, exponential kinetics.

The 'ruggedness' of the landscape is responsible for the transient trapping of structures that are either partially folded or misfolded. In other words, there is a glass transition which can occur either before the thermodynamic barrier or after it. If the landscape is sufficiently smooth, the traps are shallow and there is no significant accumulation of intermediate structures that may result in slow and/or multiphasic kinetics. In this case, only two states (folded and unfolded) are observed in both kinetic and equilibrium experiments (fig.6). Activation to an ensemble of states near the top of the free energy barrier is the rate-determining step. Another very important factor, which controls the search for the native state, is the geometric constraints coming from the chain connectivity and determining in particular the native state accessibility. They force the sequence to assemble by following an average sequential order. The protein initially collapses and starts a reconfiguration process until some critical residues are properly assembled. Once this occurs, the probability of formation of other native contacts is enhanced and the folding results in an increasingly rapid collapse to the native state.

From the point of view of statistical thermodynamics[55] the important question is whether the principle of minimal frustration, postulated in the random energy model of energy landscape,[28] can be realized with realistic Hamiltonians by the choice of appropriate sequences. It is postulated that the size of the energy gap between the native state and the unfolded conformations determines the stability and the folding rates of proteins. According to this idea, large gaps are associated with fast folding and stable sequences. Combination of design and folding makes it possible to investigate protein folding and evolution separately from the problem of finding the correct potential functions for the Hamiltonian. As in the random energy model, the energies of the conformations, which are not similar to the target structure of a given designed sequence, are statistically equivalent to those of a random heteropolymer. This leads to the existence of a threshold energy E_C such that the probability to find conformations with an energy well below E_C is extremely small. In other words, if the probability function $W(E)$ is such, that $W(E)\Delta E$ gives the number of conformations with energies belonging to the interval $(E, E+\Delta E)$, than the energy spectrum of a heteropolymer has a quasi-continuous part at $E > E_C$ such that $W(E) >> 1$ and a lower discrete part at $E \leq E_C$ where $W(E) \sim 1$. Therefore the successful design should create sequences whose native energy E_N is well below E_C. In this way random conformations will not have energies close to that of the native conformation and therefore will not serve as deep energetic traps for folding. Another

important factor, which is necessary for the reproduction of the most universal feature of the thermodynamics of real proteins, is cooperativity. Cooperativity suggests that proteins, at conditions in which their native state is thermodynamically stable, follow a first-order-like phase transition. This is supported by simulations on simplified model lattices. In these simulations more cooperative transitions between unfolded and native states (with little structural similarity between them) were found to correlate with sequences designed to have large energy gaps and faster folding at temperatures where their native state is thermodynamically stable. In contrast, sequences having a noncooperative-folding transition fold very slowly at the condition in which their native state is stable (a temperature lower than that of the glass transition is required to stabilize native conformations in this case). As was already discussed above, the double-well free energy profile (fig.6) suggests two relaxation times: one for the motion of a protein chain in the free energy minimum corresponding to the unfolded state (which is very fast); and second to overcome the free energy barrier between unfolded and folded states. In the fast folding part there is a reconfiguration via a collapsed 'burst-phase' intermediate, but this strongly depends on the conditions. Thus at higher temperature, at which the entropy contribution is more pronounced, partly structured intermediates are favorable. However at lower temperatures, in which the enthalpic contribution to the free energy barrier becomes dominant, low energy intermediates are disfavored. And finally, the cooperativity suggests that kinetically the protein folding transition follows a nucleation mechanism. The nucleus is the lowest of all the free energy barriers in the transition state ensemble of conformations. It is represented by certain nonlocal native contacts which lead to the formation of a critical fragment after which the subsequent dynamics lead unidirectionally to the native state.

Recent experiments and theoretical (computational) developments have drastically improved our understanding of what happens during the folding of protein molecules and how this depends on such characteristics as environment conditions (concentration of denaturants, pH, temperature etc.) and amino acid composition. In recent years, it has also been realized that the search for a unique structure in a globular proteins involves the discrimination between different overall folded structures, and that the collapse into a globule having secondary structure does not by itself solve the problem of the search for a unique three-dimensional structure.[56] Thus the question still remains how this understanding of the protein folding problem will help to solve the problem of protein tertiary structure predictions. There is no

improvement in predicting the correct energetics of the residue-residue interactions. This difficulty can be overcome by sequence design, but is limited in its applications to simplified lattices. Improvement is needed in the direction of more realistic representation of protein chain conformations including secondary structure formation and side chain packing which are crucial for the formation and stabilization of the native state. It is the aim of the present thesis to make a first step in this direction.

REFERENCES

1 Pauling, L., Corey,R. B., and Branson, H. R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **37**: 205-211, 1951.

2. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein. Chem.* **34**, 167-339 (1981).

3 Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**: 95-99, 1963.

4 Pauling, L., and Corey,R. B. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc. Natl. Acad. Sci. U.S.A.* **37:** 729-740, 1951.

5 Leszczynsky, J. F., and Rose, G. D. Loops in globular proteins: a novel category of secondary structure. *Science*, **234**: 849-855, 1986.

6 Richards, F. M., Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**: 151-176, 1977.

7 Richards, F. M. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**: 1-14, 1974.

8 Richardson, J. S., and Richardson, D. C. In "Prediction of protein structure and the principles of protein conformation." Ed. Fasman, G. D., New York: Plenum, 1989.

9 Kauzman, W. Denaturation of protein and enzymes. In : McElroy W. D., Glass, B., eds. *The mechanism of enzyme action.* Baltimor: Johns Hopkins Press. Pp. 70-120, 1954.

10 Kauzman, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, **14**: 1-63, 1959.

11 Chotia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**: 537-572, 1984.

12 Sternberg, M. J. E., and Thornton, J. M. On the conformation of proteins: the handedness of the β-strand-α-helix-β-strand unit. *J. Mol. Biol.* **105:** 367-382, 1976.

13 Cohen, F. E., Richmond, T. J., and Richards, F. M. Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132:** 275-288, 1979.

14 Richmond, T. J., and Richards, F. M. Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* **119**: 537-555, 1978.

15 Janin, J., and Chothia, C. Packing of α-helices onto β-pleated sheets and the anatomy of α/β proteins. *J. Mol. Biol.* **143**: 95-128, 1980.

16 Cohen, F. E., Sternberg, M. J. E., and Taylor, W. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**: 821-862, 1982.

17 Barlow, D. J., and Thornton, J. M. Helix geometry in proteins. *J. Mol. Biol.* **201**: 601-619, 1988.

18 Chothia, C. Levitt, and Richardson, D. Helix to helix packing in proteins. *J. Mol. Biol.* **145**: 215-250, 1981.

19 Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R. Analysis of the tertiary structure of protein β-sheet sandwiches. *J. Mol. Biol.* **148**: 253-272, 1981.

20 Levitt, M., and Chothia, C. Structural patterns in globular proteins. *Nature*, **261**: 552-558, 1976.

21 Chothia, C. One thousand protein families for the molecular biologist. *Nature* **357**, 543-544 (1992).

22 Chothia, C., Hubbard, T., Brenner, S., Barns, H., and Murzin, A. Protein folds in the all-β and all-α classes. *Annu. Rev. Biophys. Biomol. Struct.* **26:** 597-627, 1997.

23 Murzin, A. G., Brenner, S. E., Hubbard, T., and Chithia, C. SCOP: a structural clasification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536-540, 1995.

24 Lesk, A. M., and Chotia. C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**: 225-270, 1980.

25 Chotia. C., Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826 (1986).

26 Schulz, G. E., and Schirmer, R. H. Principles of protein structure. Springer-Verlag, New York, Heidelberg, Berlin. 1979.

27 Dill, K. A., Bromberg, S., Yue, K., Feibig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. Principles of protein folding -a perspective from simple exact models. *Protein Science* **4**: 561-602, 1995.

28 Wolynes, P. G., Luthey-Schulten, Z., and Onuchic, J. N. Fast-folding experiments and the topography of protein folding energy landscapes. *Chemistry & Biology*, **3**: 425-432, 1996.

29 Eaton, W. A., Thompson, P. A., Chan, C. K., Hagen, S. J., and Hofrichter, J. Fast events in protein folding. *Structure*, **4**: 1133-1139, 1996.

30 Plaxco, K. W., and Dobson, C. M. Time-resolved biophysical methods in the study of protein folding. *Curr. Opin. Struct. Biol.*, **6**: 630-636, 1996.

31 Pascher, T., Chesick, J. P., Winkler, J. R., Gray, H. B. Protein folding triggered by electron transfer. *Science*, **271**: 1558-1560, 1996.

32 Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R., and Gray, H. B. Cytochrome c folding triggered by electron transfer. *Chem. Biol.*, **3**: 491-497, 1996.

33 Zwanzig, R. Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 148-150, 1997.

34 Chan, C. K., Hu, Y., Takahashi, S., Rousseau, D. L., Eaton, W. A., and Hofrichter, J. Submillisecind protein folding kinetics studied by ultrarapid mixing. *Proc. Natl. Acad. Sci. U.S.A.* 1997, in press.

35 Takahashi, S., Yen, S. R., Das, T. K., Chan, C. K., Gottfried, D. S., and Rousseau, D. L. Folding of cytochrome c initiated by submillisecond mixing. *Nat. Struct. Biol.*, **4**: 44-50, 1997.

36 Chan, C. K., Hofrichter, J., and Eaton, W. A. Optical triggers in protein folding. *Science*, **274**: 628-629, 1996.

37 Fersht, A. R. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.*, **5**: 79-84, 1995.

38 Privalov, P. L. Stability of proteins. Small globular proteins. *Adv. Protein. Chem.* **33**: 167-241, 1979.

39 Privalov, P. L. Stability of proteins. Proteins which do not present a single cooperative system. *Adv. Protein. Chem.* **35**: 1-104, 1982.

40 Privalov, P. L., and Gill, S. J. Stability of protein structure and hydrophobic interaction. *Adv. Protein. Chem.* **39**: 191-234, 1988.

41 Privalov, P. L., and Makhatadze, G. I. Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. *J. Mol. Biol.* **232**: 660-679, 1993.

42 Privalov, P. L., Tiktopulo, E. I., Venyaminov, S. Yu., Grico, Yu., V., Makhatadze, G. I., and Khechinashvili, N. N. Heat capacity and conformation of proteins in the denatured state. *J. Mol. Biol.*, **205**: 737-750, 1989.

43 Wang, Y, and Shortle, D. The equilibrium folding pathway of staphylococcal nuclease: identification of the most stable chain-chain interactions by NMR and CD spectroscopy; *Biochemistry*, **34**: 1585-1505, 1995.

44 Wishart, D. S., Sykes, B. D., and Richards, F. M. Relationships between nuclear magnet: ic resonance chemical shift and protein secondary structure. *J. Mol. Biol.*, **222**: 311-333, 1991.

45 Williams, K., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H., and Dyer, R. B. Fast events in protein folding: helix melting and formation in small peptide. *Biochemistry*, **35**: 691-697, 1996.

46 Gilmanshin, R., Williams, S., Callender, R. H., Woodruff, W. H., and Dyer, R. B. Fast events in protein folding: relaxation dynamics of secondary and tertiary structure in native apomyoglobin. *Proc. Natl. Acad. Sci. U. S. A.*, 1997 in press.

47 Thompson, P. A. Laser temperature jump for the study of early events in protein folding. In Techniques in protein chemistry. Edited by Marshak D. R. San Diego: Academic Press; 1997 : in press.

48 Jones, C. M., Henry, E. R., Hu, Y., Chan, C. K., Luck, S. D., Bhuyan, A., Roder, H., Hofrichter, J., Eaton, W. A.: Fast events in protein folding initiated by nanosecond laser photolysis. *Proc. Natl. Acad. Sci. U. S. A.*, **90**: 11860-11864, 1993.

49 Chan, C. K., Hofrichter, J., and Eaton, W. A. Optical triggers in protein folding. *Science*, **274**: 628-629, 1996.

50 Hagen, S. J., Hofrichter, J., Szabo, A., and Eaton, W. A. Diffusion-limited contact formation in unfolded cytochrom c: estimating the maximum rate of protein folding. *Proc. Natl. Acad. Sci. U. S. A.*, **93**: 11615-11617, 1996.

51 Hagen, S. J., Hofrichter, J., and Eaton, W. A. The rate of intrachain diffusion of unfolded cytochrome c. J. Phys. Chem. 1997 in press.

52 Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

53 Villegas, V., Azaga, A., Catasus, L. l., Reverter, D., Mateo, P. L., Aviles, F. X., & Serrano L. (1995). Evidence for two-state Transition in the Folding Process of the Activation Domain of Human Procarboxypeptidase A2. *Biochemistry* **34**, 15105-15110.

54 Tan, Y-J., Oliveberg, M., & Fersht, A. R. (1996). Titration Properties and Thermodynamics of the Transition State for Folding: Comparison of Two-state and Multi-state Folding Pathways. *J. Mol. Biol.* **264**, 377-389.

55 Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

56 Onuchic, J. N., Zaida Luthey-Schulten, & Wolynes, P. G. (1997). Theory of Protein Folding: The Energy Landscape Perspective. *Ann. Rev. Phys. Chem.* **48**, 545-600.

3 SELF-CONSISTENT FIELD APPROACH TO PROTEIN STRUCTURE AND STABILITY. I. PH DEPENDENCE OF ELECTROSTATIC CONTRIBUTION.

Roumen. A. Dimitrov and Robert. R. Crichton

Published in: *Proteins: Structure, Function and Genetics* 4:576-596, 1997

SUMMARY

Starting from the simple case of an external field acting on non interacting particles, a formulation of the self-consistent field theory for treating proteins and unfolded protein chains with multiple interacting titratable groups is given. Electrostatic interactions between the titratable groups are approximated by a Debye-Huckel expression. Amino acid residues are treated as polarizable bodies with a single dielectric constant. Dielectric properties of protein molecules are described in terms of local dielectric constants determined by the space distribution of residue volume density around each ionized residue. Calculations are based on average charges of titratable groups, distance of separation between them, on their pKa's, residue volumes and on the local dielectric constant. A set of different residue volumes is used to analyze the influence of the permanent dipole of polar parts of the residue on calculated titration curves, electrostatic contribution to the free energy of protein stability and pK shifts. Calculations with zero volumes- which means that charged portions of protein molecules are viewed as part of the high dielectric medium- give good agreement with experimental data. The theory was tested against most accurate approaches currently available for the calculation of the pKa's of ionizable groups based upon finite difference solutions of the Poisson-Boltzmann equation (FDPB). For 70 theoretically calculated pKa's in a total of six proteins the

accuracy of the approach presented here is assessed by comparison of computed pKa's with that measured. The overall root-mean-square error is 0.79 compared to the value 0.89 obtained by FDPB approach given in the paper of Antosiewicz et. all.[1] The test of Debye-Huckel approximation for the electrostatic pair interactions shows that it is in excellent agreement with experimental data as well as the calculations of the FDPB and Tanford-Kirkwood methods on the pK shifts of His64 in the active site of subtilisin over the whole range of ionic strengths.[2,3] The theory was also analytically and numerically tested on a simple models where the exact statistical mechanical treatment is still simple.[4,5]

INTRODUCTION

From early studies[6] where proteins are assumed to be impenetrable spheres with charges uniformly distributed over their surfaces it was widely believed that the destabilizing effects of pH on protein stability were mainly due to the repulsive interactions between titratable groups. These result from the large net charge that accumulates on the protein when it is far from its isoelectric point. However recent experimental advances have shown that the main difference between the folded and unfolded protein states comes from a small number of amino acids with anomalously shifted pKa's in its native state.[7,8] Since these important experimental results were recognized, the main goal of current theoretical approaches is to represent the electrostatic contribution to total protein stability in terms of individual free energy contributions of acidic and basic groups, as well as to understanding the physical basis and estimating the degree to which the protein environment of each group alters its pKa's relative to the pKa's of the groups which surround it.

Various methods are currently available for computing the electrostatic energy of polypeptide or protein molecules but not one method is perfect for all applications. It is important that in all these methods, the only specific characteristic of the protein molecule is that in its native state, the protein forms a compact spherical-like region, the surface of which separates two regions that differ in their composition. The molecular interior is generally considered to be of low dielectric constant.[9,10] Outside the molecular boundary there is an aqueous medium of high dielectric constant. Analytical solution of such a model for natural shapes of protein molecules is not yet available because of the complex behaviour of the dielectric constant and ionic strength through space. The only existing analytical solution which

deals with hypothetical spherical molecules has been proposed by Kirkwood.[11] In one of the most popular versions of this approach[12] the protein is assumed to be a sphere with charges located at discrete sites immersed at some fixed distance beneath the surface of the sphere. This model has been modified to incorporate depth parameters and surface accessibility[13,14,15] Attempts were made to deal with the detailed microscopic dielectric effects of the protein and the surrounding water molecules. The application of the method known as PDLD-"protein-dipoles Langevin-dipoles"[16], gives reasonable results. There are also models which treat proteins with arbitrary shapes. They can be separated into two groups with the use of Poisson-Boltzmann (PB) differential equations on the basis of a finite-difference approach[17,18,19,20] or with an integral formulation of the problem on the basis of an appropriate density distribution of induced polarization charge on the protein-solvent boundary.[21] Recently, the finite-difference solutions of PB and particularly linearized PB equations appear to be the natural approach for computing detailed electrostatic fields in and around macromolecules.[1,2,22,23,24,25] An important recent application[26] examines the contribution of salt bridges to protein stability. From these results, it follows that the native state is weakly destabilized relative to the denatured one. In the last few years an attempt was made to incorporate the PB model in conformational-search algorithms.[27,28]

Because of the complexity of averaging both conformational and charged states for protein chain and each ionised residue respectively, previous investigations have attempted to attack one or other aspect separately. A well known approximation which averages over the ionization state of charged residues is the iterative method of Tanford and Roxby.[29] In this method each ionizable group interacts with the average charge of the other groups. Recently it has been shown that this method is a mean field one and a reduced site method has been proposed.[5] Here, only the pKa's of residue groups which are near to the pH of interest are treated with the full statistical mechanical expressions; the other ionizable groups are treated by a mean field approach with some modifications. However at some pH values all of the acidic or basic groups must be included in the rigorous statistical mechanical summation. Some improvement was obtained by introducing the so called hybrid statistical mechanical/Tanford-Roxby approximation[4] and "cluster method".[30] A powerful Monte Carlo method has also been developed which is very apropriate for large systems.[31]

A theory was developed which deals with the folding process of a fictitious two-state thermodynamic pathway using a porous sphere model for unfolded protein chains.[32,33] The

potential for the unfolded state is calculated using the Poisson-Boltzmann relation and the interior of the porous sphere is taken to be a linear function of the chain density. The folded state is accounted for by the approximation of the Linderstrom-Lang model. The theory successfully accounted for pH and salt effects on myoglobin stability.[34]

In this present paper we present an approach which deals with both the problem of conformational state and of ionisation of individual residues. The electrostatic interactions between the protein titratable groups are considered on the basis of a self-consistent field approach. The method is based on the approximation of electrostatic interactions between the titratable groups with the interaction of each of them with a self-consistently determined molecular field. The field can be different for different titratable groups or for identical groups placed at different spatial coordinates.

The explicit treatment of heterogeneity of residue sequence makes it possible to determine all specific dielectric properties of ionized residue environments in the native as well as in the denatured state. The use of molecular field approximation gives the possibility to replace the complex dependence of electrostatic interactions on the space distribution of residue charges with a sum over separated energy terms, each of which depends only on the coordinates of the corresponding ionized residue. This allows, in a simplified form, to take into account the averaging of electrostatic interactions over the different conformational states of the protein chain. In this paper we consider the difference of electrostatic free energy between the native and the denatured state of the protein chain. The native state is characterized using spatial coordinates taken from the Brookhaven Protein Data Bank. Electrostatic interactions in the denatured state are only between residues which are close to one another along the chain but not further than three residues, and strongly dependent on protein chain sequence.

In order to clarify the accuracy of the theory presented here the self-consistent field approach is applied to determine 70 individual pKa's in a total of six proteins with currently available experimental data- bovine pancreatic trypsin inhibitor (BPTI), ribonuclease A (Rnase A), ribonuclease T1 (Rnase T1), two crystal forms of hen egg white lysozyme (HEWL), T4 lysozyme and barnase. Our results are also extensively compared with the most accurate methods existing now based on the finite difference Poisson-Boltzmann method[1,4,22,23] as well as analytically and numerically tested on a simple models where the exact statistical mechanical treatment is still simple.[4,5]

THEORY

Statement of the problem

The protein molecule is seen as a linear system with a given amino acid residue sequence along its chain. Amino acid residues are treated as polarizable bodies with a single dielectric constant. Dielectric properties of the protein molecule at any specific chain conformation are described in terms of local dielectric constants determined by the space distribution of residue volume density around each ionized residue. Electrostatic interactions are long-range ones so it is not convenient to chose any particular conformation of the protein molecule as a reference state of zero free energy. For the reference state we chose the state in which the residue groups along the chain are disconnected from each other and separated at a distance where they cannot interact and where there is no dissolved electrolyte. The ionization properties of separated residues are described by their intrinsic pKa's which define the free energy E_i^o of charging the ionized groups. The pKa values in folded or unfolded protein chains are affected by the pair charge-charge interactions-E_{ij}, the desolvation of ionized groups and their interactions with polar but neutral groups-E_i. For a particular chain conformation the partition function of the electrostatic part of free energy is a sum over the 2^N statistical weights $\exp\left(-\frac{\sum_i (E_i^o + E_i) + \frac{1}{2}\sum_{\substack{ij \\ i\neq j}} E_{ij}}{RT}\right)$ for a chain with N ionized residues. The pK shift of a particular ionized residue i is defined as follows: the partition function is separated in two terms $Z_i(0)$ and $Z_i(q_i^o)$ which represent the charge or uncharged state of residue i respectively and q_i^o is 1 when the residue i is basic or -1 if it is acidic. It is important to remember that $Z_i(0)$ represents the sum over the statistical weights of 2^{N-1} ionization states of all other ionizable groups along the protein chain. So we have:

$$\Delta pK_i = \frac{1}{2.303} \times \ln\left(\frac{Z_i(0)}{Z_i(q_i^o)}\right) - q_i^o \times (pH - pK_i) \qquad (1)$$

The titration properties are defined as:

$$Q = \sum_i \langle q_i \rangle = \sum_i q_i^o \times \frac{Z_i(q_i^o)}{Z_i(q_i^o) + Z_i(0)} \qquad (2)$$

The equations (1) and (2) give the formal solution to the problem of pH dependent properties in protein molecules but are of little value unless it is possible to evaluate the partition functions $Z_i(0)$ and $Z_i(q_i'')$. The aim of this paper is to evaluate these terms on the basis of a self-consistent field approach.

Self-consistent field

If a protein has N ionizable residues, a given ionization state of the protein can be defined in terms of the vector $\{q_i\}$,i=1 to N, where q_i is 0 when the group i is neutral and 1 for basic or -1 for acidic groups when it is charged. The Hamiltonian of electrostatic interactions at some given ionization state of the protein molecule can be represented in the form:

$$E(\{q_i\}) = \sum_i^N E_i^o + \sum_i^N E_i + \frac{1}{2}\cdot\sum_i^{N-1}\sum_{j\neq i}^{N-1} E_{ij}$$

$$= \sum_i^N \mu_i \cdot q_i + \sum_i^N \lambda_i(\vec{r}_i)\cdot(q_i)^2 + \frac{1}{2}\cdot\sum_i^{N-1}\sum_{j\neq i}^{N-1} q_i \cdot f_{ij}(\vec{r}_i,\vec{r}_j)\cdot q_j \qquad (3)$$

where μ_1 is the change in electrostatic energy of charging the group i in water, $\lambda_i(\vec{r}_i)$ is the change in electrostatic energy of the ionized group i in the all neutral protein environment relative to the aqueous phase, $f_{ij}(\vec{r}_i,\vec{r}_j)$ is the energy of electrostatic interaction between residues i and j, in their charged state.

Following equation (3) for the free energy F of electrostatic interactions between the titratable groups in the protein chain we can write:

$$F = -RT\cdot\ln\left(\sum_{\{q_i\}}\exp\left(-\frac{E(\{q_i\})}{RT}\right)\right)$$

In many cases when the interparticle interactions have multiple characters it is appropriate to assume a simplified form for them, in such a way that the main properties of the system are conserved. The required modifications can be obtained on the basis of some minimization rules. For our purpose as a starting point for the minimization procedure the most appropriate is the well known classical statistical mechanics Gibbs-Bogoliubov inequality.[37,55]

Let $\{E_n\}$ be the energies of the states for the investigated system and $\{E_n'\}$ be the energies of the same states for the modified one. All properties of the systems are described in terms of the probability distribution of their states. We have:

$$P_n = \exp\left(\frac{F - E_n}{RT}\right) \text{ and } P_n' = \exp\left(\frac{F' - E_n'}{RT}\right) \tag{4}$$

where

$$F = -RT \cdot \ln\left(\sum_n \exp\left(-\frac{E_n}{RT}\right)\right) \text{ and } F' = -RT \cdot \ln\left(\sum_n \exp\left(-\frac{E_n'}{RT}\right)\right)$$

are the free energies of the systems. Their probabilities have to fulfill the usual normalization conditions:

$$\sum_n P_n = \sum_n P_n' = 1$$

Using (4) the Gibbs-Bogoliubov inequality can be written in the form:

$$F \le F' + \langle E - E' \rangle_{P'} \tag{5}$$

The inequality (5) states that the free energy of the system of interest is less than or equal to the free energy of the model system plus the average value of the deviation of the energy levels of the system of interest relative to the coresponding energy levels of the model system as calculated in the model system. Free energy F of electrostatic interactions can be obtained by the minimization over the energy E' of the model system in the form:

$$F = \min_{E'}\left\{F' + \langle E - E' \rangle_{P'}\right\} \tag{6}$$

Inequality (5) gives only the upper limit to the free energy of the system of interest. Therefore the choice of the model system and it's adjustable parametrers are restricted by the requirement that the limit on the right of inequality (5) be as small as possible. In the most simple case the model system is chosen in such a way that real interactions are approximated by the molecular field acting at each charged residue. In this case the last term in equation (3) is represented in the form:

$$\frac{1}{2} \cdot \sum_i \sum_{j \ne i} q_i \cdot f_{ij}(\vec{r}_i, \vec{r}_j) \cdot q_j \approx \sum_i q_i \cdot \Phi(\vec{r}_i)$$

where $\Phi(\vec{r}_i)$ is the molecular field.

Therefore the Hamiltonian of the model system at some given ionization state $\{q_i\}$ of the protein molecule is:

$$E'(\{q_i\}) = \sum_i \left(\mu_i \cdot q_i + \lambda_i(\vec{r}_i)\cdot(q_i)^2 + q_i \cdot \Phi(\vec{r}_i)\right)$$

In a molecular field, ionized residues are independent from each other, so for the free energy F' of the model system we can write:

$$F' = -RT \cdot \ln Z,\ Z = \prod_i \left(1 + \exp\left(-\left(\mu_i \cdot q_i^o + \lambda_i(\vec{r}_i)\cdot\left(q_i^o\right)^2 + q_i^o \cdot \Phi(\vec{r}_i)\right)\Big/RT\right)\right),\ q_i^o = \begin{cases} -1 \\ 1 \end{cases}$$

The probability that residue i is in the charged or uncharged form when it is placed at space coordinate $\vec{r}_i$ is:

$$P'(\vec{r}_i) = \frac{\exp\left(-\left(\mu_i \cdot q_i^o + \lambda_i(\vec{r}_i)\cdot\left(q_i^o\right)^2 + q_i^o \cdot \Phi(\vec{r}_i)\right)\Big/RT\right)}{1 + \exp\left(-\left(\mu_i \cdot q_i^o + \lambda_i(\vec{r}_i)\cdot\left(q_i^o\right)^2 + q_i^o \cdot \Phi(\vec{r}_i)\right)\Big/RT\right)} \text{ and } 1 - P'(\vec{r}_i)$$

The probability of some given set $\{q_i^o\}$ is:

$$P'\left(\{q_i^o\}\right) = \frac{\prod_{\{q_i^o\}} \exp\left(-\left(\mu_i \cdot q_i^o + \lambda_i \cdot \left(q_i^o\right)^2 + q_i^o \cdot \Phi(\vec{r}_i)\right)\Big/RT\right)}{\prod_i \left(1 + \exp\left(-\left(\mu_i \cdot q_i^o + \lambda_i \cdot \left(q_i^o\right)^2 + q_i^o \cdot \Phi(\vec{r}_i)\right)\Big/RT\right)\right)}$$

Thus from equation (6) for the free energy of electrostatic interactions between the titratable groups in the protein molecule we have:

$$F = F' + \sum_{\{q_i^o\}} P'\left(\{q_i^o\}\right)\cdot\left(\frac{1}{2}\sum_i \sum_{j\neq i} q_i^o \cdot f_{ij}(\vec{r}_i, \vec{r}_j)\cdot q_j^o - q_i^o \cdot \Phi(\vec{r}_i)\right)$$

$$= F' + \frac{1}{2}\sum_i \sum_{j\neq i} q(\vec{r}_i)\cdot f_{ij}(\vec{r}_i, \vec{r}_j)\cdot q(\vec{r}_j) - \sum_i q(\vec{r}_i)\cdot\Phi(\vec{r}_i) \qquad (7)$$

where

$$q(\vec{r}_i) = 0\cdot\left(1 - P'(\vec{r}_i)\right) + q_i^o \cdot P'(\vec{r}_i)$$

For the self-consistent solution from $\frac{\delta F}{\delta\Phi(\vec{r}_i)} = 0$ and $\frac{\delta F}{\delta q(\vec{r}_i)} = 0$ we obtain:

$$\delta F'/\delta\Phi(\vec{r}_i) = q(\vec{r}_i) \text{ and } \Phi(\vec{r}_i) = \sum_{j\neq i} f_{ij}(\vec{r}_i, \vec{r}_j)\cdot q(\vec{r}_j) \qquad (8)$$

where $\Phi(\vec{r}_i)$ plays the role of a self-consistent field. Minimization of free energy in (7) is carried out by an iteration process. We can begin with some arbitrary chosen molecular field or some distribution of average charges. In each step of the iteration procedure using the equation (8) we define the molecular field which acts at each position in the globular structure. This

field depends on the distribution of mean charges obtained at the previous step. Thus at equilibrium the charge distribution obtained by the molecular field must coincide with the charge distribution by which the molecular field is defined. Depending on the sharpness of the minimum it is possible that the free energy will increase if the step is too long and the minimum is passed. This is overcome by introducing a parameter λ $(0<\lambda<1)$ in the form:

$$\Phi(\vec{r})=\lambda\cdot\left(\Phi(\vec{r})_l-\Phi(\vec{r})_{l-1}\right)+\Phi(\vec{r})_{l-1}$$

where l is the number of iteration steps. If free energy increases λ is decreased by half.

Cluster field approach

The main difficulty of any kind of mean field method is that it is not possible to account for states in which one or several groups are charged and others neutral. The mean field method enforces symmetry and neglects correlations and as a consequence all the ionized groups are seen as simultaneously charged. In the case of dominant repulsive interactions above some critical value the mean field method leads to errors which are analyzed below. Here we give some modification of the self-consistent field approach described above which can take into account the charged and uncharged state of different ionized groups.

For a particular ionized residue i the other ionized residues are separated in groups or clusters each of which include the residue i. We assume that the different cluster fields do not act on residue i at the same time. So they represent different states of the system. We will illustrate the modification scheme in the most simple case of two cluster groups: the first cluster- includes all charged groups; the second cluster- includes only ionized residue i all other ionized residues are kept neutral. The second cluster is directly connected to the so-called 'Null' model-the assumption that the protein environment- in our case the electrostatic charge-charge interactions- do not shift pKa's of ionized residues at all. One has to keep in mind that only the electrostatic interactions between the titratable groups are treated self-consistently. The resulting free energy of residue i over the cluster fields is given by $-RT\ln\left(1+\exp\left(-\frac{q_i''\times\Phi_i}{RT}\right)\right)$, as a consequence the probability that residue i is in the charged form is given by the formula:

$$P_i = \frac{\exp\left(-\left(E_i^o + E_i - RT\ln\left(1+\exp\left(-\frac{q_i^o \times \Phi_i}{RT}\right)\right)\right)/RT\right)}{1+\exp\left(-\left(E_i^o + E_i - RT\ln\left(1+\exp\left(-\frac{q_i^o \times \Phi_i}{RT}\right)\right)\right)/RT\right)}$$

$$= \frac{\exp\left(-\frac{E_i^o + E_i}{RT}\right) + \exp\left(-\frac{E_i^o + E_i + q_i^o \times \Phi_i}{RT}\right)}{1+\exp\left(-\frac{E_i^o + E_i}{RT}\right) + \exp\left(-\frac{E_i^o + E_i + q_i^o \times \Phi_i}{RT}\right)}$$

The partition function of the residue i in the presence of an external field is defined as a sum over the clusters with nonzero external fields, therefore for the free energy we have:

$$F/RT = -\ln\left(1+\exp\left(-\frac{E_i^o + E_i + q_i^o \times \Phi_i}{RT}\right)\right)$$

The external field Φ_i is again determined by the equation (8) but now the probability P_i is given by the new formula defined above.

Electrostatic free energies

As was already mentioned above the protein molecule is seen as a linear system which consists of different amino acid residues, some of which may ionize. Residue groups are treated as polarizable bodies with a single dielectric constant. The charged portions of residue groups are taken as spheres with charges placed at their centers. The protein chain is immersed in a solvent medium which is treated as a dielectric continuum. The solvent can contain a dissolved electrolyte. Any particular chain conformation for a protein molecule with N ionized groups is characterized by the set of 2^N electrostatical energy levels-which are simply represented by all possible combinations of charged and uncharged forms for the ionized groups. Taking into account the reference state the charged ionized residues are characterized with an internal free energy given by:

$$\mu_i \cdot q_i^o = 2{,}3 \cdot RT \cdot (pK_i - pH) \cdot q_i^o \,, \qquad q_i^o = \begin{cases} -1, & \textit{acidic group} \\ 1, & \textit{basic group} \end{cases} \tag{9}$$

An assumption is made that electrostatic potentials caused by the charged ionized groups are governed by the linearized Poisson-Boltzmann differential equation[52]:

$$\nabla\varepsilon(r)\nabla\Phi(r) - \varepsilon(r)k^2(r)e\Phi(r) + 4\pi\rho(r) = 0$$

$$\varepsilon k^2 = \frac{8\pi e^2 N_a I}{k_B T}$$

where N_a, e, k_B, T, Φ, I are Avogadro's number, the charge on the proton, Boltzmann's constant, absolute temperature, electrostatic potential, and ionic strength, respectively. As a consequence of the above assumption the electrostatic energy levels for a particular protein chain conformation are given by the formula:

$$E\left(\{r_i\},\{q_i\}\right) = \frac{1}{2}\sum_i q_i \Phi\left(\{r_i\}\right),$$

where $\{r_i\}$ and $\{q_i\}$ define the spatial coordinates of amino acid residues and the distribution of charged and uncharged ionized residues along the protein chain respectively.

The usual numerical or analytical (for some simple assumption about the shape of the protein molecule) way to solve the PB equation is based on the assumption that the protein charges are separated from the aqueous phase by the protein surface and as a consequence are part of the interior of the protein molecule. Therefore for the interior of the protein molecule the solution of the Poisson-Boltzmann equation for the electrostatical potential produced at coordinate r_j by a unit charge placed at coordinate r_i is given in the form $\Phi\left(r_i, r_j\right) = \frac{1}{\varepsilon_p \left|r_i - r_j\right|} + \hat{\Phi}\left(r_i, r_j\right)$, where the first term is the usual Coulomb term with the dielectric constant equal to that in the interior of the protein molecule. The second term is connected with the dielectric interface between the protein interior and the aqueous phase. The electrostatic potential $\Phi\left(r_i, r_j\right)$ outside the protein molecule can be separated into two basic terms which correspond to the interaction of the ionized groups of the protein with the polarization charges which they cause in the aqueous phase, and with the atmosphere of the mobile ions around the protein molecule. A detailed description of the variety of terms in which the electrostatic potential can be separated, their physical interpretation and the accuracy of the overall scheme of calculation compared to the experimental data and simple exact soluble models is well documented.[22,23,25,38,39]

Our approach differs from the above only in its assumption that the ionized groups are seen as part of the aqueous phase rather than the interior of the protein molecule. This point of view is supported by recent experimental and theoretical molecular dynamic calculations of dielectric constant of protein molecules.[40,41] The mathematical background is the same linearized Poisson-Boltzmann differential equation, but the solutions now are different. We are interested in physically simple model solutions of PB equation rather than exact considerations. This follows from the basic assumption already made in the derivation of self-consistent field. The external field in equation (8) which acts on ionized groups is averaged over their charge state. Moreover in the average procedure we do not take into account the correlation between the ionized residues. Therefore we can use such solutions of the PB equation which are connected with the most probable charge states for the individual ionized groups. For charged groups there are three possibilities: to be highly extended toward the aqueous phase; to be part of the interface between the protein interior and the solvent and to be immersed in the low dielectric environment of protein interior. It was already shown in the literature that the last case for single and even for closely interacting charged groups is very unfavorable.[26,30,42] In most of the cases they will be neutralized, or if the charged forms are stabilized by some additional interactions, such as for example hydrogen bonds, they will not contribute to the pH dependent properties of the other charged groups. This follows from the fact that the electrostatic potential, which two opposite interacting charges produce, must decrease when charges approach each other.

In the case when ionized residues are seen as part of the high dielectric medium the following approximations are made: firstly, charged ionized groups do not interact with the polarization caused by the other charged groups, secondly, the ionic atmosphere around each charged ionized group is not affected by the other charged groups and lastly, we neglect the position-dependence for the contribution of the atmosphere of the mobile ions. As a consequence the solution of the PB equation outside the spheres of the charges can be approximated with the well known Debye-Huckel expression and the third term in the Hamiltonian of electrostatic interactions takes the form:

$$q_i^a f(r_i, r_j) q_j^a = q_i^a \frac{e^2 \exp\left(-\frac{|r_j - r_i| - a}{r_d}\right)}{\varepsilon |r_j - r_i| \left(1 + \frac{a}{r_d}\right)} q_j^a \qquad (10)$$

where a is the charge exclusion radius and r_d is the Debye radius. Inside the charged spheres the solution has the form:

$$\Phi(r_i, r) = \frac{q_i^o e}{\varepsilon|r - r_i|} - \frac{q_i^o e}{\varepsilon} \times \frac{1}{r_d + a}$$

From the first term, taking the potential on the surface of the charges one can easily derive the expression for the self-energy of the ionized residues caused by the polarization of the solvent medium $\frac{1}{2}\frac{(q_i^o e)^2}{\varepsilon a}$ and from the second term, the self-energy caused by the mobile ion atmosphere $-\frac{1}{2}\frac{(q_i^o e)^2}{\varepsilon} \times \frac{1}{r_d + a}$. Taking into account the reference state the second term in the Hamiltonian of the electrostatic energy levels is given by:

$$\lambda_i (q_i^o)^2 = -\frac{1}{2}\frac{(q_i^o e)^2}{\varepsilon} \times \frac{1}{r_d + a} \tag{11}$$

Different considerations are needed when ionized residues are seen as part of the dielectric boundary between the protein interior and the aqueous phase. The boundary is treated as a region with non zero thickness described in terms of local density which characterizes the packing of protein chain portions around ionized groups and as a consequence the dielectric properties of the ionized residue environment. The modification, which can be derived from simple physical considerations, concern the charge exclusion radius which is taken to be an effective radius rather than the actual dimension of the charge portion of the ionized residue, and the Debye radius of the mobile ion atmosphere.. This can be illustrated in the following simple example. Let us look at the case when a charge is immersed in the centre of a dielectric sphere with radius R and dielectric constant ε_2 surrounded by a solvent medium with dielectric constant ε_1. The difference of the charge self-energy caused by the polarization of the dielectric sphere and the surrounding solvent medium is given by the expression:

$$\Delta F_i = \frac{(q_i^o e)^2}{2}\left(\frac{1}{\varepsilon_2} - \frac{1}{\varepsilon_1}\right) \cdot \left(\frac{1}{a} - \frac{1}{R}\right) = \frac{(q_i^o e)^2}{2a_{eff}}\left(\frac{1}{\varepsilon_2} - \frac{1}{\varepsilon_1}\right)$$

If R >> a, $a_{eff} \approx a$. It follows that the dielectric properties of local regions around each residue in the folded protein chain attains that of the infinite bulk medium with the same local dielectric constant not far from the residue charge. Therefore we can neglect the boundary effects associated with R, or account for it by the effective radius of the residue charge.

The above simple analysis allows us to assume the following model, from which a Debye screening length can be obtained. Let n_o be the concentration of mobile ion in the solvent medium far from the region occupied by the chain. An assumption is made that the interface between the interior of the protein molecule and the aqueous phase is penetrable for the mobile ions. The concentration of mobile ions around some ionized residue i is:

$$n = n_o \cdot \exp\left(-\frac{e^2}{2a_{eff}RT}\cdot\left(\frac{1}{\varepsilon_i}-\frac{1}{\varepsilon_w}\right)\right)$$

where ε_w is the dielectric constant of the solvent. Thus for the Debye length of screening we have:

$$r_d \approx \left(\frac{\varepsilon_i}{n}\right)^{\frac{1}{2}}$$

As a consequence of the above considerations, for the second term in the Hamiltonian of the electrostatic energy levels we can again use the Debye-Huckel solution, but now the basic parameters as charge radius and Debye length are function of the residue spatial coordinates. For each residue there is a sphere with specific radius from which one can derive the corresponding effective parameters for the Debye-Huckel solution. For charge radius the coordinate dependence is very small such that we will keep this parameter constant. Hence the second term in the Hamiltonian takes the form:

$$\lambda_i \cdot \left(q_i^o\right)^2 = \frac{e^2}{2}\cdot\left(\frac{1}{a_{eff}}\cdot\left(\frac{1}{\varepsilon_i}-\frac{1}{\varepsilon_i^o}\right)-\frac{1}{\varepsilon_i}\cdot\frac{1}{r_d+a_{eff}}\right)\cdot\left(q_i^o\right)^2 \tag{12}$$

where a_{eff} ($\frac{1}{a_{eff}}=\frac{1}{a}-\frac{1}{R}$) is the effective charge radius; ε_i is the local dielectric constant of residue i in the folded state of the protein molecule, ε_i^o is the local dielectric constant of residue i in aqueous solution in the absence of all other residues, e is the electronic charge.

The third term in the Hamiltonian now is approximated by the expression:

$$\left(q_i^o\right)\cdot f_{ij}\left(\vec{r}_i,\vec{r}_j\right)\cdot\left(q_j^o\right)=\left(q_i^o\right)\cdot\frac{e^2}{\left|\vec{r}_i-\vec{r}_j\right|\cdot\varepsilon_{ij}\left(1+\frac{a_{eff}}{r_d}\right)}\cdot\exp\left(-\frac{\left|\vec{r}_i-\vec{r}_j\right|-a_{eff}}{r_d}\right)\cdot\left(q_j^o\right) \tag{13}$$

where $\varepsilon_{ij} = \frac{1}{2} \cdot (\varepsilon_i + \varepsilon_j)$ and $r_d \approx \left(\frac{\varepsilon_{ij}}{n(\varepsilon_{ij})} \right)^{\frac{1}{2}}$. In Debye-Huckel theory the term $\frac{a_{eff}}{r_d}$ begins to be important at high ionic strength where the Debye length reaches values of a few angstroms. But taking into account that part of the space in the vicinity of an ionized residue is restricted for the mobile ions from the protein body, we will account for this by neglecting the term $\frac{a_{eff}}{r_d}$ in the denominator. Further we will neglect a_{eff} unless otherwise specified.

The advantage of the expression (13) is that now the energy between two residues depends on the local packing around and between the interacting charges from residue groups, which are far along the chain but close in space to the interacting pair, rather than on the shape of the molecule.

The interactions between the ionized groups when they are seen as part of the interior of the protein molecule are reduced to zero because of the zero probability of individual ionized groups to be in charged form. This strongly reduces the possible errors connected with the above approximations which in the other case, when the charges are fixed and ionized residue do not change, their ionization state will be very great when we move from the aqueous phase toward the interior of the protein molecule.

Local dielectric constant and local packing density

The main idea of the model for calculation of dielectric constant $\varepsilon(\vec{r})$ is that a dielectric constant ε, equal to that in the protein core, is ascribed to each amino acid residue. Chain conformations form different kinds of residue distribution in space .So we have the following picture: small particles with intrinsic dielectric constant ε_1 are distributed randomly in a bulk medium with dielectric constant ε_2. The electrical field is averaged in a volume bigger than the irregularity of residue distribution. For such a mean electrical field the bulk medium together with randomly distributed particles is treated as isotropic and a dielectric constant may be involved.[43] An exact result can be obtained in two cases:

1) $n(\vec{r}) << 1$, ε_1 and ε_2 are arbitrary

$$\varepsilon(\vec{r}) = \varepsilon_1 + n(\vec{r}) \cdot \frac{3(\varepsilon_1 - \varepsilon_2)}{\varepsilon_2 + 2\varepsilon_1} \tag{14}$$

where $\varepsilon(\vec{r})$ is expanded about $n(\vec{r})$ and corrected to first order.

2) $|\varepsilon_1 - \varepsilon_2| << \varepsilon_1 + \varepsilon_2$

$$\varepsilon(\vec{r}) = \bar{\varepsilon} - \frac{3\overline{\delta(\varepsilon)^2}}{3\bar{\varepsilon}} \ , \ \bar{\varepsilon} = \varepsilon_1 + n(\vec{r})\cdot(\varepsilon_2 - \varepsilon_1) \tag{15}$$

Equation (15) may be transformed in the form:

$$\varepsilon(\vec{r}) = \bar{\varepsilon} - n(\vec{r})\cdot\left(1 - n(\vec{r})\right)\cdot\frac{(\varepsilon_1 - \varepsilon_2)^2}{3\bar{\varepsilon}} \tag{16}$$

where

$$\overline{\delta(\varepsilon)^2} = \left(1 - n(\vec{r})\right)\cdot(\delta\varepsilon_1)^2 + n(\vec{r})\cdot(\delta\varepsilon_2)^2$$

$$\delta\varepsilon_1 = \varepsilon_1 - \bar{\varepsilon} = n(\vec{r})\cdot(\varepsilon_2 - \varepsilon_1)$$

$$\delta\varepsilon_2 = \varepsilon_2 - \bar{\varepsilon} = \left(1 - n(\vec{r})\right)\cdot(\varepsilon_2 - \varepsilon_1)$$

On the basis of equations (14) and (16) we take the following approximated form:

$$\varepsilon(\vec{r}) = \bar{\varepsilon} - n(\vec{r})\cdot\left(1 - n(\vec{r})\right)\cdot\frac{(\varepsilon_1 - \varepsilon_2)^2}{\bar{\varepsilon} + \varepsilon_1 + \varepsilon_2} \tag{17}$$

where $\varepsilon(\vec{r})$ runs between ε_1 and ε_2 when $n(\vec{r})$ runs between 0 and 1 (see fig.1).

For the determination of local density each residue is placed in the center of a sphere with suitable radius and the average volume density is taken in the form:

$$n_i(\vec{r}_i) = \sum_j \frac{v_j}{V}\cdot\theta\left(R - |\vec{r}_i - \vec{r}_j|\right)\cdot f\left(|\vec{r}_i - \vec{r}_j|, R\right) \tag{18}$$

$$f\left(|\vec{r}_i - \vec{r}_j|, R\right) = 1 - \frac{|\vec{r}_i - \vec{r}_j|}{R}$$

$$V = \int 4\pi\Delta\vec{r}^2\cdot f(\Delta\vec{r}, R)d\Delta\vec{r} \ , \ 0 \le |\vec{r}_i - \vec{r}_j| \le R$$

$$\theta\left(R - |\vec{r}_i - \vec{r}_j|\right) = 0 \text{ if } |\vec{r}_i - \vec{r}_j| > R \text{ and } \theta\left(R - |\vec{r}_i - \vec{r}_j|\right) = 1 \text{ if } |\vec{r}_i - \vec{r}_j| \le R$$

where i is the number of the residue along the chain; v_i is the volume of residue i; $\theta\left(R - |\vec{r}_i - \vec{r}_j|\right)$ takes into account only this group of the chain which contributes to the density inside the sphere, $f\left(|\vec{r}_i - \vec{r}_j|, R\right)$ is the weight radial factor inside the sphere.

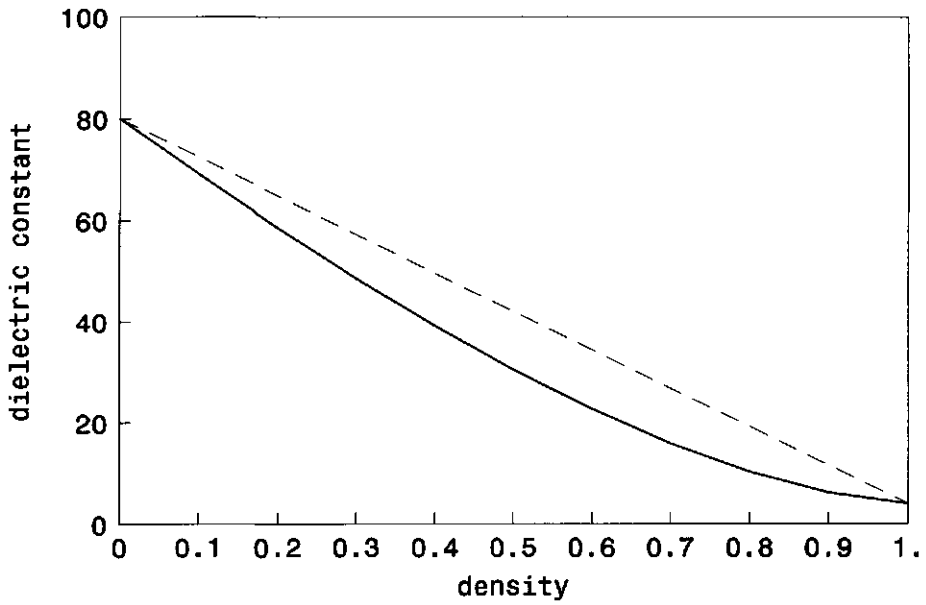


Figure-1 Dielectric constant as a function of local density. The theoretical curve is represented as (——) (see equation (17)). Linear relationship (– – –) between the $\varepsilon = 4$ and $\varepsilon = 80$ is given as a comparison and has the form: $\varepsilon = \varepsilon_w + n \cdot \left(\varepsilon_p - \varepsilon_w\right)$, where n is the local density.

RESULTS AND DISCUSSIONS

Numerical test on small model systems

Two interacting titratable groups.

We consider a detailed physical picture of a self-consistent field approach on a simple model of two interacting titratable groups.[4,5] The Tanford-Roxby approximation[29] works quite well for weakly (≤ 1 kcal/mol) interacting groups but breaks down when groups with similar pKa's interact strongly. The Monte Carlo sampling technique was also tested on the model of two interacting titratable groups.[4] The usual Metropolis algorithm is shown to converge far more slowly because of the energy barriers between states of low energy, causing the Monte Carlo trajectory to be trapped in a local minimum. When two ionized groups change their states at one Monte Carlo step the results are in excellent agreement with the exact solution. Our approach is also a mean field one, but it differs from the Tanford-Roxby method and gives fully compatible results with exact solutions up to coupling energies of 2.3 kcal/mol. For higher

coupling energies the results are similar to those of the Tanford-Roxby approach. However for coupling energies between 2 and 2.3 kcal/mol the minimization process becomes unstable. The comparison between the results of the present study and that of Tanford-Roxby[4] as well as the exact solution on the individual average charges at 2 and 2.76 kcal/mol is given in fig.2 A,B.

The main result from fig.2 is that above 2.3 kcal/mol the curves diverge from each other in the region of pH values where they are half ionized. From equations (7) and (8) one can find a simple analytical expression for the equilibrium free energy as a function of the average charges of the titratable groups relative to the reference state of zero energy interaction:

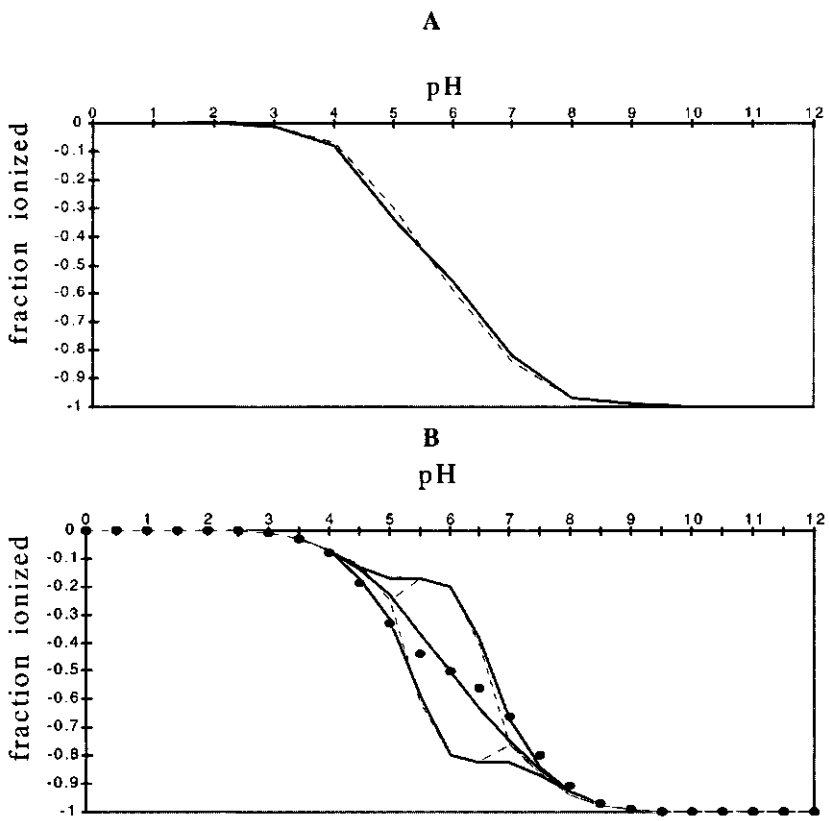


Figure-2. The individual titration curve of two identical ionized residue having $pK_{int} = 5.0$ and coupling energy 2 and 2.76 kcal/mol. Fig.2A-titration curve at coupling energy 2 kcal/mol; (----) calculation was by the self-consistent field approach; (——) calculation was done with the complete statistical method. Fig.2B-titration curve at coupling energy 2.76 kcal/mol; (----) calculation was by the Tanford-Roxby[4] approximation; (——) calculation was by the self-consistent field approach, the middle line represents the case where the minimization of free energy start with equal fractional charges for the ionized groups; (●) calculation was done with the complete statistical method.

$$\frac{F}{RT} = \ln\left(\frac{1.+\exp\left(-2.303q_1 \times (pH - pK_1)\right)}{1.+\exp\left(-2.303q_1 \times (pH - pK_1)\right) \times \exp\left(-\frac{q_1 \times E \times \langle q_2 \rangle}{RT}\right)}\right) +$$

$$\ln\left(\frac{1.+\exp\left(-2.303q_2 \times (pH - pK_2)\right)}{1.+\exp\left(-2.303q_2 \times (pH - pK_2)\right) \times \exp\left(-\frac{q_2 \times E \times \langle q_1 \rangle}{RT}\right)}\right) - \frac{\langle q_1 \rangle \times E \times \langle q_2 \rangle}{RT} \qquad (19)$$

where pK_1 and pK_2 are the intrinsic pKa's of the groups while E is the sign-independent interacting energy in kcal/mol. Because we are interested in the case where ionized groups are half ionized, there is an additional condition which has the form $\langle q_1 \rangle + \langle q_2 \rangle = 1$. The equilibrium free energies for coupling energies of 2 and 2.76 kcal/mol, as a function of the average charges, are shown in the fig.3 A,B.

At coupling energies between 2 and 2.3 kcal/mol the free energy minimum becomes very wide and the minimization process does not converge well. At higher coupling energies (Figure 3B) a second free energy minimum is attained. If the coupling energy is increased further, the separation between the minimums also increases: when we consider the fractional ionization of the groups, we reach a situation where one of them is fully charged while the other is uncharged. After the first 1 or 2 steps the minimization trajectory entirely lying on the free energy curves obtained by the equation (19) along the main diagonal of the square phase space as is shown on fig.3A,B. Depending on the starting conditions the minimization trajectory can be trapped in one of the local free energy minimums, or can be kept on the top of the free energy barrier. As a consequence of the entirely symmetrical potentials, which each group produce on the other the minimization trajectory, degeneration into a single point occurs.

At low pH values the fractional ionization of the groups is small and as a consequence the effective average coupling energy is smaller than 2.3 kcal/mol. Therefore at these pH values the free energy has one minimum. At high pH values the fractional ionizations of the ionized residues are close to 1, but in these conditions the coupling energy is only a small perturbation to the individual probabilities of each of the ionized residues to be in the charged state when there is no interaction between them. Hence at these pH values the free energy also has one minimum. Physical interpretation of self-consistent minimization of free energy can be given on the basis of the definition of the entropy of nonequilibrium systems.[44] The minimization

process can be seen as some kind of equilibrium kinetic which depends on the starting point in the phase space.

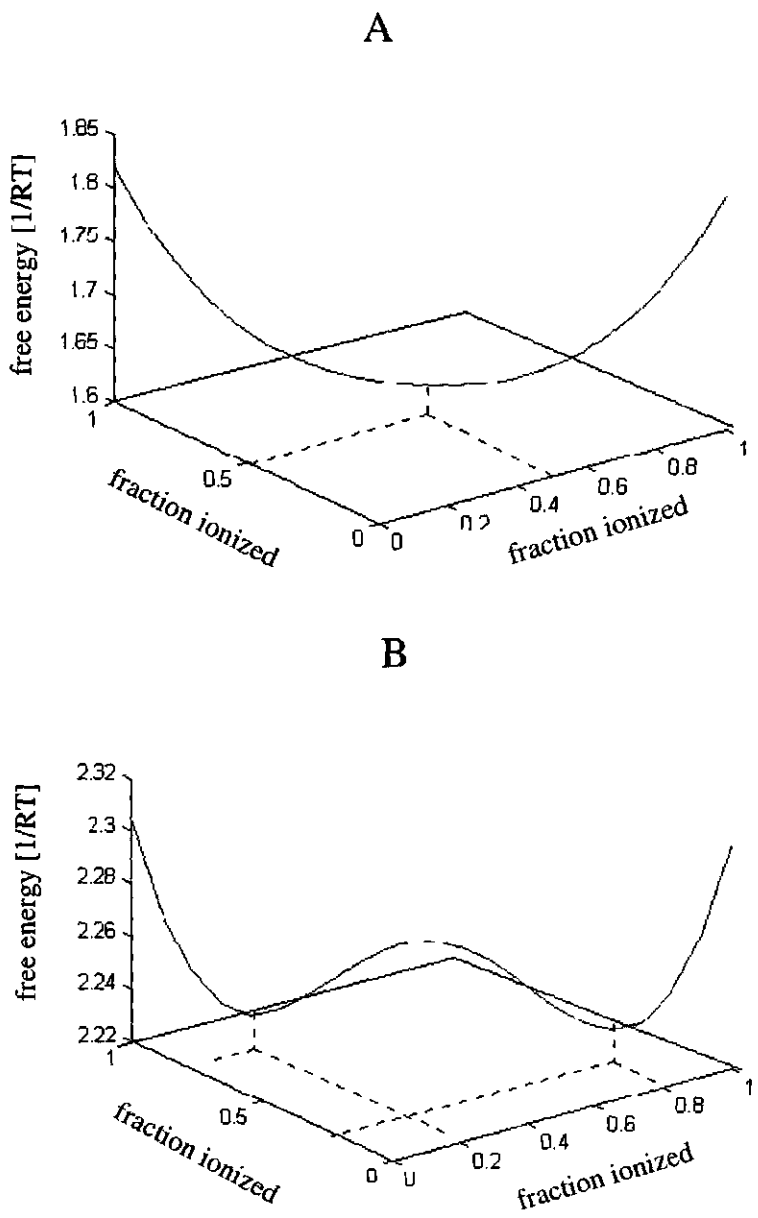


Figure-3. Equilibrium free energy as a function of the average charges of two identical ionized groups with $pK_{int} = 5.0$; fig.3A-coupling energy 2 kcal/mol and fig.3B-coupling energy 2.76 kcal/mol. The coupling energy 2 kcal/mol represents the critical value above which the free energy minimum is degenerate.

Each point on the minimization trajectory represents a nonequilibrium state, whose selection is realized under a self-consistently determined molecular field. This field restricts the system in some region in the phase space, the volume of which represents the entropy of the nonequilibrium state. Therefore, the entropy of the system in the external field can be represented in the form:

$$S^{noneq} = \left(F^{ext.field} - \langle \Phi \rangle\right) / RT$$

where the external field Φ is averaged using the Hamiltonian in which the external field is already included. Thus, the non equilibrium free energy of the system has the form:

$$F^{noneq.} / RT = \langle E \rangle_{\Phi} / RT - S_{\Phi}^{noneq.} \tag{20}$$

When one of the groups is acidic and the other basic there is no upper limit, self-consistent and exact solutions are identical everywhere. The exact solution according to the chosen reference state can be represented as follows:

$$\frac{F}{RT} = -\ln\left(1.+W_1 \times W_2 \times \left(\exp(-\frac{q_1 \times E \times q_2}{RT} - 1.\right)\right)$$

where $W_{1,2} = \dfrac{\exp\left(-2.303 \times q_{1,2} \times \left(pH - pK_{1,2}\right)\right)}{1.+\exp\left(-2.303 \times q_{1,2} \times \left(pH - pK_{1,2}\right)\right)}$ are the probability of the ionized groups in the reference state.

Ten Interacting Titratable Groups

The model system of ten ionized groups was first proposed and investigated in detail by Bashford and Karplus[5] who introduced the reduced-site approximation. Thereafter the model was widely used as a test for novel approaches. Therefore the model is useful for a comparison between different approaches. The basic idea of the reduced-site approximation method is to use the exact statistical mechanical calculation over a subset of titratable sites. The reduction of the number of titratable groups is based on the fact that not all will titrate at a pH of interest if their pKa's have a broad distribution over the range 4 to 10, which is typical for protein molecules.

The basic assumptions of the model are as follows: pKa's of the groups are randomly and uniformly distributed in the range 4.0 to 10.0, and the groups are randomly assigned as basic or acidic with equal probability. All self energies are set to zero. The accuracy test is assessed by

the root-mean-square and maximum deviation relative to the exact solution. The test procedure is separated in two parts: in the "strong coupling" case-the pair interactions are randomly and uniformly distributed in the range of 0.69-2.1 kcal/mol.; in the "moderate coupling" case- the pair interactions are randomly and uniformly distributed in the range of 0.138-1.1 kcal/mol.

Because of the additive way in which the parameters of the system, pKa's of the different groups, pH and the pair interaction terms E_{ij}, appear in the Hamiltonian of the system, there is some kind of translational symmetry relative to the pH. For example pK and pH always are in combination $(pH - pK)$ multiplied by some constant. Therefore if all the pKa's are changed by the same value Δ, the system can be kept in the same state by changing the pH in the opposite direction -Δ. Therefore all calculations are made at pH=7 over the total of 3000 sets of randomly generated pKa's and E_{ij} for the strong and moderate cases respectively.

The results for strong and moderate cases are shown in fig.4A,4B. The histogram bars represent the results taken over all 10 titratable groups in all 3000 'molecules' relative to the average charge θ of the individual groups.

From fig.4A and fig.4B we see that for the strong and moderate cases, the maximum errors of the self-consistent field approach are similar to that of the Tanford-Roxby mean field approach. But the average errors are similar to that of reduced-site approximation, especially for the moderate case, where at the extremes of θ intervals, the self-consistent field approach gives better results.

This can be easily explained. The upper limit for the site energy in the case of repulsive external field is 2 kcal/mol above which the self-consistent field fails. On the other hand, the results depend strictly on the distribution of titratable sites as either acidic or basic. For uniform probability distribution, the most probable is the 50%:50% acidic against basic groups. As such conditions depend on the character of the distribution of pair energy interactions E_{ij} the titratable sites are influenced by the attractive or close to zero repulsive fields, where the self-consistent field approach and the exact solution give essentially the same results. The maximal errors in fig.5A,5B show that at high repulsive interactions the self-consistent and Tanford-Roxby approaches give very similar results. The assumption that for the uniform distribution in most cases, the resulting field which acts on a particular site is close to zero, or is an attractive field, can be illustrated by the Cluster field approach described above. In this

approach the states with the resulting repulsive fields are ignored because the unshifted pKa states are more probable. The deviations from the unshifted pKa states are taken into account only in the resulting attractive fields. The results are summarized in Table I.

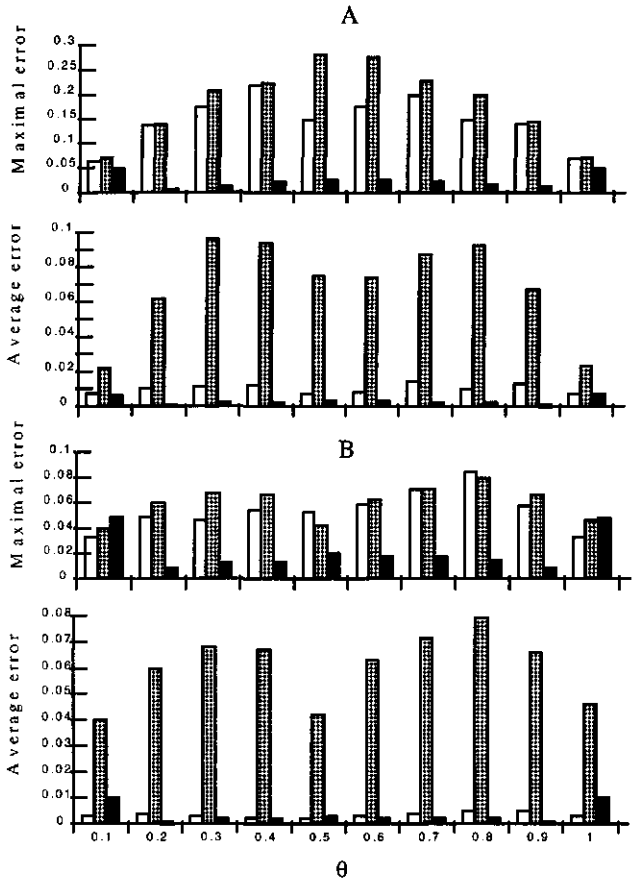


Figure-4. fig.4 Average and maximum errors of Tanford-Roxby[5] (gray bars), reduced-site approximation[5] (black bars) and self-consistent field (white bars) for ten interacting groups. Errors are defined as the difference between the exact θ values and that of the corresponding approximation. For self-consistent field the calculations are made for all of the groups of the 3000 random 'molecules' as explained in the text. Fig.4A,B represent the strong and the moderate case of coupling energy between the titratable sites, respectively.

The alternative way to view these results is directly connected to the central idea of the reduced-site approximation.[5] From equation (8) it follows that the field acting on a particular site is a sum of terms each of which is proportional to the probability that that particular site is in the charged state. Therefore at any given pH only sites which are more likely to be in the charged state will contribute.

Table I
Test on small model Systems

Method	Coupling	charg fraction		free nergy kcal/mol	
		Average error	Maximum error	Average error	Maximum error
Predominant state approx	Strong			0.240	1.200
	Moderate			0.097	0.520
Cluster method	Strong	0.022	0.170	0.077	0.310
	Moderate	0.006	0.062	0.024	0.140
Self-consistent field	Strong	0.044	0.220	0.290	2.400
	Moderate	0.015	0.096	0.106	0.850

This is also very close to the cluster and predominant methods.[30] Cluster method separates titratable groups into clusters and treats intra-cluster interactions exactly, but intercluster interactions approximately. The predominant method takes into account single highly occupied ionization states. The comparison with these methods allow us to include not only the fractional charges but also the free energy of the system. The results are summarized in Table II.

Table II
Test on small model Systems

	Self-consistent field				Cluster field			
	strong coupling		moderate coupling		strong coupling		moderate coupling	
Average	Average	Maximum	Average	Maximum	Average	Maximum	Average	Maximum
charge	error	error	error	error	error	error	error	error
0.0-0.1	0.007	0.064	0.003	0.033	0.060	0.980	0.034	0.620
0.1-0.2	0.010	0.137	0.004	0.049	0.046	0.790	0.038	0.600
0.2-0.3	0.011	0.176	0.003	0.046	0.030	0.760	0.029	0.590
0.3-0.4	0.012	0.220	0.002	0.054	0.047	0.650	0.026	0.500
0.4-0.5	0.007	0.149	0.002	0.053	0.027	0.480	0.024	0.420
0.5-0.6	0.008	0.176	0.003	0.060	0.025	0.047	0.027	0.430
0.6-0.7	0.014	0.198	0.004	0.096	0.030	0.390	0.023	0.340
0.7-0.8	0.010	0.148	0.005	0.084	0.018	0.250	0.020	0.270
0.8-0.9	0.013	0.139	0.005	0.058	0.017	0.180	0.015	0.180
0.9-1.0	0.007	0.069	0.003	0.032	0.009	0.095	0.008	0.080

The free energy obtained by the predominant and the self-consistent field approaches are very close to each other because both are based on the most probable state of the system. But

for the average charges the self-consistent approach gives very close results to that of the cluster method both for strong and moderate cases. For the free energy of the system the cluster method is considerably more accurate.

Test of Debye-Huckel Expression for the Electrostatic Pair Interactions on Aliphatic Dicarboxylic Acids: Determination of pK_a Shifts

The model calculations are based on the following assumptions: first-carboxylic groups of the dibasic acids are seen as low dielectric regions which at close distance give the contribution to the atom density around each of their effective charges; second-only the closest CH2 group gives contribution to the atom density around carboxylate charges when the number of intervening methylene groups is greater than 2. The radius of the sphere inside which the atom density around the charges is calculated by the equation (18) is $5\,\mathring{A}$. Because of the symmetry at this radius the carboxylic groups in oxalic acid are half screened by each other. The infinite long dibasic acid is chosen as the reference state.

The calculations are carried out using their X-ray coordinates of the dibasic acids. A dielectric constant 2 is assigned to each atom of the dibasic molecule and the atoms are represented by their van der Waals radiuses: $R_C = 1.7\,\mathring{A}$, $R_{CH2} = 1.9\,\mathring{A}$, $R_O = 1.5\,\mathring{A}$. The surrounding solvent is treated as a continuum with a dielectric constant 80.0. The charges of carboxylate groups are placed in the middle between the anionic oxygen atoms. Electrostatic interactions between the charges are approximated by the Debye-Huckel expression and because there are not dissolved electrolytes in the solvent the Hamiltonian of the system takes the form:

$$E = \frac{332}{\varepsilon \times r} + \frac{166}{a}\left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon_a}\right),$$

where dielectric constant ε is determined by the equation (17), r is the distance between the charges of carboxylate groups, ε_a is the dielectric constant of carboxylate groups at their reference states and $a = 2\,\mathring{A}$ is the effective radius of the charges.

The exact statistical mechanic and self-consistent field expressions for the pK shifts are given by:

$$\Delta pK^{exact} = -\frac{1}{2.303}\ln\left(\frac{\exp\left(-\frac{2.303\times q\times(pH-pK)}{RT}\right)+\exp\left(-\frac{2.303\times q\times(pH-pK)+E}{RT}\right)}{1+\exp\left(-\frac{2.303\times q\times(pH-pK)}{RT}\right)}\right)-q\times(pH-pK)$$

$\Delta pK^{self-cons.} = \frac{q\times\Phi}{2.303\times RT}$, where Φ is defined by equation (8).

The results are summarized in Table III-A where for comparison we give the calculations obtained by the FDPB method, Coulomb pair interactions with dielectric constants $\varepsilon = 80.0$ and $\varepsilon = r\times 10$ taken from the paper of Rajasekaran et. All.[42]

Table III-A.
pKa shifts in dibasic acids

dibasic acid(n)	experimental	pKa shifts self-onsistent field	exact	FDPB	ε=80.0	ε=r
oxalic acid(0)	2.36	1.71	1.70	5.41	0.98	25.49
malonic acid(1)	2.26	2.41	2.40	2.42	0.83	18.47
succinic acid(2)	0.84	0.70	0.71	0.82	0.58	9.00
glutaric acid(3)	0.47	0.54	0.56	0.49	0.47	5.98
adipic acid(4)	0.38	0.44	0.47	0.37	0.40	4.14
pimelic acid(5)	0.34	0.41	0.41	0.30	0.35	3.15
suberic acid(6)	0.28	0.36	0.36	0.27	0.31	2.49
azelaic acid(7)	0.26	0.32	0.32	0.24	0.27	1.89

The results are in very good agreement with the experimental data. The deviation in oxalic acid as already was noted are more probably connected with the sensitivity of the calculations to the radius of the atoms when the charges are close. In the work of Rajasekaran et. all[42] a value 1.87 have been proposed for the C-atom and a value 2.55 for the pK shift was obtained. With this correction our calculations give value 2.37 for the pK shift which is in an excellent agreement with the experiment. A little deviation in the case of long dibasic acids comes from the fact that the effective radius inside which the atom density is calculated is constant. In real dibasic acid with increasing the length between the carboxylic groups the dielectric constant have rapidly to increase to that of the solvent 80.0 as can be seen in Table III-A in the column of pK shifts calculated by the dielectric constant 80.0 and their correspondence to the experimental data.

Test of Debye-Huckel Expression for the Electrostatic Pair interactions on Subtilisin: Determination of pK_a Shifts of His64 on Removal of Asp99 and Glu156

A set of different residue volumes is used to analyze the influence of the permanent dipole of polar parts of the residue groups. The volume of amino-acid residues and some of their atomic groups are given in table III.[45] Spatial coordinates of residues are given by the centre of mass of their heavy atoms. The volumes used in calculations are divided into four groups. In the first group residues are taken as a whole; in the second group the volume of the polar part of side groups are subtracted; in the third group the volume of amino and carboxyl groups are subtracted; in the fourth group residue volumes are taken to be zero. In each group the residue centre of mass is defined between the remaining heavy atoms after deletion of the polar parts. In such a way it is possible to account for the influence of the polar parts of the residue on the dielectricconstant.

Table III

Residue pK[54] and volumes[45]

	volume($\AA^3$)			
residue	group 1	group 2	group 3	pK
A	88	88	55	
C	108	77	44	9.3
D	111	80	47	4.5
E	138	107	74	4.6
F	189	189	156	
G	60	60	27	
H	153	121	88	6.2
I	166	166	133	
K	168	155	122	10.4
L	166	166	133	
M	162	162	129	
N	117	104	71	
Y	193	184	151	9.7
P	122	122	89	
Q	143	130	97	
R	173	131	98	12.0
S	89	80	47	
T	116	107	74	
V	139	139	106	
W	227	227	194	

The centre of mass of the polar parts of the residue represents the spatial cordinate of residue charges and is defined between the heavy polar atoms. Local density around each

residue charge is calculated in the sphere with radius $8\,\mathring{A}$ with the space coordinate of nonpolar parts. The internal dielectric constant of residue groups in the case of non zero volumes is 4, the dielectric constant of the solvent medium is 80. The effective radius of the charge of ionized residues is $2\,\mathring{A}$.

A detailed experiment using site-directed mutagenesis has been carried out on subtilisin in order to evaluate the pK shifts on His64 in the active site of the molecule, caused by the replacement of Asp99 and Glu156 by serine residues.[3,35] Subtilisin is an excellent example for our test for the following reasons: Firstly, it is a relatively large protein with 275 residue groups among which 44 can ionize including the 'N' and 'C' terminus and the distances between the different pairs vary between $4\,\mathring{A}$ and $45\,\mathring{A}$. Secondly, this protein is already tested[2] on the ability of various theoretical methods to reproduce experimental data and this allows comparison of the present approach with already existing methods. Thirdly, the available experimental data obtained by Fersht and coworkers[3,35] covered a very wide range of ionic strengths which is very suitable for estimating the ability of the present approach to reproduce the ionic strength effects. The results of this approach and that of other methods, taken from the paper of Gilson and Honig[2], are summarized in table IV and table V. Calculations are carried out on the X-ray coordinates of the subtilisin molecule taken from the file 1sbt from the PDB data bank. In the file, 1sbt Asp99 is inverted with Ala98. In order to make our calculations compatible with that in the paper of Gilson and Honig we have used the same coordinates for Asp99 obtained by these authors. Table IV includes the results of the removal of Asp99, while in Table V, the results are summarized for the removal of Glu156. From the paper of Gilson and Honig the results chosen are obtained by the Tanford-Kirkwood (TK) and detailed PB methods. The TK method is carried out at three different depths for the charges, which is very similar to our test on different volumes of residue groups. What is more important is that all three methods are solutions of the linearized PB equation. TK and detailed PB are based on the assumption that charges are part of the low dielectric environment of the protein molecule and TK approximates the protein as a sphere, while the detailed PB takes advantage of detailed protein shape. Our present approach is based on the assumption that charges are part of the high dielectric medium. From the physical point of view, the solutions of PB outside and inside the protein molecule have to converge when charges are close to the protein boundary from both sides. This is exactly what one can see in table V and table VI. Our

results with zero volumes are close to the TK with depths 0.1 Å while those with volumes of group-3 are close to the TK with depths 1.0 Å. In the tables V and VI the values obtained in parenthesis include the Stern layer around each ionized residue. The thickness of the Stern layer cannot be varied to any great extent because of the restriction from the packing of residue groups. The maximal value is 1.5 Å. The results obtained with the Stern layer are better and cover the experimental results over the whole range of ionic strengths. The root-mean-square error of self-consistent and cluster field approaches relative to the experimental data in the case of Asp99 removal is 0.08 and in the case of Glu156 removal is 0.06, which can be compared with the experimental error ± 0.06.[36] For TK and DPB these values are 0.097 and 0.09 for Asp99 and 0.06 and 0.05 for Glu156. In conclusion, from the present results it follows that the solutions of PB with charges placed outside the molecule interior can be approximated with sufficient accuracy to the modified Debye-Huckel expression which depends on local characteristics rather than on detailed protein shape.

Table IV

His64 pK shift caused by the removal of Asp99

	Ionic strength (M)				
Method	0.005	0.010	0.025	0.100	0.500
Experiment[3,35]	0.38	0.42	0.36	0.26	0.10
Self-consistent field (zero residue volumes)	0.19 (0.20)	0.17 (0.16)	0.11 (0.14)	0.06 (0.10)	0.01 (0.02)
Self-consistent field (residue volumes from group 3)	0.31 (0.34)	0.28 (0.31)	0.22 (0.26)	0.12 (0.16)	0.05 (0.05)
Cluster-field (zero residue volumes)	0.21 (0.21)	0.16 (0.16)	0.13 (0.13)	0.05 (0.08)	0.01 (0.02)
Cluster-field (residue volumes from group 3)	0.31 (0.34)	0.28 (0.31)	0.22 (0.26)	0.12 (0.16)	0.03 (0.05)
Detailed Poisson-Boltzmann (Stern layer=2.0 Å, $\varepsilon_p = 2$, $\varepsilon_s = 80.0$)	0.31	0.29	0.25	0.18	0.10
Tanford-Kirkwood (depths=0.1 Å, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.24	0.19	0.15	0.09	0.04
Tanford-Kirkwood (depths=0.5 Å, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.26	0.21	0.17	0.12	0.07
Tanford-Kirkwood (depths=1.0 Å, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.32	0.27	0.24	0.19	0.14

ε_p - an interior dielectric constant of the protein molecule; ε_s - solvent dielectric constant; pK shifts of His64 are represented by their absolute values.

Table V
His64 pK shift caused by the removal of Glu156

Method	Ionic strength (M) 0.001	0.010	0.025	0.100
Experiment[3,35]	0.39	0.42	0.41	0.25
Self-consistent field (zero residue volumes)	0.24 (0.25)	0.17 (0.18)	0.14 (0.17)	0.07 (0.11)
Self-consistent field (residue volumes from group 3)	0.42 (0.44)	0.33 (0.37)	0.27 (0.32)	0.16 (0.21)
Cluster-field (zero residue volumes)	0.27 (0.29)	0.14 (0.18)	0.12 (0.16)	0.05 (0.09)
Cluster-field (residue volumes from group 3)	0.43 (0.45)	0.33 (0.37)	0.27 (0.32)	0.16 (0.22)
Detailed Poisson-Boltzmann (Stern layer=2.0 $\mathring{A}$, $\varepsilon_p = 2$, $\varepsilon_s = 80.0$)	0.42	0.37	0.34	0.27
Tanford-Kirkwood (depths=0.1 $\mathring{A}$, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.28	0.23	0.19	0.12
Tanford-Kirkwood (depths=0.5 $\mathring{A}$, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.31	0.26	0.22	0.16
Tanford-Kirkwood (depths=1.0 $\mathring{A}$, $\varepsilon_p = 2$, $\varepsilon_s = 78.0$)	0.39	0.34	0.31	0.25

Titration Curves

The self-consistent field approach is applied to determine the total average charge of native proteins at different pH and room temperature as well as the free energy difference between the native and denatured states. We chose three proteins which had been well-studied in previous experiments: lysozyme[46], ribonuclease-A[47] and ribonuclease-T1.[48] For the last two proteins (r nasa A and r nasa T1) there are excellent experimental studies[7] which allow comparison of the theoretical results with those of the experimental data of the electrostatic contribution to protein stability. For lysozyme only the titration curves of the native and denatured forms are used.[46] The crystallographic data are taken from the Protein Data Bank: lysozyme file 7lyz,[53] ribonuclease-A file 7rsa,[49] ribonuclease-T1 file 3rnt.[50]

Titration curves of ribonuclease-A and ribonuclease-T1, in addition to that of lysozyme, at different ionic strengths, are shown in fig.5 and fig.6 respectively.

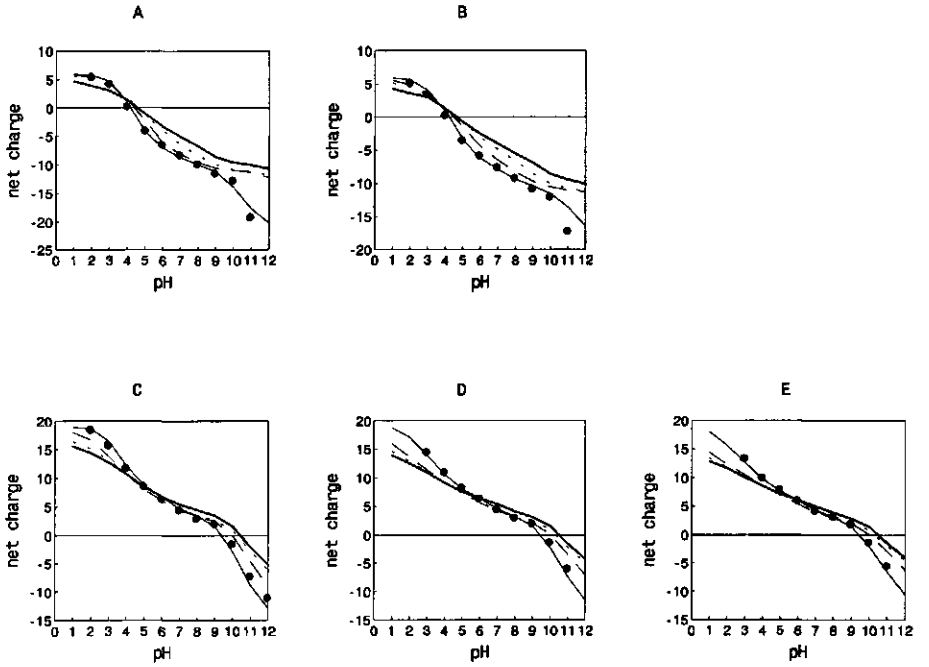


Figure-5. Experimental (●) and calculated titration curves of Ribonuclease A[47] and Ribonuclease T1[48] at different ionic strengths: 0.1(A); 0.01(B); 0.15(C); 0.03(D); 0.01(E). Theoretical curves: (——), calculated at volume group 1 (see table III); (- - - -) calculated at volume group 2; (– – –) , calculated at volume group 3; (——), calculated at zero volumes.

The pH dependence of total average charge is obtained from the expression of free energy in the molecular field: $F^{'} = \sum_i F_i^{'}$, where $F_i^{'}$ is the free energy of residue i in the molecular field $\Phi(\vec{r}_i)$.

$$Z = \frac{1}{2,3} \cdot \frac{\delta F^{'}}{\delta pH} = \frac{1}{2,3} \cdot \sum_i \frac{\delta F_i^{'}}{\delta pH} = \sum_i P_i^{'} \cdot q_i^{o}$$

It is seen that in the case of R Nasa A and R Nasa T1 the best fit with experimental results is obtained with zero residue volumes. Therefore the ionized residues are seen as part of the high dielectric medium. Lysozyme shows some deviation from this picture. The best results are obtained at second and third volume groups, as it shows lower polarity. This difference between the dielectric properties of proteins can be explained if the distributions of ionized residues among the local dielectric constants are observed. These distributions are shown in the fig.7 where the proteins are compared at each volume group.

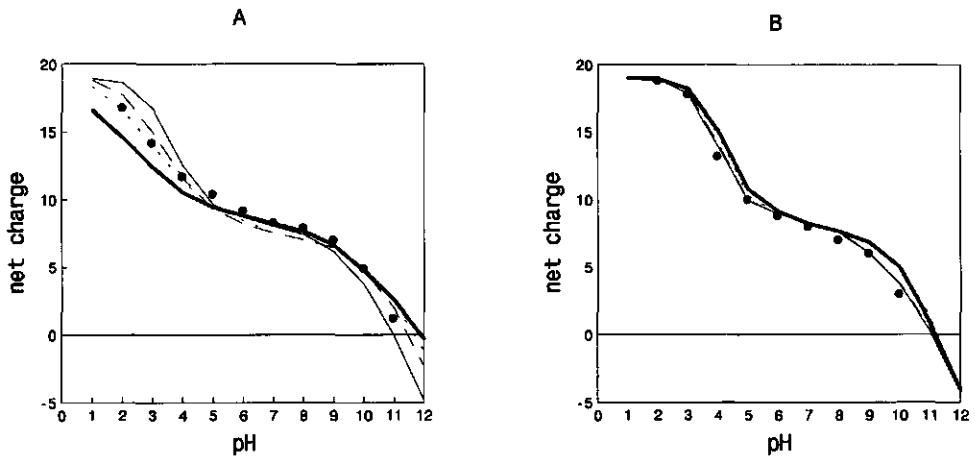


Figure-6. Experimental[46] and theoretical titration curves of hen egg lysozyme. For the native form experiment (●) is at ionic strength 0.1(A), for the denatured form at 6M GuHCl. Theoretical curves are calculated as follows: (——), calculated at volume group 1; (- - - -), calculated at volume group 2; (– – –), calculated at volume group 3; (——), calculated at zero volumes.

In fig.7A,B,C the distributions are obtained over all of the residues along the chain:

$$N(\varepsilon) = \sum_{\varepsilon_k=4}^{\varepsilon} \frac{n(\varepsilon_k)}{N} \text{ , where } 4 \le \varepsilon \le 80$$

where $n(\varepsilon_k)$ is the number of residues with the same local dielectric constant ε_k and N is the length of the protein chain. In fig.7D,E,F the distributions are only over the ionized residues. Although lysozyme is more compact as a whole, as is seen from fig.7A,B,C, it's ionized residues are associated predominantly with the aqueous phase rather than in R Nasa A and R Nasa T1-fig.7D,E,F. There is also an influence from the screening effect of the mobile ion atmosphere. Therefore a lower polar environment around ionized residues is necessary to compensate for the decrease of electrostatic interactions. As can be seen from fig.5A and fig.6 this effect is greater below and above $pH = 4$ and $pH = 10$, respectively. Within these ranges we have $2{,}3 \cdot RT \cdot q_i^{\circ} \cdot (pH - pK_i) \ge 0$ for both. At low salt concentration the Debye screening length is much longer than the effective radius of residue charges, such that the interaction of ionized residues with mobile ions can be neglected. Thus the probability of the charged form of the residue when $pH \approx pK$ depends only on the self energy of the Born transferring value and the electrostatic interactions between the titratable groups. Born effects always decrease the

robability of the charged form, whereas electrostatic interactions stabilize the charged form of acidic residues at $pH \leq 4$ and for basic residues at $pH \geq 10$.

Regions of the chain which are rich in ionized residues (loops or bends) are located on the surface of the protein globule. The deviation between different proteins comes from the depth of immersion of these parts inside the globule. It can be observed from fig.7D,E,F, where $N(\varepsilon)$ changes approximate linearly with ε and from one protein to the other, there is a shift along the ordinate.

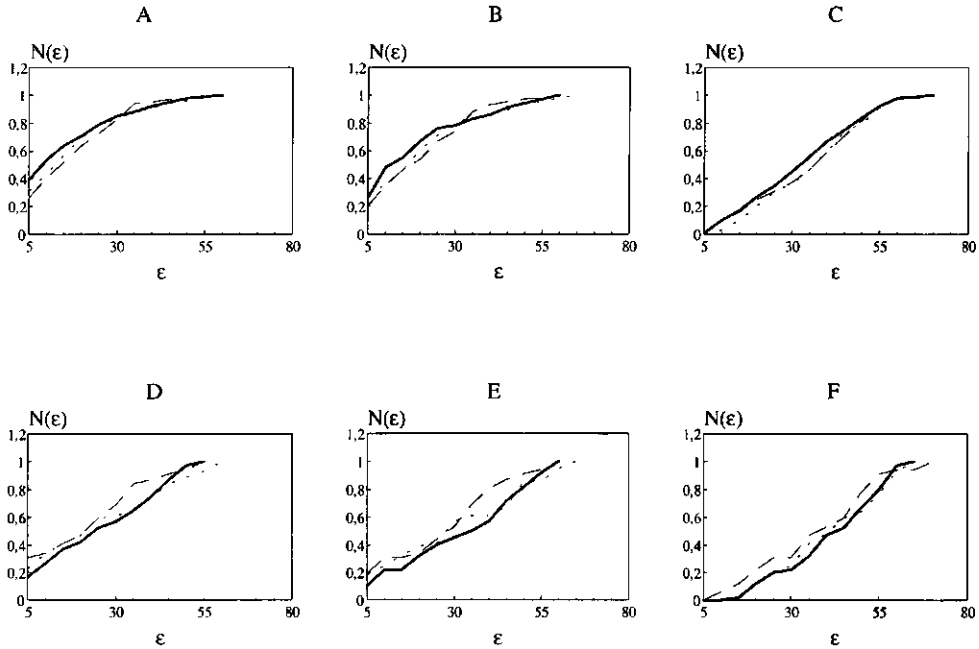


Figure-7. Local dielectric constant and it's distribution along the protein chains of Lysozyme (——), RNasa A (- - - -) and R Nasa T1 (— — —). In the case of A, B and C all residues along the chains are taken into account while in D, E and F only ionized residues are taken into account. Volume groups are represented as follows: A and D - group 1; B and E - group 2; C and F - group 3.

In fig.5B the experimental titration curve of the denatured form of lysozyme is given in 6M GuHCl. In this case, the Debye screening length is approximately $1\,\mathring{A}$. In these circumstances, even the ionized residues closest along the chain cannot interact. The overall charge of lysozyme depends on the individual titration of each ionized residue. In contrast, in the situation of low salt concentration, the Debye-Huckel self energy term from the mobile ion

environment now contributes a significant role to the stabilization of the charged form of residues.

Protein Stability

Free energy differences between the native and denatured states of R Nasa A and R Nasa T1, shown in fig.8, are obtained by the minimization of equation (7):

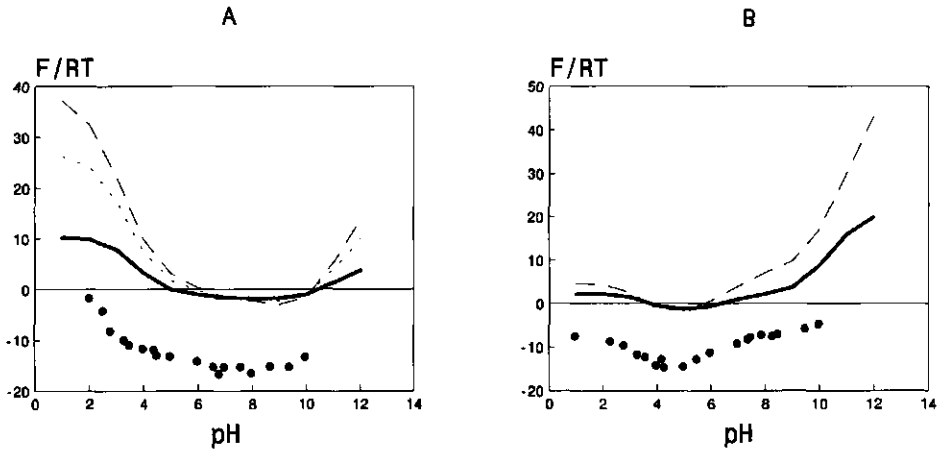


Figure-8. Electrostatic part of the free energy difference between the native and denatured states of R Nasa A and R Nasa T1. Experimental data[7] (●) for R Nasa A(A) are at ionic strength 0.15, for R Nasa T1(B) at ionic strength 0.1. Curves are calculated at zero residue volume: (———) ionic strength 0.15(A) and 0.1(B); (- - - -) ionic strength 0.03(A); (— —) ionic strength 0.01(A and B).

The use of PDB space coordinates shows that local motions of residues do not disrupt the good relationship between the theoretical and experimental results (fig.5 and fig.6). However in denatured states the picture is different. Long-range pair contacts have an entropic contribution which is restricted only by the chain length placed between them. The characteristic length connected with the electrostatic interactions can be introduced on the assumption that $\frac{\Delta F_{ij}}{RT} \approx 1$. Therefore we have $\Delta r_{ij} = \frac{\langle E_{ij} \rangle - RT \cdot S_{ij}}{RT}$, where E_{ij} is the electrostatic interaction between the residues i and j. A simple estimation based only on the $\frac{\langle E_{ij} \rangle}{RT}$ shows that $(\Delta r_{ij})_{max} \approx 8\,\mathring{A}$. In the denatured form, only the electrostatic interactions

between residues can be considered which are close within the chain but not further than 3 residues apart. Because in this situation the deviation of $\vec{r}_{ij}$ (distance between the ionized residues along the chain) is small, relative to that in the native state, it is possible to approximate the denatured state using the coordinates from the native form. Minimization of free energy is also based on equation (7). Thus, the electrostatic part of the free energy difference between the native and denatured form is expressed as:

$$\Delta F^{el} = F^{native} - F^{denat}$$

The experimental results in fig.8 are obtained assuming that the total free energy difference (including not only the electrostatic interactions) is proportional to the concentration of denaturant. The expected values of free energy can be obtained by extrapolation to zero concentration of denaturant. If it is assumed that the non electrostatic part of the protein chain interactions in the denatured state does not change on average from one conformation to the other it is possible to anticipate that the electrostatic part of the free energy is shifted at constant values ≥ 0 relative to the experimental one. This is observed in fig.8. With decreasing ionic strength the free energy minimum becomes sharper and deeper.

pKa Shifts

For several proteins such as bovine pancreatic trypsin inhibitor, ribonuclease A , ribonuclease T1, hen egg white lysozyme and T4 lysozyme, detailed experimental data on the individual pKa's are available. In this study, we present the results based on the self-consistent field and cluster field approaches, and, where available, the results reported by the other authors. The calculated pKa sets for each protein are compared with the experimental data using the root-mean square deviation. From the analyses of the section on titration curves and protein stability, the pK calculations are carried out as follows: for hen egg white lysozyme the volumes of group-3 are used; for all other proteins zero volumes are used. Apart from ribonuclease A where the Debye radius is $6.8\,\mathring{A}$, a value $8.0\,\mathring{A}$ is used for other proteins in order to fit the experimental conditions.

Table VI A, B presents the results on pK shift calculations for triclinic and tetragonal crystal forms of HEWL. In both cases, self-consistent field gives higher values for the r.m.s. compared to the null model. Explanation for this result is the incorrect predicted pK shifts of Tyr23 and Tyr53. This occurs because close to each of these tyrosine groups there are basic ionized

groups as follows: Arg44(5.32 Å) and Arg60(7.71 Å) around Tyr23; Arg61(12.9 Å), Arg45(9.9 Å) and Arg68(8.7 Å) around Tyr53.

Table VI

A. pKa's in triclinic crystal structure of lysozyme

Residue	Expt.[56,22]	Model					
		Null	Self-consistent field	Cluster field	Antosiewicz et al	Yang and Honig[23]	Bashford and Karplus[22]
Lys1	10.6	10.4	11.3	11.4	10.1	10.2	9.6
Glu7	2.6	4.4	2.5	2.5	3.1	3.4	2.1
Lys13	10.3	10.4	11.6	11.8	10.7	12.0	11.6
His15	5.8	6.3	5.4	5.9	5.1	6.7	4.0
Asp18	2.8-3.0	4.0	3.0	3.0	2.9	4.0	3.1
Tyr20	10.3	9.6	10.6	10.1	10.2		14.0
Tyr23	9.8	9.6	14.4	11.6	9.7		11.7
Lys23	10.4	10.4	9.9	10.3	10.5	10.2	9.6
Glu35	6.1	4.4	4.1	4.0	4.4	5.6	6.3
Asp48	4.3	4.0	2.8	2.8	3.4	1.6	1.0
Asp52	3.4-3.7	4.0	3.9	3.7	3.4	5.2	7.0
Tyr53	12.1	9.6	16.2	12.1	11.0		20.8
Asp66	1.5-2.5	4.0	3.4	3.4	2.1	3.1	1.7
Asp87	3.5-3.75	4.0	2.6	2.6	2.6	1.6	1.2
Lys96	10.7	10.4	10.4	10.8	11.2	10.5	10.4
Lys97	10.1	10.4	10.7	10.9	11.0	11.4	10.6
Asp101	4.0-4.5	4.0	3.2	3.1	3.7	6.4	7.9
Lys116	10.2	10.4	10.1	10.5	10.0	10.6	9.9
CTER	2.7-3.1	3.8	2.8	2.7	2.9	2.6	2.3
r.m.s.d.		1.0	1.65	0.95	0.67	1.3	2.7

Expt. - experimental pKa's (Kuramitsu and Hamaguchi[56], Bashford and Karplus[22]); Null - intrinsic pKa's; r.m.s.d. - root-mean-square deviation; CTER ,NTER - 'C' and 'N' terminus of the chain; "<x" - the deviation is set to zero and computed relative to "x" if the calculated value is greater than "x"; - "<x<" - the deviations are computed relative to the midpoint.

This is the situation when at low dielectric environment and high repulsive interactions the self-consistent field leads to errors. If these two groups are not accounted for in the r.m.s. estimation the result is exactly that of the null model. This implies that only the closest amino acids in the space contribute to such effects.

The cluster field approach can correct for such basic environment of tyrosine groups but nevertheless the overall r.m.s. deviation is still slightly lower than that of the null model. One of the reasons is that all ionized groups are seen simultaneously as part of the solvent or boundary. This is appropriate for our analysis of the approximations made in the solution of the

PB equation but at lower dielectric constants, overestimation of some of the short-range interactions occurs.

Table VI

B. pKa's in tetragonal crystal structure of lysozyme

Residue	Expt. [56,22]	Model					
		Null	Self-consistent field	Cluster field	Antosiewicz et al	Yang and Honig[23]	Bashford and Karplus[22]
Lys1	10.6	10.4	11.4	11.5	10.8	12.1	10.8
Glu7	2.6	4.4	2.6	2.6	3.0	2.4	1.2
Lys13	10.3	10.4	11.8	11.9	11.4	11.5	10.1
His15	5.8	6.3	4.9	5.9	4.8	5.6	2.4
Asp18	2.8-3.0	4.0	3.1	3.1	3.0	3.0	2.6
Tyr20	10.3	9.6	12.2	11.1	9.6		12.1
Tyr23	9.8	9.6	15.5	12.0	9.4		10.1
Lys23	10.4	10.4	9.8	10.3	10.2	10.4	7.7
Glu35	6.1	4.4	4.2	4.1	4.4	5.4	6.2
Asp48	4.3	4.0	3.4	3.3	3.3	3.6	1.6
Asp52	3.4-3.7	4.0	4.2	3.9	5.2	7.2	8.5
Tyr53	12.1	9.6	15.7	11.8	11.2		18.8
Asp66	1.5-2.5	4.0	3.8	3.8	2.8	5.6	2.2
Asp87	3.5-3.75	4.0	2.8	2.7	2.9	2.2	0.8
Lys96	10.7	10.4	10.1	10.4	10.9	10.0	8.9
Lys97	10.1	10.4	10.7	11.0	10.8	11.0	8.4
Asp101	4.0-4.5	4.0	3.3	3.3	4.3	6.8	4.3
Lys116	10.2	10.4	10.0	10.4	10.4	10.4	9.7
CTER	2.7-3.1	3.8	2.5	2.5	2.0	0.5	2.2
r.m.s.d.		1.0	1.84	1.0	0.83	1.7	2.5

It is clear that the same difficulties also exist for other methods as can be seen from table VI A,B. Only the PB approach of Antosiewicz et al[1] significantly beats the null model. Table VII,VIII and IX show the results on pK shift calculations for BPTI, R-Nasa A and R-Nasa T1 respectively. The results are in excellent agreement with the experimental data. For BPTI in table VIII the overall r.m.s. deviation is 0.31(0.27) for self-consistent field and 0.29(0.24) for cluster field approach compared to the r.m.s. of the null model 0.60. In parentheses are shown results on r.m.s. when experimental pKa's of Asp3 and Asp50 are reversed because the NMR could not distinguish these two groups. The results obtained with self-consistent and cluster field approaches can be compared with the best PB model of Antosiewicz et al..[1]

Table VII
pKa's in BPTI

Residue	Expt. [57,58,59]	Model				
		Null	Self-consistent field	Cluster field	Antosiewicz et al	Yang et al. [4]
NTER	8.1	7.5	7.6	7.9	7.1	7.0
Asp3	3.0	4.0	3.5	3.4	3.4	3.6
Glu7	3.7	4.4	3.7	3.6	3.7	3.4
Lys15	10.6	10.4	10.5	10.8	10.5	10.7
Lys26	10.6	10.4	10.6	10.9	10.5	10.8
Lys41	10.8	10.4	11.2	11.3	10.8	10.3
Lys46	10.6	10.4	10.5	10.8	10.2	10.3
Glu49	3.8	4.4	4.1	3.9	3.9	4.5
Asp50	3.4	4.0	3.2	3.1	2.7	1.7
CTER	2.9	3.8	3.3	3.2	3.5	3.6
r.m.s.d.		0.60	0.31	0.29	0.47	0.77
r.m.s.d.		0.60	0.27	0.24	0.40	0.66

Table VIII
pKa's in RNasa A

Residue	Expt. [60]	Model			
		Null	Self-consistent field	Cluster field	Antosiewicz et al
NTER	7.6	7.5	7.5	7.9	7.0
Glu2	2.8	4.4	3.2	3.0	2.5
Glu9	4.0	4.4	4.1	3.8	4.1
His12	6.2	6.3	5.9	6.5	4.5
Asp14	<2.0	4.0	3.1	3.0	1.9
Asp38	3.1	4.0	3.0	2.8	2.8
His48	6.0	6.3	6.7	6.9	6.5
Glu49	4.7	4.4	4.2	3.9	4.6
Asp53	3.9	4.0	3.8	3.5	3.6
Asp83	3.5	4.0	3.2	3.0	2.0
Glu86	4.1	4.4	3.9	3.7	3.8
His105	6.7	6.3	6.6	6.8	6.2
Glu111	3.5	4.4	4.0	3.8	3.8
His119	6.1	6.3	6.5	6.7	5.9
Asp121	3.1	4.0	3.1	2.9	1.5
CTER	2.4	3.8	3.0	2.8	2.3
r.m.s.d.		0.86	0.44	0.50	0.75

Table IX
pKa's in RNasa T1

Residue	Expt. [61,62]	Model			
		Null	Self-consistent field	Cluster field	Antosiewicz et al
His27	7.3	6.3	7.1	7.2	6.7
His40	7.9	6.3	6.9	7.0	7.5
His92	7.8	6.3	6.6	6.9	6.9
Glu58	4.3	4.4	3.3	3.3	2.5
r.m.s.d.		1.2	0.93	0.81	1.1

The results of these authors are as follows: for BPTI 0.47(0.40); for R-Nasa A 0.75 and for R-Nasa T1 1.1. For BPTI, R-Nasa A and R-Nasa T1 the self-consistent and cluster field approaches significantly beat both the null and the best FDPB approaches.

Table X
pKa's in T4 lysozyme

Residue	Expt. [63]	Model				
		Null	Self-consistent field	Cluster field	Antosiewicz et al.[1]	Yang et al.[4]
His31	9.1	6.3	7.5	8.7	8.0	8.0
Asp70	<1.4	4.0	0.6	0.6	2.6	3.3

Table X shows the results for the salt bridge His31-Asp70 in T4 lysozyme. The calculated pK shifts are in excellent agreement with the experimental data and beat the null model as well as the PB approaches of Antosiewicz et al.[1] and Yang et al..[4] Our present results are also consistent with the fact that the bridge is buried. The best agreement with the experiment is obtained with volumes of group 2. This shows that the account of hydrophobic environment is very important to fit the experimental data. This is confirmed by recent investigations on pK shifts in synthetic peptides.[51]

In conclusion, a self-consistent field theory for treating the electrostatic interactions in proteins is presented. This includes, not only the native state, but arbitrary chain conformations and the possibility of predictions of a variety of pH dependent properties: individual pK and pK shifts for the ionized groups in addition to the wide range of ionic strengths; titration curves and the free energy of the electrostatic part of protein stability. The main results are: Firstly, the approximate solutions of PB equations, in the case when ionized residues are seen as part of the high dielectric medium, rather than the interior of the protein molecule are in excellent agreement with the experimental data. This is also supported by the recent theoretical calculations of the dielectric constant of proteins.[40,41] The results obtained show that when charged portions of the charged side chains are viewed as part of the solvent medium the theoretical analysis and molecular dynamic simulations are consistent. Secondly, the solutions of PB equation outside the protein interior, depend on local characteristics, such as the packing

of chain portions around ionized residues rather than on the detailed shape of the protein molecule. Lastly, the contribution of electrostatic interactions at neutral conditions to the free energy difference between the unfolded and folded states of protein molecules is close to zero. This indicates that the main driving forces for folding of protein molecules under these conditions are hydrophobic and backbond-backbond hydrogen bonding interactions.

REFERENCES

1 Antosiewicz, J., McCammon J., A., Gilson, M., K. Prediction of pH-dependent Properties of Proteins J. Mol. Biol. 238:415-436, 1994.

2 Gilson, M. K., Honig, B. Energetics of charge-charge interactions in proteins. Proteins: Struct. Func. Genet. 3:32-52, 1988.

3 Russell, A. J., Thomas, P. G., Fersht, A. R. Electrostatic Effects on Modification of Charged Groups in the Active Site Cleft of Subtilisin by Protein Engenieering. J. Mol. Biol. 193:803-813, 1987

4 Yang, A. S., Gunner, M. R., Sampogna, R., Sharp, K., Honig, B. On the calculation of pKa's in proteins. Proteins: Struct. Func. Genet. 15:252-265, 1993.

5 Bashford, D., Karplus, M. Multiple-site titration curves of proteins: An analysis of exact and approximated methods for their calculation. J. Phys. Chem. 95:9556-9561, 1991.

6 Linderstrom-Lang, K. C. R. Trav. Lab. Carlsberg 15:7-14, 1924

7 Pace, C. N., Laurents, D. V., Thomson, J. A. pH Dependence of the Urea and Guanidine Hydrochloride Denaturation of Ribonuclease A and Ribonuclease T1. Biochemistry 29:2564-2572, 1990

8 Hu, C.-Q., Sturtevant, J. M., Thomson, J. A., Erickson, R. E., Pace, C. N. Thermodynamics of ribonuclease T1 denaturation. Biochemistry 31:4876-4882, 1992

9 Nakamura, H., Sakamoto, T., Wada, A. A theoretical study of the dielectric constant of protein. Protein Engineering 2:177-185, 1988

10 Gilson, M. K., Honig, B. H. The dielectric constant of a folded protein. Biopolymers 25:2097-2119, 1986.

11 Kirkwood, J. G. Theory of solutions of molecules containing widdely separated charges with special applications to zwiterions. J. Chem. Phys. 2:351-361, 1934.

12 Tanford, C., Kirkwood, J. G. Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres. J. Amer. Chem. Soc. 79:5333-5339, 1957

13 Shire, S. J., Hanania, G. I. H., Gurd, F. R. N. Electrostatic Effects in Myoglobin. Hydrogen Ion Equilibria in Sperm Whale Ferrimyoglobin. Biochemistry 13:2967-2974, 1974

14 Friend, S. H., Gurd, F. R. M. Electrostatic Stabilization in Myoglobin. pH Dependence of Summed Electrostatic Contributions. Biochemistry 18:4612-4619, 1979

15 Lee, B., Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. J. Mol. Biol. 55:379-400, 1971

16 Russell, S. T., Warshel, A. Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. J. Mol. Biol. 185:389-404, 1985.

17 Orttung, W. H., Extension of the Kirkwood-Westheimer Model of Substituent Effects to General Shapes, Charges, and Polarizabilities. Application to the Substituted Bicyclo[2.2.2]octanes. J. Amer. Chem. Soc., 100:4369-4375, 1978

18 Warwicker, J., Watson, J. H. C. Calculation of the Electric Potential in the Active Site Cleft due to α-Helix Dipoles. J. Mol. Biol. 157:671-679, 1982

19 Klapper, I., Magstrom, R., Fine, R., Sharp, K., Honig, B. Focusing of Electric Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-Acid Modification. Proteins 1:47-51, 1986

20 Rogers, N. K., Moore, G. R., Sternberg, M. J. E. Electrostatic Interactions in Globular Proteins: Calculations of the pH Dependence of the Redox Potential of Cytochrom c_{551}. J. Mol. Biol. 182:613-616, 1985

21 Zauhar, R. J., Morgan, R. S. A New Method for Computing the Macromolecular Electric Potential J. Mol. Biol. 186:815-820, 1985

22 Bashford, D., Karplus, M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry, 9:327-335, 1990.

23 Yang. An-S., Honig, B. On the pH Dependence of Protein Stability J. Mol. Biol. 231:459-474, 1993.

24 Oberoi, H., Allewell, N. M., Multigrid solution of the nonlinear Poisson-Boltzmann equation and calculation of titration curves. Biophys. J. 65:48-55, 1993.

25 Gilson, M. K., Rashin, A., Fine, R., Honig, B. On the Calculation of Electrostatic Interactions in Proteins J. Mol. Biol. 183:503-516, 1985.

26 Hendsch. Z. S., Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. Proteins Science 3:211-226, 1994.

27 Smith, K. C., Honig, B. Evaluation of the conformational free energies of loops in proteins. Proteins: Struct. Func. Genet. 18:119-132, 1994.

28 Abagian, R., Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983-1001, 1994.

29 Tanford, C., Roxby, R. Interpretation of protein titration curves. Aplication to lysozyme. Biochemistry 11:2192-2198, 1972.

30 Gilson, M. K. Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups. Proteins: Struct. Func. Genet. 15:266-282, 1993.

31 Beroza, P., Fredkin, D. R., Okamura, M. Y., Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides*. Proc. Nat. Acad. Sci., U. S. A. 88:5804-5808, 1991.

32 Stitger, D., Dill, K. A. Charge Effects on Folded and Unfolded Proteins. Biochemistry 29:1262-1271, 1990.

33 Stitger, D., Alonso, D. O. V., Dill, K. A. Protein stability: Electrostatics and compact denatured states. Proc. Nat. Acad. Sci., U. S. A. 88:4176-4180, 1991.

34 Breslow, E., Gurd, F. R. N. J. Biol. Chem. 237:371-381, 1962.

35 Thomas, P.G., Russell, A. J., Fersht, A. R. Tailoring the pH dependence of enzyme catalysis using protein engineering. Nature 318:375-376, 1985.

36 Gilson, M. K., Honig, B. H. Calculation of electrostatic potentials in an enzyme active site. Nature 330:84,1987

37 Gibbs, J. W. "Elementary Principles in Statistical Mechanics", New Haven, 1902.

38 Gilson, M. K., Honig, B. Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvetion Energies, Binding Energies, and Conformatinal Analysis Proteins: Struct. Func. Genet. 4:7-18, 1988.

39 Gilson, M. K., Sharp, K. A., Honig, B. Calculation the Electrostatic Potential of Molecules in Solution: Method and Error Assessment J. Comp. Chem. 9:327-335, 1987.

40 Smith, P. E., Brunne, R. M., Mark, A. E., van Gunsteren, W. F. Dielectric Properties of Trypsin Inhibitor and Lysozyme Calculated from Molecular Dynamics Simulations. J. Phys. Chem. 97:2009-2014, 1993

41 Simonson, T., Perahia, D. Internal and interfacial dielectric proparties of cytohrome c from molecular dynamics in aqueous solution. Proc. Nat. Acad. Sci., U. S. A. 92:1082-1086, 1995.

42 Honig , B., Hubbell, W. Stability of "salt bridges" in membrane proteins Proc. Natl. Acad. Sci. USA 81:5412-5416, 1984.

43 Landau, L. D., Lifshitz, E. M. "Electrodynamics of Continuous Media", Nauka Publishers: Moscow, 1992 (in Russian)

44 Leontovich, M. L. "Introduction to thermodynamics. Statistical physics.", Nauka Publishers: Moscow, 1983 (in Russian)

45 Zamyatnin, A. A. Protein volume in solution. Progress in Biophysics and Molecular Boilogy 24:109-123, 1972

46 Tanford, C., Roxby, R. Interpretation of protein titration curves. Application to lysozyme. Biochemistry 11:2192-2198, 1972

47 Tanford, C., Haunstein, J. D. Hydrogen Ion Equilibria of Ribonuclease. J. Amer. Chem. Soc. 78:5287-5291, 1956

48 Iida, S., Ooi, T. Titration of Ribonuclease T_1. Biochemistry 8:3897-3902, 1969

49 Wlodawer, A., Svensson, L. A., Sjolin, L., Gilliland, G. L. Structure of Phosphate-Free Ribonuclease A Refined at 1.26Å. Biochemistry 27:2705-2717, 1988

50 Kostrewa, D., Choe, H. W., Heinemann, U., Saenger, W. Crystal Structure of Guanisine-Free Ribonulease T_1, Complexed with Vanadate(V), Suggests Conformational Change upon Substrate Binding. Biochemistry 28:7592-7600, 1989

51 Urry, D. W., Gowda, D. C., Peng, S., Parker, T. M., Jing, N., Harris, R. D. Nanometric design of extraordinary hydrophobic-induced pKa shifts for aspartic acid: relevance to protein mechanisms. Biopolymers 34:889-896, 1994.

52 McQuarrie, D. A. "Statistical Mechanics", Harper and Row, New York, 1976.

53 Herzberg, O., Sussman, J. L. Protein model building by the use of a constrained-restrained least-squares procedure. J. Appl. Crystallogr. 16:144-150, 1983

54 Greighton, T. E. "Proteins: Structures and Molecular Principles", W.H.Freeman, New York, 1983

55 Callen, H. B. "Ttermodynamics and an introduction to thermostatistics", John Wiley & Sons, New York, 1985.

56 Kuramitsu, S., Hamaguchi, K. Analysis of the acid-base titration curve of hen lysozyme. J. Biochem. 87:1215-1219, 1980.

57 Brown, L. R.., Marco, A. D., Wagner, G., Wüthrich, K. A study of the lysyl residues in the basic pancreatic trypsin inhibitor using ^{1}H nuclear magnetic resonance at 360 Mhz. Eur. J. Biochem. 62:103-107, 1976.

58 Brown, L. R.., Marco, A. D., Richarz, R., Wagner, G., Wüthrich, K. The influence of a single salt bridge on static and dynamic features of the globular solution conformation of the basic pancreatic trypsin inhibitor. Eur. J. Biochem. 88:87-95, 1978.

59 Richarz, R., Wüthrich, K. High-field ^{13}C nuclear magnetic resonance studies at 90.5 Mhz of the basic pancreatic trypsin inhibitor. Biochemistry, 17:2263-2269, 1978.

60 Rico, M., Santoro, J., Gonzalez, C., Bruix, M., Neira, J. L. Solution structure of bovine pancreatic ribonuclease A and ribonuclease-pyrimidine nucleotide complexes as determined by ^{1}H NMR. In Structure, Mechanism and Function of Ribonucleases. Procedings of the 2nd International Meeting held in Sant Felin de Guixols, Girona, Spain, 1990. (Cuchillo, C. M., de Liorens, R., Nogués, M. V., Parés, X. Eds.). pp. 9-14. Bellatrra, Spain. Department de Bioquimica i Bioilogia Molecular and Institut de Biologia Fonamental Vicent Villar Palasi, Universitat Autonomià de Barcelona.

61 Inagaki, F., Kawaano, Y., Shimada, I., Takahashi, K., Miyazawa, T. Nuclear magnetic resonance study on the microenvironments of histidine residues of ribonuclease T1 and carboxymethylated ribonuclease T1. J. Biochem. 89:1185-1195, 1981.

62 Shirley, B. A., Stanssen, P., Steyaert, J., Pace., C. N. Conformational stability and activity of ribonuclease T1 and mutants. J. Biol. Chem. 264:11621-11625, 1989.

63 Anderson, D. E., Becktel, W., J., Dahlquist, F. W. pH-induced denaturation of proteins a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry, 29:2403-2408, 1990.

4 THE NMR SOLUTION STRUCTURE AND CHARACTERIZATION OF PH DEPENDENT CHEMICAL SHIFTS OF THE β-ELICITIN, CRYPTOGEIN.

Paul R. Gooley, Max A. Keniry, Roumen A. Dimitrov, Danny E. Marsh, David W. Keizer, Kenwyn R. Gayler & Bruce R. Grant

In press: *Journal of Biomolecular NMR*

SUMMARY

The NMR structure of the 98 residue β-elicitin, cryptogein, which induces a defence response in tobacco, was determined using ^{15}N and $^{13}C/^{15}N$ labelled protein samples. In aqueous solution conditions in the millimolar range the protein forms a discrete homodimer where the N-terminal helices of each monomer form an interface. The structure was calculated with 1047 intrasubunit and 40 intersubunit NOE derived distance constraints and 236 dihedral angle constraints for each subunit using the molecular dynamics program DYANA. The twenty best conformers were energy-minimized in OPAL to give a root-mean-square deviation to the mean structure of 0.82 Å for the backbone atoms and 1.03 Å for all heavy atoms. The monomeric structure is nearly identical to the recently derived x-ray crystal structure, (backbone rmsd 0.86Å for residues 2 to 97) and shows five helices, a two stranded antiparallel β-sheet and an Ω-loop. Using $^{1}H,^{15}N$ HSQC spectroscopy the pKa of the N- and C-termini, Tyr-12, Asp-21, Asp-30, Lys-61, Asp-72, Tyr-85 and Lys-94 were determined and support the proposal of several stabilizing ionic interactions including a salt bridge between Asp-21 and Lys-62. The hydroxyl hydrogens of Tyr-33 and Ser-78 are clearly observed indicating that these residues are buried and hydrogen bonded. Two other tyrosines, Tyr-47 and -87, are also removed from the solvent and show pKa's > 12, however, there is no indication that their hydroxyls are hydrogen bonded. Calculation of theoretical pKa's show general agreement with the experimentally determined values and are similar for both the crystal and solution structures.

INTRODUCTION

One of more challenging questions in the biology of the interactions between plants and pathogens, is the identification of the molecules that interact to initiate defence responses. Molecules from each partner in the interaction presumably bind, resulting in the induction of signal cascades which lead to the generation of an antimicrobial environment and successful repulsion of the attempted invasion. Generally, the complexity and heterogeneity of the molecules produced by many microorganisms which elicit such plant response, has hindered the determination of their structures. However, members of a subgroup of such elicitors which are peptides of proteins have been purified, cloned and/or sequenced, thus enabling the examination of their structure/function properties (De Wit, 1992; Kooman-Gersmann et al., 1996; Jones et al., 1994).

One group of protein elicitors, the elicitins, have properties which make them particularly well-suited as model elicitors to use in the study of structure/function relationships in their interaction with plants. Elicitins are small proteins that are secreted by members of the genus *Phytophthora* and by a limited number of species in the genus *Pythium* (Huet et al., 1995; Gayler et al., 1997). Preliminary studies have shown that they belong to the group of proteins described as cysteine-knot proteins (Myers et al., 1993), but show no homology to other known families of proteins. While the role of elicitins in the biology of the microorganisms which secrete them remains in doubt (Grant et al., 1996) these proteins have the capacity to act as elicitors of both the hypersensitive response and systemic acquired resistance in tobacco (Bonnet et al., 1996) and in some members of the family *Cruciferae*, particularly radish (Kamoun et al., 1993). Their capacity to act as elicitors in these species, together with their small size and highly conserved amino acid sequences, make elicitins ideally suited as compounds in which the effect of substitution of specific amino acids on the three dimensional structures can be determined and related to alterations in biological activity.

Moreover, in radish the presence of different genotypes which show differences in their response to elicitins (Keizer et al., 1998) opens the possibility of the isolation of specific proteins which confer elicitin-sensitivity, and in turn the possibility of determining how these latter proteins act as receptors for elicitins, and the ways in which elicitin and such receptors interact. Recent studies by Wendehenne et al (1995) have indicated specific elicitin binding sites in membrane proteins isolated from tobacco and it seems probable that the same will

prove to be true in the radish system.

Determination of the three-dimensional structure and the functionally important residues of the elicitins is a problem of importance. The crystal and solution structures of cryptogein have been reported by Boissy, et al (1996) and Fefeu et al. (1997), respectively. The crystal structure showed a unique fold which consisted of six helices, a two stranded β-sheet and an Ω-loop, where the loop and sheet formed a «beak motif». The protein also showed two salt bridges: the N-terminus to Asp-72, and Asp-21 to Lys-62. The solution structure was similar, but showed some differences in the 64-70 and 83-88 region and the salt bridge between Asp-21 and Lys-62 was not apparent.

The N-terminus, Asp-21 and Asp-72 are amongst the least conserved residues of which include other ionizable residues, Lys-13, Lys-39, Lys-48, Lys-61, Tyr-85 and Lys-94. At least some of these surface residues are expected to be functionally important, for example, substitution of Lys-13 with Val reduces the activity of cryptogein in tobacco by 100-fold (O'Donohue et al., 1995). Characterization of the surface properties, such as the ionization states of surface residues, of elicitins will assist in understanding the molecular interactions which these proteins can undergo. NMR spectroscopy is the method of choice for determining pKa's of individual residues (Forman-Kay et al., 1992; Nakamura, 1996). As the N-H bond is more easily polarized than the C-H bond, the ^{15}N resonance shows large chemical shift changes with pH and therefore the 2D ^{1}H, ^{15}N HSQC experiment is an excellent method for analysing pH dependencies of chemical shift.

In this report we describe the refined solution structure of cryptogein, and show that under millimolar solution conditions the protein forms a dimer. The individual subunits, however, are nearly identical to the crystal structure. The experimental pKa's of a number of ionizable residues are determined and compared to the theoretical pKa's calculated with both the solution and crystal structures.

MATERIALS AND METHODS

Phytophthora Growth Conditions and Cryptogein Purification

Cryptogein was purified from liquid cultures of *P. cryptogea* Pethybr & Laff (Isolate P7407,

supplied by Professor M. Coffey, University of California). For isotopic labelling, *P. cryptogea* was grown in 500 ml of modified High Phosphate Ribeiro medium (Fenn & Coffey, 1984) pH 6.2, which contained 50 mM ^{13}C-glucose as a carbon source, 3 mM ^{15}N-ammonium chloride and 1.5 mM potassium ^{15}N-nitrate, as nitrogen sources. For ^{15}N labelled cryptogein, cultures of *P. cryptogea* with only ^{15}N-labelled substrates were prepared. Cultures were grown in the light for 21 days at 26°C.

Culture medium was harvested by filtration through two GF/C filters (Whatman) to separate mycelia from the culture filtrate. Crude culture filtrate was concentrated and subjected to ammonium sulfate (55%) precipitation. Cryptogein remained in the supernatant, which was desalted by dialysis against 10 mM sodium acetate. Cryptogein was then separated by cation exchange chromatography using a S Hyper D column (Beckman).

NMR spectroscopy

NMR measurements were performed at 40°C on either ^{15}N-labelled or ^{15}N,^{13}C double labelled protein on Varian INOVA-400 and Varian INOVA-600 spectrometers. Cryptogein was dissolved to a concentration of 1 mM in 90% H_2O/10% ^2H_2O, at pH 6.8. The following experiments were collected using the ^{15}N-labelled sample at 400 MHz: 2D ^{1}H,^{15}N-HSQC, 3D ^{1}H,^{15}N-TOCSY-HSQC (40 ms mixing time), NOESY-HSQC (100 ms mixing time) (Zhang et al., 1994), HSQC-NOESY-HSQC (100 ms mixing time) (Frenkiel et al, 1990), HNHA (Vuister and Bax, 1993) and HNHB (Archer et al., 1991); and the following experiments were collected with the ^{13}C,^{15}N-double-labelled sample at 600 MHz: 3D CBCA(CO)NH (Muhandiram and Kay, 1994), ^{13}C-HCCH-TOCSY (Kay et al., 1993), ^{13}C HSQC-NOESY (100 ms mixing time) (Majumdar and Zuiderweg, 1993) and ^{13}C HSQC-ROESY (25 ms mixing time). 2D and 3D ^{13}C-edited, ^{12}C-filtered NOESY spectra (120 ms mixing time) (Vuister et al., 1994; Folmer et al., 1995) were acquired on samples of mixed labelled (1:1 of ^{13}C/^{15}N and ^{12}C/^{14}N). All spectra were processed with NMRPipe (Delaglio et al., 1995). ^{1}H chemical shifts were referenced to residual water at 4.64 ppm relative to TSP at 40°C. The corresponding ^{13}C and ^{15}N reference frequencies were calculated from the ^{1}H spectrometer frequency. All spectra were analysed by NMRView (Johnson and Blevins, 1994).

Structure Calculations

Structures were calculated with the torsion angle dynamics program, DYANA (Güntert et al., 1997). NOEs were converted to distance constraints with the macro CALIBA and angle constraints were derived from intraresidue NOE data and coupling constants with HABAS (Güntert et al., 1991). The calculations included the disulphide bonds previously determined for capsicein: Cys-3 to Cys-71, Cys-27 to Cys-56 and Cys-51 to Cys-95 (Bouaziz, et al 1994). Typically, 100 structures were calculated with the 20 structures showing lowest residual target function being retained for further NOE assignment. The 20 structures from the final calculation were energy minimized with OPAL (Luginbühl et al., 1996) using the default values, except a dielectric constant of 4 and 2500 steps of minimization were used. Structure analysis and all color figures were performed in MOLMOL (Koradi et al., 1996) or Insight (MSI, Inc).

pH titrations

The effect of pH on chemical shift was followed in 2D $^{1}H,^{15}N$ HSQC spectra over the range pH 1.5 to 11.2 in approximately 0.5 pH units. In addition, the tyrosine resonances were followed in 1D and 2D ^{15}N-filtered spectra (Ikura and Bax, 1992) and 2D NOESY spectra over the range pH 2 to 11.2. Appropriate buffers (ionic strength, 45 mM) were used for each pH: for pH 1.45, 2.03 and 2.66 (potassium chloride/ hydrochloric acid), pH 3.36, 3.75, 4.24, 4.76, and 5.20 (d_3-acetate), pH 6.13, 6.80, 7.54 and 7.96 (phosphate), pH 8.52 and 9.03, (boric acid/borax), pH 9.47, 9.87 and 10.44 (carbonate/bicarbonate), pH 10.8 and 11.15 (phosphoric acid/sodium hydroxide).

pKa's were determined by fitting to the Henderson-Hasselbach equation expressed in the form:

$$\delta = [\delta_{acid} + \delta_{base} 10^{(pH-pKa)}]/[1 + 10^{(pH-pKa)}] \quad (1)$$

where δ is the observed chemical shift of a resonance, δ_{acid} and δ_{base} are the chemical shifts at extreme low and high pH, respectively. The program Microsoft Excel (Solver version) was used to perform non-linear least-squares fits of the data to (1). Data from all titration curves of nuclei that sensed the ionizable group and showed a reasonable shift were fitted. Based on

structural analysis, apparent pKa's were assigned to single ionizable groups, and thus the average pKa for each of these titrations is reported.

Theoretical pKa Calculations

Theoretical pKa values were determined using a novel method for treating the pH dependent properties in proteins with multiple interacting titratable groups in the frame of Macroscopic Continuum model (Dimitrov and Crichton, 1997). The method is based on the approximation of electrostatic interactions between the titratable groups with the interaction of each of them with a self-consistent determined molecular field. The field can be different for different titratable groups or for identical groups placed at different spatial coordinates. The use of molecular field approximation gives the possibility to replace the comlex dependence of electrostatic interactions on the space distribution of residue charges with a sum over separated energy terms, each of which depends only on the coordinates of the coresponding ionized residue. Calculations are based on the average charges of titratable groups (Dimitrov and Crichton, 1997), the distance of separation between these groups, their intrinsic pKa's (which define the free energy of charging the ionized groups when they are disconnected from each other and separated at a distance where they cannot interact and where there is no dissolved electrolyte), on residue volumes (Zamyatnin, 1972) and the local dielectric constant (Dimitrov and Crichton, 1997). The charge portions of residue groups are taken as spheres with charges placed at their centers and defined as a center of mass between the heavy polar atoms. Dielectric properties of protein molecules are described in terms of local dielectric constants determined by the space distribution of residue volume density around each ionized residue. The volume used in calculations are obtained by subtraction of the volume of the polar part of side groups as well as the volume of amino and carbonyl groups from the whole volume of residue groups. The residue center of mass is defined between the remaining heavy atoms after deletion of the polar parts. Local density around each residue charge is calculated in the sphere with radius 8Å with the space coordinate of nonpolar parts. The internal dielectric constant of residue groups is 4, the dielectric constant of the solvent medium is 80 and the effective radius of the charge of ionized residues is 2Å.

RESULTS AND DISCUSSION

Characterisation of Aggregation

Line widths in the initial ^{1}H NMR spectra of cryptogein suggested that the protein may have a tendency to aggregate. Pulsed field gradient NMR diffusion experiments (Dingley et al., 1995) showed that 1mM solutions of cryptogein have diffusion coefficients of about 0.9 x 10^{-6} cm^2s^{-1} compared to that of lysozyme, 1.04 x 10^{-6} cm^2s^{-1} and therefore showed an apparent molecular weight of ~20 kDa.

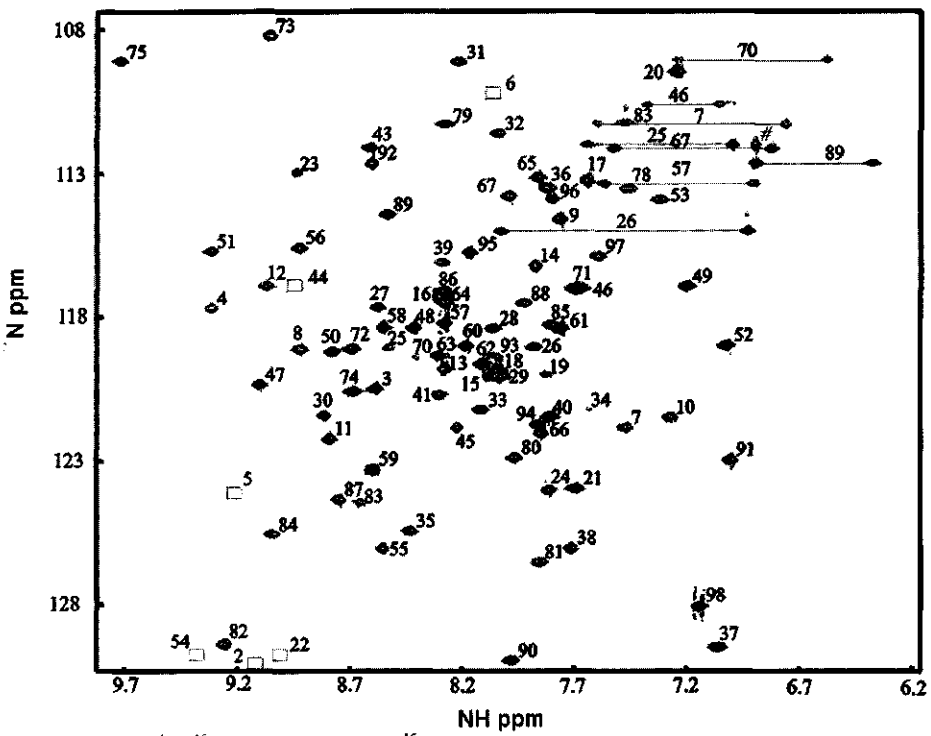


Figure 1. 2D ^{1}H, ^{15}N HSQC spectrum of ^{15}N labelled cryptogein at pH 6.8 and 40°C. All resonances are assigned except for the peak (#) near 6.9 ppm (^{1}H) and 112 ppm (^{15}N) which would belong to either the side chain amide resonances of either Gln-8 or Asn-93. These are the only ^{15}NH resonances not assigned. The correlations of Thr-37, Leu-73, Val-75 and Gly-90 are folded in this spectrum. Boxes indicate the chemical shift positions of the ^{15}NH resonances of Ala-2, Ala-5, Thr-6, Ala-22, Thr-44 and Thr-54, which are observed at pH 4, but due to exchange with water are not observed at pH 6.8.

Acquiring spectra, either at pH values of 4, 5 or 6.8, at low or high ionic strength (0, 200 and 400 mM NaCl in 10 mM phosphate, pH 6.8) and in the presence of 5, 10, 15 or 20 mM

CHAPS (pH 6.8, 45 mM phosphate) did not substantially change the diffusion coefficient. Ultracentrifugation experiments on 20 μM samples at pH 5 or 7.5 showed that the protein had apparent molecular weights of 14 kDa or 11 kDa, respectively, indicating that dimerization is both pH and concentration dependent. We concluded that under most solution conditions and at 1 mM, cryptogein is predominantly a dimer.

Collection of $^{1}H,^{15}N$ HSQC spectra at pH 4 or 6.8, and the eventual complete assignment, showed a single set of resonances was present (Figure 1), showing that if the predominant form is a dimer then it must be symmetric. In the analysis of ^{13}C-filtered spectra (Figure 2) of samples of mixed label (1:1 of $^{13}C/^{15}N$ and $^{12}C/^{14}N$ labelled sample) a number of intermolecular NOEs were assigned, proving that the protein dimerizes.

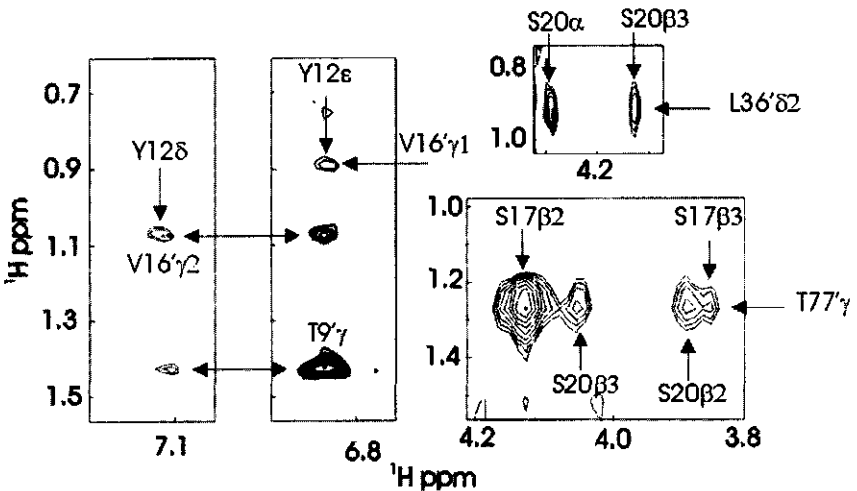


Figure 2. Sections of a 3D ^{13}C-edited, ^{12}C-filtered NOESY spectrum that show key intermolecular NOEs that define the dimer interface. The spectrum was acquired on a sample of 1 mM cryptogein in the ratio 1:1 of $^{13}C/^{15}N$:$^{12}C/^{14}N$ dissolved in 45 mM phosphate, pH 6.8, 100% $^{2}H_2O$, at 45°C.

Structure calculation of the dimer of cryptogein

The sequence specific assignments, intramolecular and intermolecular NOEs of cryptogein were determined using a standard combination of three-dimensional spectra acquired on three samples: a ^{15}N labelled and a $^{13}C/^{15}N$ labelled sample in 90% H_2O/10% $^{2}H_2O$, and a mixed sample of labelled and unlabelled protein. The complete assignment of all observable resonances is summarized in the 2D $^{1}H,^{15}N$ HSQC spectrum in Figure 1.

Table 1. Structural statistics for Cryptogein Before and After Energy Minimization

Residual target function (Å^2)	3.22 ± 0.14	n.a	
Total Energy (kcal/mol)	1141 ± 65	-1013 ± 254	
Van der Waal's (kcal/mol)	713 ± 48	-769 ± 102	
Distance violations [a]			
	RMSD	0.031 ± 0.002	0.030 ± 0.002
	Sum (Å)	20.5 ± 0.6	23.3 ± 1.8
	Maximum (Å)	0.32 ± 0.01	0.10 ± 0.00
Angle violation [b]			
	RMSD	1.19 ± 0.58	0.44 ± 0.02
	Sum (°)	46.9 ± 6.0	66.2 ± 7.2
	Maximum (°)	4.2 ± 0.7	2.6 ± 0.4
Rmsd deviations from ideality			
	Bond angles (°)	n.a	1.660 ± 0.101
	Bond length (Å)	n.a	0.0054 ± 0.0004
Rmsd of atom coordinates of dimer [c]			
	All heavy backbone atoms		
	(Cα,C',N,O) (Å)	0.80	0.82
	All heavy atoms (Å)	0.99	1.03
PROCHECK [d]			
	Residues in most favoured region		
	of Ramachandran Plot A, B and L (%)	87.9 ± 1.3	89.3 ± 1.5
H-bond energy std dev		1.1 ± 0.0	0.7 ± 0.0
Bad contacts/ 100 residues		8.5 ±1.5	0.0

(a) The final 2134 NOEs are categorised as: 1047 intrasubunit NOEs (105 intraresidue, 321 sequential, 305 short range and 323 long range) and 40 intersubunit NOEs (b) 236 dihedral angles (c) For each monomer the rmsd of the backbone and all heavy atoms are 0.49 and 0.79Å, respectively (d) PROCHECK parameters determined by PROCHECK and PROCHECK-NMR.

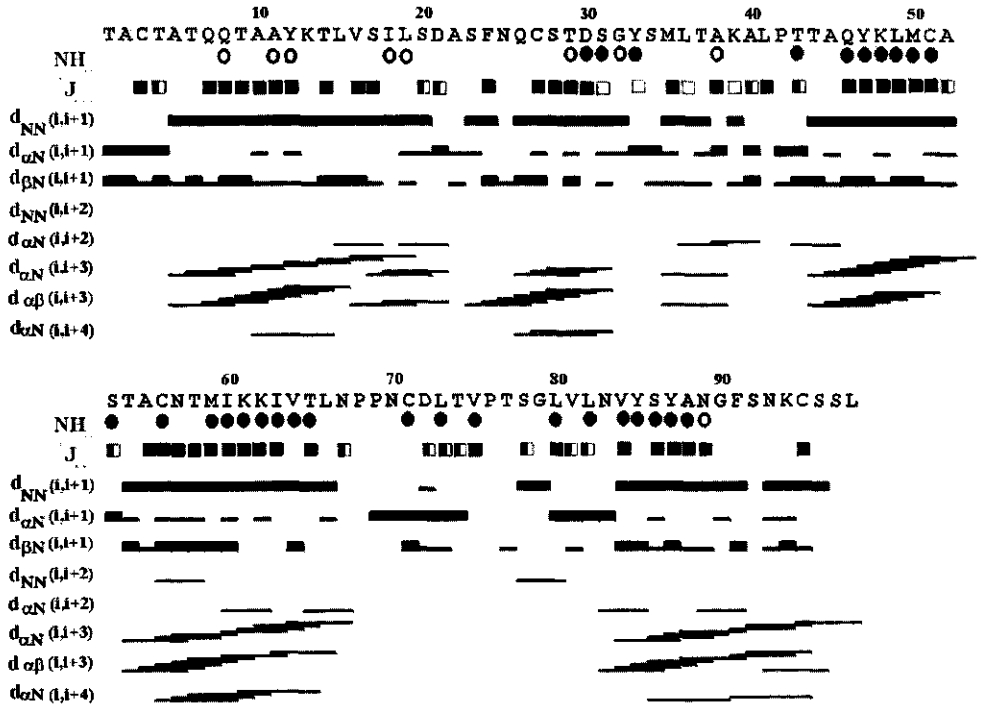


Figure 3. Amino acid sequence of cryptogein and summary of sequential and short range NOE data, $^3J_{HN\alpha}$ coupling constants and NH exchange. The thickness of the bar for the sequential NOE data indicates approximate intensity of the NOE with thick, medium or thin bars for strong, medium and weak NOEs. For short range data presence of the NOE, and not intensity, is indicated. $^3J_{HN\alpha}$ coupling constants determined in 3D HNHA experiments are classified as closed squares (<6 Hz), half-closed squares (6-8 Hz) and open squares (>8 Hz). NH exchange rates determined at pH 6.8, 25°C are qualitatively indicated as closed circles (persisted after 3 hours), open circles (present after 1 to 2 hours, but no longer observed at 3 hours).

The distance constraint data was accumulated from five spectra. The NH_i to NH_j ($j \geq i+1$) NOE data were assigned in 3D $^1H,^{15}N$ HSQC-NOESY and HSQC-NOESY-HSQC spectra. All other interresidue NOEs, including all NH_i to CH_j ($j \geq i+1$) were assigned in a 3D ^{13}C NOESY-HSQC recorded in 90% H_2O/10% ^2H_2O. The sequential assignment data are

summarized in Figure 3. Intraresidue NOEs were assigned in a 3D ^{13}C ROESY-HSQC recorded in 90% H_2O/10% ^2H_2O, but only ROEs that clearly indicated preferred rotamers were included. Intersubunit NOE data were assigned in either a 3D ^{13}C-separated, ^{12}C-filtered NOESY or 2D 13C filtered spectra (Figure 2). After trial structure calculations several NOE violations (Val-16 to Pro-76) assigned to intrasubunit NOEs were clearly intermolecular and were reassigned. These peaks were not observable in the X-filtered spectra, presumably due to the lack of sensitivity of these experiments on mixed label samples. The final 2134 NOEs (Table 1) were complemented with 236 angle constraints per subunit derived from sequential and intraresidue NOEs, 58 $^3J_{HNH\alpha}$ couplings (Figure 3), determined from 3D HNHA spectra, and 29 pairs of $^3J_{HNH\beta}$ couplings from 3D HNHB. The coupling constant data from the HNHA experiments were considered quantitative and were given the limits of ± 1 Hz, whereas the couplings from the HNHB experiment were treated qualitatively and given the limits of ±2 Hz. Stereoassignments of $C\beta H_2$ were determined by considering intraresidue ROE, sequential NOE data and assignments of CβH resonances in the 3D HNHB spectra. A total of 26 $C\beta H_2$ were unambiguously stereoassigned out of 62 spin systems. Similarly, nine pairs of Leu $C\delta H_3$ methyls and four pairs of Val $C\gamma H_3$ methyls were stereoassigned on the basis of intraresidue NOEs and ROEs (the methyl resonances of Leu-82 could not be resolved from each other nor could the methyls of Val-84).

Structures were calculated using the molecular dynamics program DYANA (Güntert et al., 1997). An iterative process of calculation was employed, where successive rounds of structure calculations were used to further the NOE assignments. After the final round of calculation with DYANA the 20 structures with the lowest residual target function were energy minimized with OPAL. Structural statistics are presented before and after minimization in Table 1.

Description of the dimer solution structure of cryptogein and comparison to the X-ray crystal structure.

The solution structure presented here was solved independently of the X-ray crystal structure (Boissy, et al, 1996) and the recently published solution structure (Fefeu et al., 1997). The solution structures of the homodimer show convergence to a single fold (Figure 4A) with the interface formed by the N-terminal helix (α1) and residues at the tip of Ω-loop and β-sheet.

Comparison of the monomers of the solution structure with the crystal structure for the full protein (1 to 98) (Figure 4B) show an rmsd of 0.94 for the backbone atoms and 1.38 for all heavy atoms, and for residues 2 to 97, an rmsd of 0.86 and 1.33 for backbone and heavy atoms, respectively. The global folds of the solution and crystal structure and analysis with PROCHECK (Laskowski et al., 1993; Laskowski et al., 1997) show that despite the solution structure forming a homodimer there are few differences observed between the two structures. Interestingly, Asn-67, which terminates helix-4, shows left-hand helix ϕ,ψ angles in both structures.

One difference to be noted is that the crystal structure has a clear break in the C-terminal helix at residues Asn-89 to Gly-90 resulting in a sixth distinct helix. For the solution structure, the NH_i-NH_{i+1} NOE data (Figure 3) this break is not observed, and the short-range NOE data shows that this helix is continuous to Ser-97.

A loop region that has been classified as an Ω-loop is observed between helices-2 and -3 encompassing Tyr-33 to Pro-42 (Figure 4). The neck of this loop is characterised by NOEs between the side chains of Tyr-33 and Pro-42. Sequential NOEs show that the region Met-35 to Ala-38 is helical-like, and the region encompassing Ala-38 to Leu-41 has alternating $d_{\alpha N}(i,i+1)$ and $d_{NN}(i,i+1)$ NOEs (Figure 3), and several short range NOEs suggesting that the loop consists of several bends. The Ω-loop is positioned over the two-stranded β-sheet to form what has been termed the beak motif (Boissy et al. 1996). As observed in the crystal structure, Tyr-33 which is at the neck of the Ω-loop, is buried in the hydrophobic core. The hydroxyl proton of this residue can be readily observed in homonuclear ^1H spectra between pH 4 and 7. The ring of Tyr-33 is stacked against the ring of Pro-42 in both the solution and crystal structures. The hydrogen bond acceptor of the OH proton of Tyr-33 is the carbonyl of Ala-40 in the crystal structure, which is observed in the family of solution structures. As previously noted (Fefeu et al., 1997) the hydroxyl of Ser-78 is observed in NMR spectra and its acceptor appears to be the carbonyl of Ala-38. These hydrogen bonds may contribute to the stability of the Ω-loop and the overall protein. Recently, the solution structure of P14a, a protein isolated from tomato, has been reported (Fernández et al., 1997).

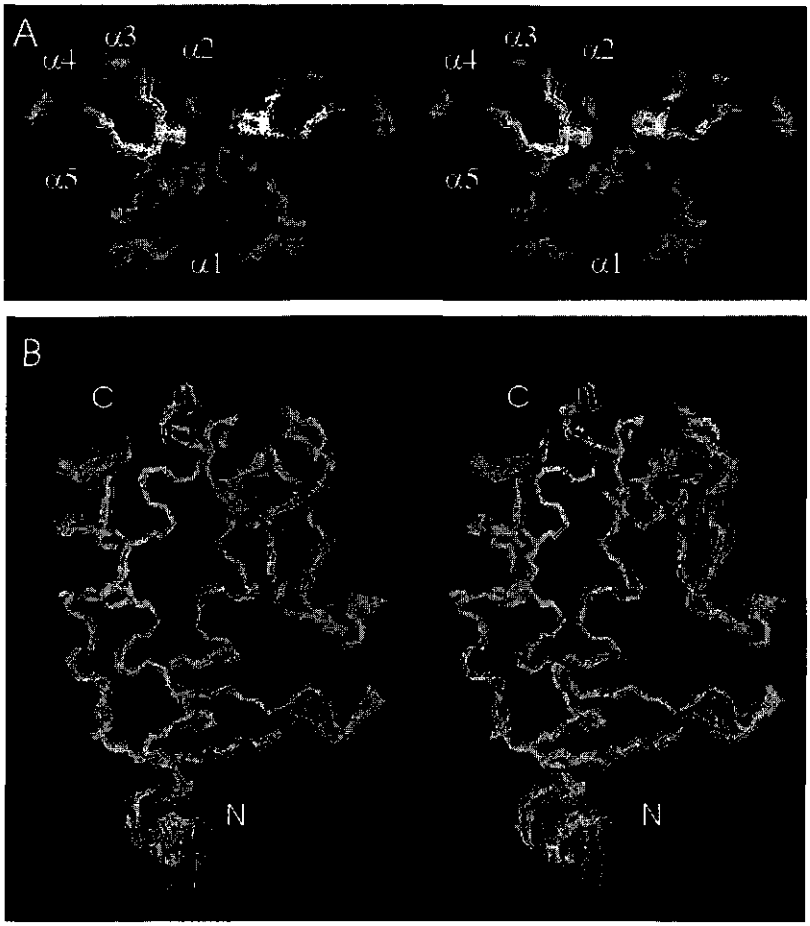


Figure 4. A) Heavy atom backbone trace of the family of 20 final structures superimposed for minimal rmsd for the residues 2 to 97 in each monomer. Helices are shown in red/yellow: α-1, Ala-5 to Ser-20; α-2, Ala-22 to Asp-30; α-3, Thr-44 to Cys-51; α-4, Thr-54 to Thr-65; α-5, Val-84 to Ser-97. The two stranded β-sheet (Asp-72 to Thr-74; Val-81 to Asn-83) is shown in cyan. The Ω-loop that spans Tyr-33 to Pro-42 is shown in yellow. The interface is characterized by helix α1 from one monomer running parallel and at an angle of XX to the same helix of the other monomer. These helices make contacts with residues of the β-sheet and Ω-loop of the opposing monomer. B) superposition of the first monomer of the family of 20 solution structures on the xray crystal structure (1beo) for minimal rmsd (0.86 Å for backbone atoms) for residues 2 to 97.

This protein is induced in response to pathogen invasion and displays antifungal activity. P14a shows a network of internal side chain hydroxyl hydrogen bonds and it is suggested that such bonds would each contribute about 5 kJ mol^{-1} to the free energy of folding. These hydrogen bonds would make a significant contribution to the stability of the protein, which is important for the integrity of the protein in the harsh extracellular environment. While the elicitins do not appear to be structurally related to P14a it is noteworthy that a similar pattern of side chain hydroxyl hydrogen bonds has been identified and these may have similar roles for structural stability. Another tyrosine residue, Tyr-87, is also buried in the protein and near to Tyr-33, with its hydroxyl withdrawn from the surface of the protein, however no hydrogen bond acceptor is observed.

In the family of solution structures, the ring of Tyr-47 appears to be either stacked or edge on to the opposite face of Pro-42 with respect to Tyr-33. In the crystal structure Tyr-47 is stacked against Pro-42. Nevertheless, this stacking and proximity of aromatic rings to Pro-42 results in significant upfield shifts of the Pro resonances and should contribute to the stability of the protein. Another unusual feature of the hydrophobic core of the protein is that the three methionines form a cluster. These residues show a substantial number of NOEs to various residues including those between Met-59 to both Met-35 and –50, which supports the formation of this cluster. The presence of buried tyrosine residues, the side chain hydroxyl hydrogen bonds and the clustered methionine residues may place unusual constraints on the packing and stability of the protein, and consequently the high conservation of these residues.

The N-terminal helix (helix-1) is associated with the edge of the sheet and Ω-loop, whereas the other helices form a globular unit. We have qualitatively used amide exchange rates at pH 6.8 and 27°C to analyse the stability of the overall protein, in particular helix-1, and to determine if Gly-90 breaks the C-terminal helix into two independent helices. Analysis of amide exchange rates (Figure 3) shows that the NH protons of helix-1, and those of the region that encompass helix-6 of the crystal structure (or the Gly-90 to Ser-96 of helix-5 of the solution structure), exchange markedly faster than the NH protons of the other secondary structure elements of cryptogein. Consequently, the N-terminal helix and the region Gly-90 to Ser-96 are less stable compared to the other helices and β-sheet of cryptogein. The Ω-loop does not appear to be stabilized by any main chain hydrogen bonds whereas the NH hydrogens internal to the β-sheet show the expected slow exchange.

Measurement of pKa's

The crystal structure shows two salt bridges: Asp-21 to Lys-62, and Asp-72 to the N-terminus, which may be structurally and/or functionally significant.

Table 2. Experimental and Calculated pKa Values for Cryptogein.

Residue	Exp pKa [a]	Xray[b]	DYANA[c]	OPAL[c]
			Calculated pKa	
N-terminal, Thr-1	7.43	8.05	7.87	7.90
Tyr-12	~11.5	10.47	11.75	11.82
Lys-13		10.93	10.94	10.90
Asp-21	2.49±0.05	2.83	3.31	3.18
Asp-30	2.57±0.12	3.90	3.86	3.88
Tyr-33	>12	11.68	11.27	11.25
Lys-39		10.91	10.94	10.93
Tyr-47	>12	12.38	12.24	12.34
Lys-48		10.88	11.02	11.10
Lys-61	10.1[d]	10.81	10.87	10.87
Lys-62		12.05	11.59	11.67
Asp-72	2.61±0.15	3.25	3.54	3.50
Tyr-85	10.35±0.15	10.40	10.41	10.23
Tyr-87	>12	12.39	12.40	11.92
Lys-94	9.4[d]	10.85	11.09	11.44
C-terminal, Leu-98	3.51±0.01	3.31	3.33	3.20

(a) Experimental pKa values determined at 40°C. Mean and standard deviations are given for pKa's determined by following the pH dependence of a number of chemical shifts: N-terminus (peptide ^{15}N of D72); Asp-21 (peptide ^{15}N of S20, D21, S23, T65 and peptide NH of A22, S23); Asp-30 (peptide ^{15}N of D30, S31, G32, T54, A55); Lys-61 (peptide ^{15}N of I63); Asp-72 (peptide ^{15}N of D72, T74, peptide NH of V84 and side chain $^{15}N\delta^2$ of N70); Tyr-85 (peptide ^{15}N of Y87, N89 and ring Hε of Y85); Lys-94 (peptide ^{15}N of T43); C-terminus (peptide ^{15}N of S96, S97, L98) (b) pKa's calculated for the X-ray crystal structure (1beo) (c) pKa calculated for the 20 lowest energy conformers determined with DYANA and after energy minimization with OPAL using a dielectric constant of 4 (d) pKa could not be assigned unambiguously, for Lys-61 this pKa may be assignable to Lys-62 and for Lys-94 this pKa may be assigned to Lys-48.

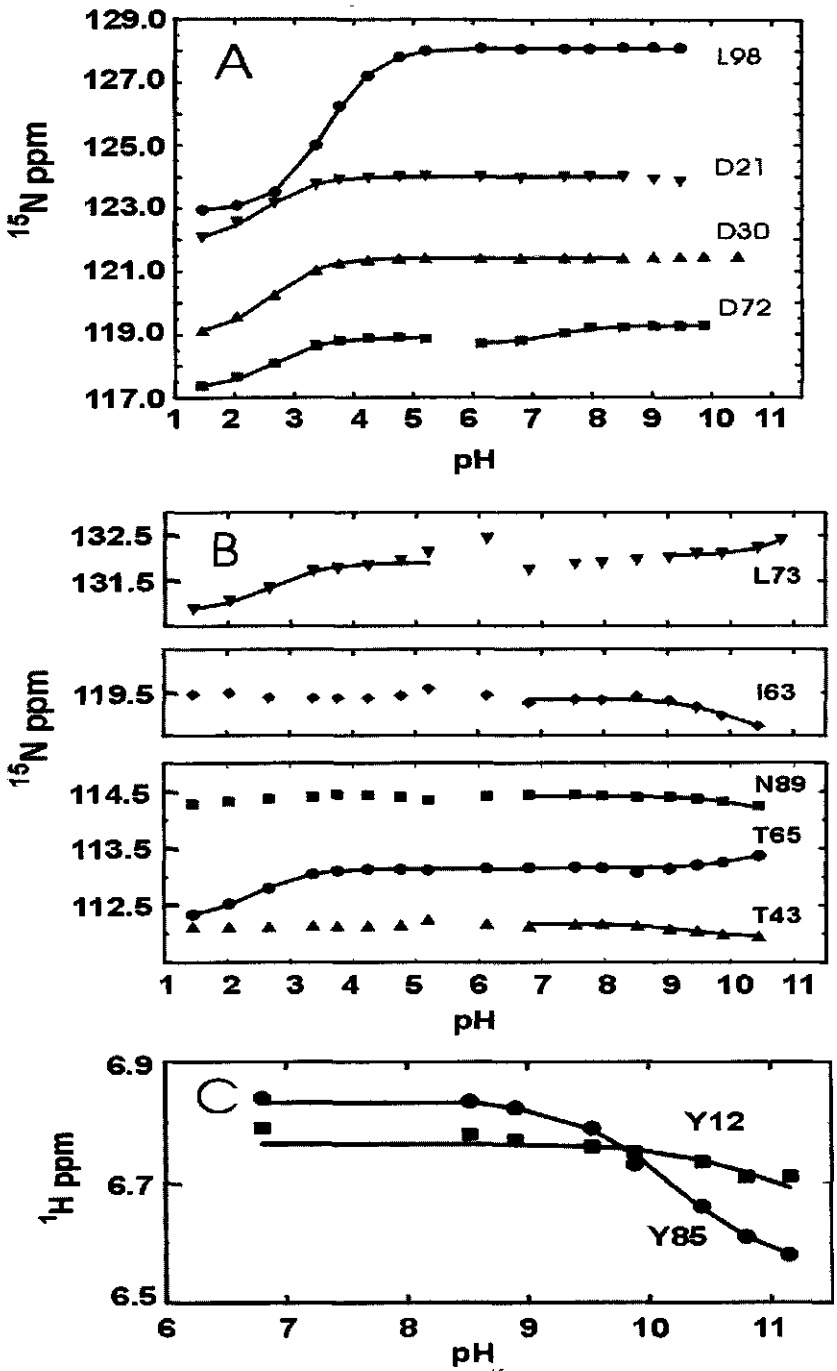


Figure 5. pH dependence of the chemical shifts of (A) the ^{15}N resonances of Asp-21, Asp-30, Asp-72 and Leu-98; (B) the ^{15}N resonances of Thr-43, Ile-63, Thr-65, Leu-73 and Asn-89; and (C) the CεH resonances of Tyr-12 and Tyr-85. Data for (A) and (B) were from 2D ^{1}H,^{15}N HSQC spectra and for (C) from 1D and 2D ^{15}N-filtered NOESY and 2D ^{1}H,^{1}H NOESY spectra. Curves are shown for only the portion of the data fitted to obtain single pKa's.

Visual inspection of the solution structures suggests that the former bridge may be present, but the disorder of the N-terminus makes it difficult to support the presence of the latter (Figure 4). Several NOEs were assigned between the N-terminal region and Asp-72, including the $C\beta H_2$ of Asp-72 to the $C\beta H_3$ of Ala-2, which supports an interaction between the N-terminus and Asp-72. Amongst the elicitins, the N-terminal residue and position 72 are homologously conserved, where the N-terminal is either Ala or Thr and position 72 is either Asp or Glu. Position 21 and 62, however, are not conserved, and although these residues are frequently Asp and Lys, respectively, position 21 may also be Glu, Lys or Thr and position 62 may also be Asn, Glu or Thr. In cryptogein, the salt bridge between Asp-21 and Lys-62 would present a stabilizing interaction between helices-2 and -4. These salt bridges and the ionization states of other residues are important elements of the surface of the protein which may play significant roles in receptor recognition.

The pKa values of the N- and C-termini, Asp-21, Asp-30, Lys-62, Asp-72, Tyr-85 and Lys-94 were determined by collecting a series of 2D $^{1}H,^{15}N$ HSQC spectra at approximately half pH units over the pH range 1.5 to 10.8 at the same ionic strength, and fitting the chemical shifts to the Henderson-Hasselbach equation. Collecting 1D ^{15}N-filtered and 2D ^{15}N-filtered NOESY spectra over the pH range 2 to 11.2, the effect of pH on Tyr-12, -33, -47, -85 and -87 could also be observed. Representative plots of nuclei that show significant shifts ($\Delta\delta > 0.1$ ppm for ^{15}N) with pH are plotted in Figure 5 and all experimental pKa's are described in Table 2. The ionizable residues in cryptogein and those residues affected by pH are shown in Figure 6.

pKa's of C-terminus and Asp residues

The C-terminus shows a typical pKa of 3.5 (Figure 5A) indicating that it does not interact with any residues, which is in agreement with the effects of its ionization being limited to the region Ser-96 to Leu-98. In contrast, all three Asp show low pKa's 2.5 to 2.6 (Figure 5A), suggesting that they interact with positive charges. These data support the presence of the two salt bridges observed in the crystal structure: Asp-21 to Lys-62 and Asp-72 to the N-terminus. Indeed, the titration of Asp-21 is reflected in the chemical shift of the ^{15}NH resonances of Met-59 and Thr-65 (Figure 5B). Asp-30 is not near any positively charged residue. However, this residue which is at the C-terminal end of the short helix-2, is located near the N-terminal end

of helix-4 and its position suggests that it may be aligned with the helix dipole, thus effectively capping that helix. The titration of Asp-30 is clearly observed on the chemical shifts of the ^{15}NH resonances of Thr-54 and Ala-55 which are the first residues of helix-4.

pKa's of N-terminus and Lys residues

The pKa of the N-terminus was determined by following the peptide ^{15}N resonance of Asp-72 (Figure 5A) and is estimated to be 7.3, similar to a pKa expected for a free N-terminus. This pKa is somewhat incompatible with the observed low pKa of the carboxylate of Asp-72, which itself supported the presence of a salt bridge between the carboxylate of Asp-72 and the N-terminus, and perhaps the low pKa of Asp-72 is due to other interactions.

Several other shifts are observed in the alkaline pH range, which may be attributable to lysine residues (Figure 5B). The ^{15}N peptide resonance of Thr-43 shows a small but significant pH dependence ($\Delta\delta$ = 0.12 ppm) with an apparent pKa of 9.4 and as Thr-43 is near Lys-94 (~8Å) we assign the pKa to this residue. However, Lys-48 is also relatively close to Thr-43 (10Å) and may have an effect and thus we can not assign this pKa unambiguously. The ^{15}N peptide resonance of Ile-63 has an apparent pKa of 10.1. As this residue is near both Lys-61 and Lys-62 this pKa also can not be assigned unambiguously. The ^{15}N peptide resonance of Thr-65 shows three pKa inflections, the first pKa is 2.4, and is assigned to Asp-21 whose carboxylate forms a salt bridge with Lys-62. The next two inflections do not give reasonable fits, but are most likely due to the N-terminus and either or both Lys-61 and -62.

pKa's of Tyr residues

The ring resonances of the tyrosine residues are substantially unaffected between pH 2 and 8.5. The CεH resonances of Tyr-85 show a large chemical shift dependence ($\Delta\delta$=0.26 ppm) which fits to a pKa 10.2. The pKa of Tyr-85 appears to be reflected in the peptide ^{15}N resonances of Tyr-87 and Asn-89 with apparent pKa's of 10.5 and 10.4, respectively. The CεH resonances of Tyr-12 show a small pH dependence between pH 7 and 11 ($\Delta\delta$=0.09 ppm) that does not appear to fit to a single ionization. The difficulty with this titration is that it may reflect the ionization of both Tyr-12 and Lys-13 in both the monomeric and dimeric species (Figure 6). Several ^{15}N peptide resonances of residues located in the β-sheet and near Tyr-12

show pH dependence in both the acid and the alkaline regions. The ^{15}N resonance of Leu-73 shows three pH inflections (Figure 5B). The first ($\Delta\delta$=1.09 ppm) is readily fitted to a carboxylate ionization with a pKa of 2.7 and is thus assigned to Asp-72. The second inflection occurs between pH 5 and 8 ($\Delta\delta$=0.24 ppm), and while it was not possible to fit this inflection it must be due to the N-terminus. The third inflection occurs above pH 9 ($\Delta\delta$=0.5 ppm). If this latter shift is fitted, an estimate of an apparent pKa of ~11.5 (with reasonable fits ranging from pKa of 11 to 12) is observed and thus we assign this pKa to Tyr-12. The CεH resonances of Tyr-33 show no shifts at all, while those of Tyr-47 and -87, show small shifts above pH 9.0 ($\Delta\delta$ = 0.02 and 0.03 ppm, respectively), suggesting that the pKa's of these three tyrosines are greater than 12.

The titration behaviour of all the tyrosine residues mostly agrees with the extent of their solvent exposure and their interactions.

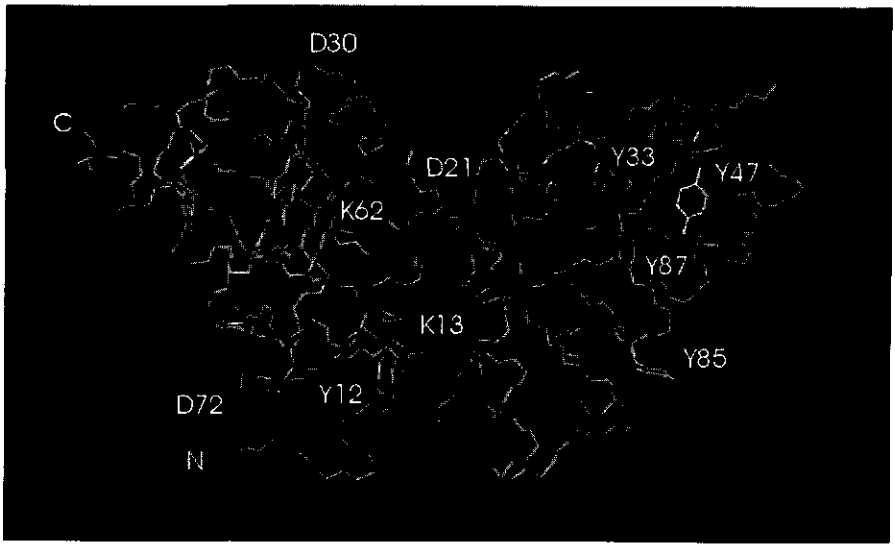


Figure 6. The dimer solution structure highlighting the ionizable residues. Aspartate residues are colored in red, tyrosine in yellow, lysine in green. The N- and C-termini are indicated on one monomer. Tyr-12 and Lys-13 are the only residues at the interface of the dimer. Tyr-12 lies parallel to the protein surface of the monomer and the side chain of Lys-13 is well exposed to the solvent. The pKa's of these residues do not appear affected by dimerization.

Tyr-85 is completely exposed to the solvent giving a typical pKa, and Tyr-12 is mostly exposed, but lies flat against the protein in both the solution and crystal state. Tyr-33 and -87 are all partially or completely buried. Tyr-33 is buried with its OH hydrogen bonded. Tyr-87 is

also predominantly buried with its hydroxyl withdrawn from the solvent, however, we have not detected this hydroxyl hydrogen in any of our spectra. The hydroxyl of Tyr-47 is exposed to the solvent, but is stacked against Pro-42 and near to Phe-91, and therefore is in a hydrophobic environment. Further this residue is near Tyr-87 and thus the deprotonation of these two residues may affect each other. All experimentally determined pKa's of the tyrosines, except the pKa of Tyr-87, are in reasonable agreement with those determined by UV difference spectroscopy (Nespoulous et al., 1994). Tyr-87 was found to have a typical pKa by this latter method.

Calculations of the pKa's based on the crystal and solution structures.

The theoretical pKa's were determined by using the crystal structure (Boissy et al., 1996), the unrefined and refined solution structures that are described above. In Table 2 the calculated pKa's of the individual ionizable groups are reported along with the experimental pKa's. In most cases, the experimental and calculated values agree well, and the data for the crystal structure is similar to both the refined and unrefined solution structures. The five tyrosine residues have values that in good agreement and consistent with their interactions and/or their exposure to the solvent as discussed above. Tyr-12 is on the protein surface, but lies flat against its own monomer. This residue is at the interface of the dimer, and is near the Lys of the opposing dimer (Figure 6). However, the pKa of Lys-13 is calculated to have a typical value and thus we do not expect an interaction between these residues. The calculated pKa of Tyr-12 is similar for both the solution and crystal structures and therefore the effect by dimerization appears minimal.

The calculated pKa for Lys-94 is markedly lower to the experimental pKa of 9.4, however, the assignment of this pKa is ambiguous (Table 2). Recent calculations of the pKa's of Lys-94 and Tyr-47 suggest that these residues interact to raise the pKa of Lys-94 to 13 and lower the pKa of Tyr-47 to 9 (Vogel & Huffer, 1998). This interaction is not apparent in our calculations nor in the experimental data for Tyr-47. The remaining Lys residues have typical pKa's (~11) except for Lys-62 which is expected to participate in a salt bridge with Asp-21. The experimental and/or theoretical pKa's for Asp-21 and Lys-62 agree with the presence of the salt bridge (Table 2). The salt bridge between Asp-72 and the N-terminus is less apparent with both a normal experimental and theoretical pKa for the N-terminus. The low experimental pKa

for Asp-72 may reflect other interactions that remain to be determined. These conclusions are supported by the calculations of the pKa's of Asp-21 and -72 reported by others (Vogel and Juffer, 1998) which are in better agreement with our experimental data. The theoretical pKa of Asp-30 is typical for an Asp residue in small peptides, which is in contrast to the experimental pKa. The method (Dimitrov and Crichton, 1997) that we have used here does not account for aligned partial charges that would occur in helix dipoles. Calculations by others using a different approach (Juffer et al., 1997; Vogel and Juffer, 1998) also show that Asp-30 has a typical pKa, and these methods do include partial charges and should therefore consider interactions with helix dipoles. While it is clear from the experimental data that this residue is not interacting with other ionizable residues, we can not conclude on the actual nature or importance of the interactions that lower the pKa of Asp-30.

Functionally Important Residues

The >70% identity of the sequences of the elicitins, and the absence of a characterized receptor reduces the conclusions that can be made with respect to which are functionally important residues. One would expect that the evolutionary pressure to conserve residues amongst the elicitins would be due to the unknown function of the elicitins within the genera *Phytophthora* and *Pythium*, and not the advantageous induction of defence responses displayed within the resistant plant. In this report, a number of potentially important structural/dynamic features have been raised: firstly, the amide exchange rates of helix-1, that include the functionally significant Lys-13, highlights the relative instability of the first helix compared to the other helices and the β-sheet; secondly, the three Asp residues show low pKa's and may be involved in both salt bridges and other interactions that stabilize protein structure; thirdly, four of the five Tyr residues show shifted pKa's and withdrawn completely or to some extent from the solvent; finally the protein shows a tendency to dimerize at concentrations above 20 μM. While it is tempting to suggest that dimerization may be important for inducing a defence response, this activity is observed at concentrations < 1 nM, where dimerization is less likely. Interestingly, the regions that appear to interact, ~20s, ~30s and the ~70s, are on the same side of the protein as is the functionally important residue 13. Whether these residues are important in protein-protein interactions can only be proved through site-directed mutagenesis studies and the isolation of the receptor.

Acknowledgments

This work benefited from the use of NMR facilities at the University of Melbourne and the Australian National University and was funded by ARC grants no. S09711445 and no. A09801407, and a joint Institute of Advanced Studies/ Australian Universities Collaborative Research Grant. We thank Ms Sieu Cleland for preparing cultures, Dr Geoff Howlett for the ultracentrifugation data and Prof Hans Vogel and Dr André Juffer for providing their manuscript prior to publication.

REFERENCES:

Archer, S.J., Ikura, M., Torchia, D.A. and Bax., A. (1991) *J. Magn. Reson.*, **95**, 636-641.

Boissy, G., de la Fortelle, E., Kahn, R., Huet, J.-C., Bricogne, G., Pernollet, J.-C. and Brunie, S. (1996) *Structure*, **4**, 1429-1439.

Bonnet, P., Bourdon, E., Ponchet, M. Blein, J.P. and Ricci, P. (1996) *Eur. J. Plant Pathol.* **102**, 181-192.

Bouaziz, S., Van Heijenoort, C., Guittet, E., Huet, J.-C., and Pernollet, J.-C. (1994) *Eur. J. Biochem.*, **33**, 8188-8197.

De Wit, P. J. G. M. (1992) *Annu. Rev. Phytopathol.*, **30**, 391-418.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biolmol. NMR*, **220**, 427-438.

Dimitrov, R. A. and Chrichton, R. R. (1997) *Proteins*, **27**, 576-596.

Dingley, A. J., MacKay, J. P., Chapman, B. E., Morris, M. B., Kuchel, P. W., Hambly, B. D. and King, G. F. (1995) *J. Biomol. NMR*, **6**, 321-328.

Fefeu, S., Bouaziz, S., Huet, J.-C., Pernollet, J.-C. and Guittet, E. (1997) *Protein Science*, **6**, 2279-2284.

Fenn, M. E. and Coffey, M. D. (1984) *Phytopathology* **74**, 606-611

Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mösinger, E. and Wüthrich, K. (1997) *J. Mol. Biol.*, **266**, 576-593.

Folmer, R.H.A., Hilbers, C.W., Konings, R.N.H. & Hallenga, K. (1995) J. Biomol. NMR **5**, 427-432.

Forman-Kay, J. D., Clore, G. M. and Gronenborn, A. M. (1992) *Biochemistry*, **31**, 3442-3452.

Frenkiel, T., Bauer, C., Carr, M. D., Birdsall, B. and Feeney, J. (1990) *J. Magn. Reson.*, **90**, 420-425.

Gayler, K.R., Popa, K.M., Maksel, D.M. Ebert, D.L. and Grant, B.R. (1997) *Mol. Plant Pathol. On-Line* http://www.bspp.org.uk/mppol/1997/0623gayler.

Grant, B.R., Ebert, D. and Gayler, K.R. (1996) *Aust. Plant Pathol.* **25**, 148-157.

Güntert, P., Braun, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 517-530.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283-298.

Huet, J.C., La Cear, J.P., Nespoulous, C. and Pernollet, J.C. (1995) *Mol. Plant-Microbe Interactions* **8**, 302-310.

Ikura, M. and Bax, A. (1992) *J. Am.Chem. Soc.*, **114**, 2433-2440.

Johnson, B. A. and Blevins, R. A. (1994) *J. Biomol. NMR*, **4**, 603-614.

Juffer, A. H., Argos, P. and Vogel, H. J. (1997) *J. Phys. Chem. B*, **101**, 7664-

Jones, D. A., Thomas, C. M., Hommon D'Kosack, K. E., Balint-Kurti, P. J. and Jones, J. D. G. (1994) *Science*, **266**, 789-793.

Kamoun, S., Young, M., Glascock, C.B. and Tyler, B.M. (1993) *Mol. Plant-Microbe Interactions* **6**, 15-25

Kay, L. E., Xu, G.-Y., Singer, A. U., Muhandiram, D. R. and Forman-Kay, J. D. (1993) *J. Magn. Reson.*, **101**, 333-337.

Keizer, D. W., Schuster, B., Grant, B. R. and Gayler, K. R. (1998) *Planta*, in press.

Kooman-Gersmann, M., Honée, G., Bonnema, G. and De Wit, P. J. G. M. (1996) *Plant Cell*, **8**, 929-938.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph*, **14**, 51-55.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) *J. Appl. Crystallogr.*, **26**, 283-291.

Laskowski, R. A., Rullman, J. A. C., MacArthur, M. W., Kaptein, R. and Thornton, J. M. (1996) *J. Biomol. NMR*, **8**, 477-486.

Luginbühl, P., Güntert, P., Billeter, M. and Wüthrich, K. (1996) *J. Biolmol. NMR*, **8**, 136-146.

Majumdar, A. and Zuiderweg, E. R. P. (1993) *J. Magn. Reson.*, **102**, 242-244.

Myers, R.A., Cruz, L.J., Rivier, J.E. and Olivera, B. (1993) *Chemical Reviews* **93**, 1923-1936

Muhandiran, D. R. and Kay, L. E. (1994) *J. Magn. Reson.*, **103**, 203-216.

Nakamura, H. (1996) *Q. Rev. in Biophys.*, **29**, 1-90.

Nespoulous, C. and Pernollet, J.-C. (1994) *Int. J. Pept. Protein. Res.*, **43**, 154-159.

O'Donohue, M. J., Gousseau, H., Huet, J.-C., Tepfer, D. and Pernollet, J.-C. (1995) *Plant Mol. Biol.*, **27**, 577-586.

Vogel, H. J. and Juffer, A. H. (1998) *Theor. Chem. Acc.* (in press).

Vuister, G. W. and Bax A. (1993) *J. Am. Chem. Soc.*, **115**, 7772-7777.

Vuister, G. W., Kim, S.-J., Wu, C. and Bax A. (1994) *J. Am. Chem. Soc.*, **116**, 9206-9210.

Weiner, , S. J., Kollman, P. A., Nguyen, D. T. and Case, D. A. (1986) *J. Comput. Chem.*, **7**, 230-252.

Wendehenne, D., Binet, M.N., Blein, J.P., Ricci, P. and Pugin, A. (1995) *FEBS Letters*, **374**, 203-207

Zamyatnin, A. A. (1972) *Prog. Biophys. Mol.*, **24**, 803-813.

Zhang, O., Kay, L. E., Olivier, J. P. and Forman-Kay, J. D. (1994) *J. Biomol. NMR*, **4**, 845-858.

5 TOPOLOGICAL REQUIREMENT FOR THE NUCLEUS FORMATION OF A TWO-STATE FOLDING REACTION. IMPLICATIONS FOR Φ-VALUES CALCULATIONS.

Roumen A. Dimitrov, Colja Laane, Jacques Vervoort and Robert R. Crichton

Submitted to: *Protein Science*

SUMMARY

The folding problem of two-state small monomeric proteins is reduced to the question of how the folding nucleus at the transition state (TS) is formed from the ensemble of rapidly interconverting partly structured conformations in the denatured state. It is shown that in the denatured state the folding is energetically favored by certain highly fluctuating nucleation regions (α-helices and(or) β-hairpins), which in the experiments based on site directed mutagenesis are revealed by their high Φ-values. In the TS the folding is favored by the packing of these nucleation regions together with some other portions of the polypeptide chain, thus leading to a broad distribution of the Φ-values. The packing process results in a nucleus with native like-topology, approximately correctly formed secondary structures and loop regions with different degrees of order. This native-like nucleus is separated from all other folding alternatives by a high free energy barrier. The calculations of the free energy of the folding nucleus are based on: 1) statistical mechanics of a linear cooperative system; 2) a self-consistent molecular mean field theory previously developed for electrostatic interactions (Dimitrov, R. A., & Crichton, R. R. (1997). Self-Consistent Field Approach to Protein Structure and Stability. I. pH Dependence of Electrostatic Contribution. *Proteins Struct. Funct. Genet.* **27**, 576-596), and 3) a lattice model based on packing of idealized α - and(or) β-secondary structures. The model was tested on a set of proteins for which the Φ-value analysis of the TS is experimentally well studied: barley chymotrypsin inhibitor 2 (CI2); association of two fragments of barley chymotrypsin inhibitor -[CI2-(20-59) and CI2-(60-83)] and association of the two domains of the Arc repressor of phage P22. Finally, the model successfully predicted the experimental Φ-values of the activation domain of human procarboxypeptidase (Ada2H) (unpublished data of L. Serrano EMBL, Heidelberg). From the

calculated Φ-values it follows that the key residues for the folding of Ada2H domain are ILE23 from the amino end of the second α-helix and ILE15 from the fourth β-strand. Together with LEU26, from the amino end of the second α-helix, and ALA52 and VAL54, from the first β-strand, they constitute the proposed nucleation site.

INTRODUCTION

Experimental analysis of the folding reaction of proteins larger than 100 residues have shown that these proteins normally display kinetic intermediates (Villegas et al., 1995). Kinetic intermediates are believed to reduce significantly the conformational freedom of the polypeptide chain and thus to speed up the search for productive interactions. On the other hand, detailed kinetic and equilibrium experimental studies on small monomeric proteins with less than 80 residues (and no disulfide bridges)- α-spectrin SH3 domain (Viguera et al., 1996), IgG-binding domain of protein G (Alexander et al., 1992), cold-shock protein CspB (Schindler et al., 1995), acyl-coenzyme A binding protein (Kragelund et al., 1995), barley chymotrypsin inhibitor 2 (CI2) (Jackson & Fersht, 1991), N-terminal domain of λ_{6-85} repressor (Burton et al., 1997) and the activation domain of human procarboxypeptidase Ada2H (Villegas et al., 1995)- have shown that these proteins fold without accumulation of kinetic intermediates.

It seems that the folding pathway of small monomeric proteins is determined by the sequence ability to stabilize only productive transition states and not by conformational restrictions in the denatured state. Thus, a thorough understanding of the structural reorganization which takes place in the TS for folding is needed. The principal approach is to convert the experimental measurements of the rate and equilibrium constants along the folding pathway to free energy profiles for the folding pathways. A Φ-value analysis has been introduced (Fersht, 1997) as a measure of the perturbation caused by mutants on the free energy profile of the wild-type protein. Φ-value analysis has shown that the TS for small monomeric proteins, which fold and unfold as a single cooperative unit, is compact but relatively uniformly unstructured, with tertiary and secondary structure being formed in parallel. The observed Φ-values range from 0 (site unfolded in the TS) to 1 (site fully folded in the TS), being in general higher in the hydrophobic core than on the surface of the protein. The distribution of the Φ-values is broad and only a few of the residues have Φ-values close to 1.

Recently, characterization of the refolding properties of a number of simple monomeric proteins has shown that the distribution of the Φ-values is strongly dependent on the topology of the final folded form of the protein. The influence of other factors, such as equilibrium stability of the native state and the chain length, are not apparent (Plaxco et al., 1998). Therefore, the important question is to understand how the topological restrictions on the folding pathway result in the experimentally observed broad distribution of the Φ-values. None of the current theoretical models addresses this subject properly. Thus, in the nucleation-condensation mechanism (Fersht, 1997) the TS is seen as an uniformly expanded form of the native state with a large diffusive nucleus. It is composed of both neighboring residues in local secondary structure and long-range tertiary interactions. However, topological properties of the TS are not specified. In the funnel picture (Bryngelson, 1995), the TS is structurally degenerated. It is represented by a diffuse ensemble of states, distributed over the top of a broad free energy barrier. The nucleus is delocalized over the sequence and tertiary contacts. Finally, in the nucleation growth mechanism (Abkevich, 1994; Ptitsyn, 1994) the topological dependence applies to the accumulation of a kinetic intermediate with a native-like overall fold.

In this study we present a statistical thermodynamic theory which properly addresses the topological restrictions on the structural properties of the TS ensemble of conformations. Our approach is based on the existence of a high free energy gap at the TS level. The role of this gap is to increase the population of conformational states with productive interactions along the folding pathway. Thus, at its lowest free energy state, the TS is dominated by conformations with native-like topology, approximately correctly formed secondary structures and flexible loops. The population of all other folding alternatives, which include both changes in topology and secondary structures, are strongly reduced. A lattice model is introduced which takes into account the mutual packing of the secondary structures. Conformational states of the polypeptide chain are described by the fluctuations of the lengths and location of the secondary structures along the sequence and in the lattice. A theoretical approach, based on the statistical mechanics of a linear cooperative system and a self-consistent molecular field theory, is developed for the calculation of the free energy of the TS. Φ-Values are determined from the expression $\Phi = \frac{\Delta\Delta F_{TS}}{\Delta\Delta F_{N-D}}$, where $\Delta\Delta F_{TS}$ and $\Delta\Delta F_{N-D}$ are the perturbations of the free energy of the TS and that of the unfolding free energy upon mutation.

The theoretical model was tested on a set of small proteins for which the experimental Φ-values are well determined- barley chymotrypsin inhibitor 2 (CI2) (Itzhaki et al.,1995); association of two fragments of barley chymotrypsin inhibitor -[CI2-(20-59) and CI2-(60-83)] (de Prat Gay et al., 1994), association of the two domains of the Arc repressor of phage P22 (Milla et al., 1994; Milla et al., 1995) and the activation domain of human procarboxypeptidase (Ada2H) (unpublished data of L. Serrano EMBL, Heidelberg). The calculated Φ-values are in good quantitative agreement with the experimental Φ-values.

THEORY

Lattice model

The lattice model is based on packing of α- and(or) β-secondary structures. Steric restrictions between the secondary structures are taken into account by averaging over the backbone coordinates of known 3D-protein structures. The volumes of amino acid residue side groups are usually much smaller than the volumes of α- and β-secondary structures and therefore cannot seriously influence the crude packing of these secondary structures. Therefore, we use a simplified model for the secondary structures. They are represented by hypothetical cylindrical surfaces on which C_α-atoms of the polypeptide backbone form a right handed spiral. The side groups of amino acid residues are treated as spheres, the center of which represents the average displacement of the side groups from the C_α-atoms. The lines which connect the C_α-atoms and the center of the side groups are perpendicular to the axes of the spirals, fig.1. Lattice conformations are described by the conformational freedom of the loops and by the fluctuations of the lengths and locations of the regular α- and (or) β-regions along the sequence and in the lattice.

Free energy of protein conformations at TS

The functional form of the Hamiltonian of the protein conformations and its lattice representation are determined, based on the following experimental observations: 1) Proteins in the TS are expanded relative to the native state by approximately 10% to 15% (Fersht, 1997).

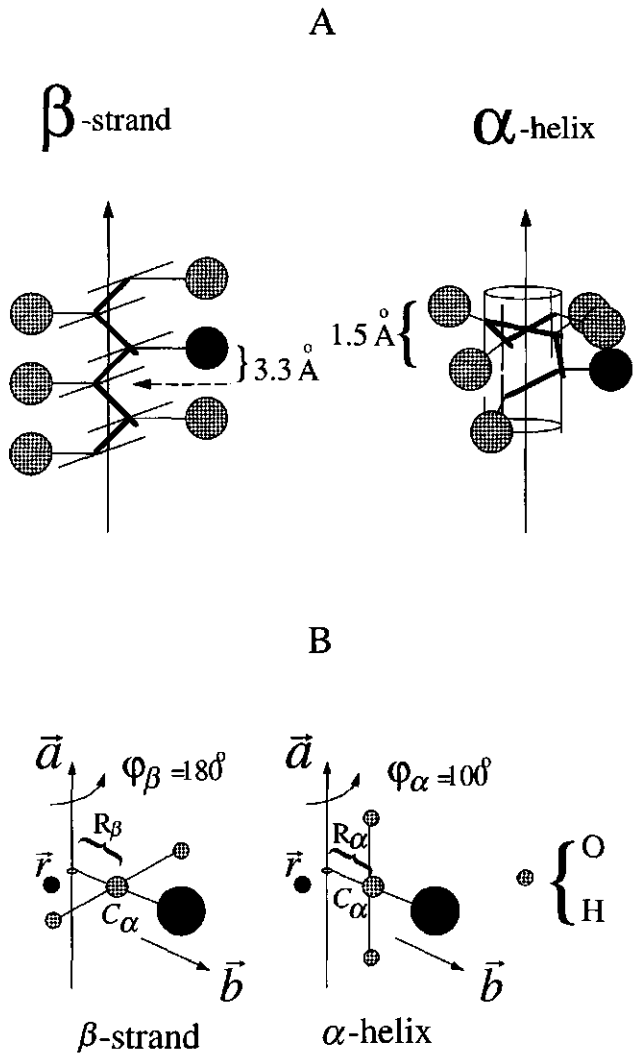


Figure.1. Geometrical representation of the idealized α - and β - secondary structures. β-strands and α - helices are represented by the right-handed spirals. The main characteristics of these idealized structures are: distance between the C_α-atoms along the helix axis- d ; helix radius- R ; rotation angle between two closest C_α-atoms along the helix axis-φ. For β-structures we have: d =3.3Å, R =0.25Å, φ=189.15°. For α - structures we have: d =1.5Å, R =2Å, φ=100°. The direction of the strand is marked by $\vec{a}$ and $\vec{b}$ is the direction of the side group placed at the central position $\vec{r}$ (●) of the strand. The effective directions of $C_\alpha \rightarrow H$ and $C_\alpha \rightarrow O$ hydrogen bonds are also shown. For α -structure effective hydrogen bonds are directed toward the helix axis, but it is important to note that they are not involved in intramolecular long-range interactions. Both for α - and β-structures the length of $C_\alpha \rightarrow O$ and $C_\alpha \rightarrow H$ is on average 2Å. The average displacement of the side groups from C_α-atoms is 3Å.

As a consequence the van der Waals interactions are greatly weakened (>50% relative to the native protein). 2) Electrostatic interactions on the surface of the protein are also weakened (Oliveberg & Fersht, 1996). 3) The enthalpy of the TS is higher than that of both the native and denatured states, which means that the gain in enthalpy resulting from the weakening or breaking of internal interactions (especially in the loop regions) in the native state is not completely compensated by the hydration effect in the TS (Segawa & Sugihara, 1984). The latter has been shown to amount for more than 80% of the total heat capacity change in going from the TS to the unfolded state (Privalov, 1979). Thus in the TS the interior of the protein is partly intact and highly hydrophobic (unsolvated) but its surface is flexible especially in the loop regions. 4) Secondary structures are not fully formed: α-helices are mainly flexible in their N- and(or) C-terminus, β-sheets are flexible in the ends of the β-strands and at their edges (Itzhaki et al., 1995; Viguera & Serrano, 1997). 5) Also as follows from kinetic experiments (Itzhaki et al., 1995; Serrano et al., 1992; López-Hernández & Serrano, 1995; Viguera et al., 1996; Milla et al., 1995), the close packing of the side chain groups in the hydrophobic core is never completely broken before the TS for unfolding or completely formed before the TS for refolding. This means that at least some partial packing constraints exist at the TS which lead to an increased rigidity of the secondary structures around their buried hydrophobic groups.

For the denatured state D the experimental results show that the sum of the differences in heat capacities ΔC_p in going from the denatured state (D) to the TS and from the native state (N) to the TS under strongly refolding and unfolding conditions agree well with the difference in heat capacity in going from D to N in equilibrium conditions, $\Delta C_{p,TS-D} - \Delta C_{p,N-TS} \approx \Delta C_{p,N-D}$ (Tan et al., 1996). This suggests that under strongly refolding conditions the D state is approximately as hydrated as the denatured state under strongly denaturing conditions. Therefore, as a first approximation the free energy of protein conformations, which populate the TS, is calculated with reference to a ground state with completely unfolded conformations. Thus, interactions between residues which are far apart in the sequence, but close in space are neglected and amino acid residues which are close in sequence are considered as energetically uncoupled.

From the above, it follows that the main contributions to the stability of the protein conformations in the TS are determined by: the free energies of the individual residues forming during the coil-α-helix and coil-β-structure transitions, free energy of residues transferring from polar to nonpolar medium, as well as the free energy of α-helix initiation, free energy of

bend formation between α-α, α-β or β-β structural regions and their non specific interactions with the rest of the polypeptide chain and the excess free energy of the β-sheet marginal chains.

Because of the increased rigidity around the hydrophobic core, the most probable fluctuations of the secondary structures, as indicated by the Φ-values analysis (Itzhaki et al., 1995; Viguera & Serrano, 1997), is at their N- and C- terminus. Taking into account that in the TS the terminus of the secondary structures are connected by relatively long and flexible loops it follows that as a first approximation it is reasonable to neglect the restrictions from the loop connectivity and to use the mean field approximation-considering the N- and C- terminus of the secondary structures as independent. Therefore, it is convenient to represent interactions between the secondary structures in terms of interactions between their N- and C- terminus. For this, each secondary structure is separated in two parts, one of which is associated with its N-terminus and the other with its C-terminus, fig.2.

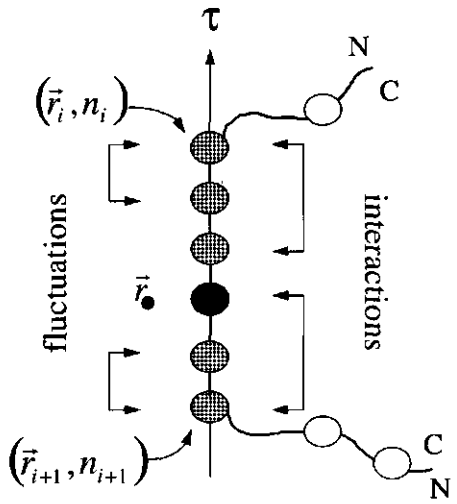


Figure.2. A schematic representation of the allowed positions for the N- and C- terminus of the secondary structures. ◍-Marks the position (τ) of the C_α-atoms along the axis of a given secondary structure, whereas ○-marks the position (n) of a given residue in the sequence. The space coordinate ($\vec{r}_0$) of ○ is variable, whereas the space coordinate ($\vec{r}_\bullet$) of ◍ is fixed. Thus, $(\vec{r}_i, n_i)$ and $(\vec{r}_{i+1}, n_{i+1})$ are the coordinates of the N- and C-terminal residues, whose C_α-atoms correspond to the entering or exiting of the polypeptide chain relative to the axis of the secondary structure. The boundaries marked at the left side of fig.2 represent the domains of fluctuation. The right side represents the separation of secondary structures in two parts, which are ascribed to N- or C- terminus. These regions participate in energy interactions between the terminus of secondary structures.

Thus the free energy can be presented in the form:

$$E(Q)=\sum_i E_i(\vec{r}_i,n_i)+\sum_i E_{i,i+1}(\vec{r}_i,n_i;\vec{r}_{i+1},n_{i+1})+\frac{1}{2}\times\sum_i\sum_{\substack{j\\ i\neq j}} E_{i,j}(\vec{r}_i,n_i;\vec{r}_j,n_j) \qquad (1)$$

where $Q\left(\{\vec{r}_i\},\{n_i\}\right)$ determines the locations of the secondary structures along the sequence and in the lattice as well as the topological connections between their N- and C-terminus. E_i is the internal energy associated with each terminus (for example the sum of free energies of elongation for α- or β-secondary structures for the individual amino acid residues which contribute to the given termini). $E_{i,i+1}$-the energy of interaction between nearest neighbors termini along the sequence (for example the bending energy of loop connections between the N- and C- terminus of adjacent secondary structures). $E_{i,j}$-non-local energy interactions between residues which are far apart in the sequence but close in space- for example hydrophobic and hydrogen bonding interactions.

Strictly $E(Q)$ is the free energy of the solution when the protein conformation is fixed which means that one has to average over the solvent degrees of freedom. Therefore, in general E_i, $E_{i,i+1}$ and $E_{i,j}$ depends on such parameters as temperature, salt concentration, dielectric constant and so on.

Statistical mechanics of the protein conformations in an external field

In the presence of an external field φ the last term in (1) is represented in the form:

$$\frac{1}{2}\times\sum_i\sum_{\substack{j\\ i\neq j}} E_{i,j}(\vec{r}_i,n_i;\vec{r}_j,n_j)\approx E_i^{\varphi}(\vec{r}_i,n_i)$$

Therefore, the free energy of the protein conformations, which populate the TS at its lowest free energy state, is presented in the form:

$$E^{\varphi}(Q)=\sum_i\left(E_i(\vec{r}_i,n_i)+E_i^{\varphi}(\vec{r}_i,n_i)\right)+\sum_{i,i+1}E_{i,i+1}(\vec{r}_i,n_i;\vec{r}_{i+1},n_{i+1})$$

where $E_i^{\varphi}(\vec{r}_i,n_i)$ is the free energy of the termini i in the presence of the external field. Hence, to each protein conformations we can ascribe a statistical weight:

$$\exp\left(-\frac{E^{\varphi}(Q)}{RT}\right)$$

The free energy of the TS can be obtained by taking the logarithm of the sum over all possible weights for the available conformational states in the TS. Because in the presence of an external field the terminus of the secondary structures do not interact with each other the calculation of the partition function Z is best carried out step by step (de Gennes, 1979; Dimitrov & Crichton 1996; Dimitrov et al., 1998). We can begin from the N-terminus of the protein chain and add a new structural terminus i at each step:

$$Z(1,i+1)=$$

$$=\sum_{(\{\vec{r}_i\},\{n_i\})} Z(1,i)\times\exp\left(-\frac{\left(E_{i+1}\left(\vec{r}_{i+1},n_{i+1}\right)+E^{\varphi}_{i+1}\left(\vec{r}_{i+1},n_{i+1}\right)\right)+E_{i,i+1}\left(\vec{r}_i,n_i;\vec{r}_{i+1},n_{i+1}\right)}{RT}\right)\times\Delta\left(n_i,n_{i+1}\right)$$

where $\Delta\left(n_i,n_{i+1}\right)=\Delta\left(n_{i+1}-\left(n_i+\Delta n^{struct.\ region}\left(r_i,r_{i+1}\right)\right)\right)$ and $\Delta n^{struct.\ region}\left(r_i,r_{i+1}\right)$ is the number of residues between the ends i and $i+1$, the $\Delta(x)$ function is equal to 1 if $x=0$ and zero if $x\neq 0$.

The same procedure can be done if we begin with the C-terminus of the chain.

$$Z(i+1,N)=\sum_{(\{\vec{r}_{i+2}\},\{n_{i+2}\})} Z(i+2,N)\times\exp\left(-\frac{E_{i+1,i+2}\left(\vec{r}_{i+1},n_{i+1};\vec{r}_{i+2},n_{i+2}\right)}{RT}\right)\times\Delta\left(n_{i+1},n_{i+2}\right)$$

where $\Delta\left(n_{i+1},n_{i+2}\right)=\Delta\left(n_{i+2}-\left(n_{i+1}+\Delta n^{loop}\left(r_{i+1},r_{i+2}\right)\right)\right)$ and $\Delta n^{loop}\left(r_{i+1},r_{i+2}\right)$ is the minimal loop length between the ends $i+1$ and $i+2$. To obtain the whole partition function we must multiply the N- and C- partition functions for some termini i and take the sum over all of it's possible states $\left(\{\vec{r}_i\},\{n_i\}\right)$.

$$Z(1,N)=\sum_{(\{\vec{r}_i\},\{n_i\})} Z(1,i)\times Z(i,N)$$

$$F^{\varphi}=-RT\times\ln\left(Z(1,N)\right)\text{ and }P^{\varphi}_i\left(\vec{r}_i,n_i\right)=\frac{Z(1,i)\times Z(i,N)}{Z(1,N)} \qquad (2)$$

Calculation of free energy

The free energy of the TS can be calculated on the basis of minimization rules described in our previous study (Dimitrov & Crichton, 1997). For our purpose as a starting point for the minimization procedure the most appropriate approach is the well known classical statistical mechanics Gibbs-Bogoliubov inequality (Gibbs, 1902; Callen, 1985), which is quite general

and does not depend on the character of the investigated system. It states that the free energy of the system of interest is less than, or equal to, the free energy of the system in the presence of an external field which approximates part or all of its internal interactions. If we represent the free energy and the energy levels of the system with and without the presence of an external field φ by F, $\{E_i\}$ and F^{φ}, $\{E_i^{\varphi}\}$ correspondingly, the Gibbs-Bogoliubov inequality takes the form:

$$F \le F^{\varphi} + \langle E - E^{\varphi} \rangle_{P^{\varphi}} \tag{3}$$

Therefore, the free energy of the system of interest can be obtained as a minimum of the right side of the inequality (3) over the external field:

$$F = \min_{E^{\varphi}} \{ F^{\varphi} + \langle E - E^{\varphi} \rangle_{P^{\varphi}} \}$$

where F^{φ} and P^{φ} are determined as in (2). It is important to note that inequality (3) gives only the upper limit to the free energy. Therefore, the choice of the adjustable parameters and the corresponding potential functions for the pairwise residue interactions are restricted by the requirement that the limit on the right of inequality (3) be as small as possible.

General charges and minimization of the free energy

From the description of free energy in terms of space and sequence coordinates of the N- and C- terminus of the secondary structures $(\{\vec{r}_i\},\{n_i\})$, we pass to the description in terms of the physicochemical characteristics of the individual residues. This gives the possibility to represent the free energy of the protein conformations in terms of some average characteristics of its sequence, such as distribution of hydrophobic and hydrophilic residues along the sequence, as well as the distribution of residue volumes which are important to account for the steric restrictions, distribution of ionized groups important in electrostatic interactions and so on. For this, it is necessary to represent the non-local interactions between the residue groups which are far apart along the sequence, but close in space in a few multipliers one of which is associated with the physicochemical characteristics of the individual residues and the other with the distance between them. The most appropriate form comes from the analogy with the electrostatic interactions proposed in our previous studies (Dimitrov & Crichton, 1997; Gooly et al., 1998). The non-local pairwise residue interactions are taken in the form:

$$\Delta F_{ij}\left(\vec{r}_i, n_i; \vec{r}_j, n_j\right) = \Delta F_i\left(n_i\right) \times f_{ij}\left(\vec{r}_i, \vec{r}_j\right) \times \Delta F_j\left(n_j\right) \tag{4}$$

where $(\vec{r}_i, n_i)$ represent all space and sequence positions inside the N- and C- terminal parts of the secondary structures; ΔF_i and ΔF_j are associated with the physicochemical properties of the individual residue; $f_{ij}\left(\vec{r}_i, \vec{r}_j\right) = f_{ji}\left(\vec{r}_j, \vec{r}_i\right)$ is the geometrical factor which characterizes the distance dependence of residue-residue interactions determined within a sphere with radius R (depending on the type of residue-residue interactions) and has the form:

$$f_{ij}\left(\vec{r}_i, \vec{r}_j\right) \approx \vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right) \cdot f\left(\left|\vec{r}_i - \vec{r}_j\right|, R\right)$$

$$\vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right) = 0 \text{ if } \left|\vec{r}_i - \vec{r}_j\right| > R$$

$$\vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right) = 1 \text{ if } \left|\vec{r}_i - \vec{r}_j\right| \le R$$

$f\left(\left|\vec{r}_i - \vec{r}_j\right|, R\right)$ is the weight radial factor inside the sphere which has to smooth out the distribution of residue side chains.

In the Gibbs-Bogolubov inequality (3) the last term $\left\langle E - E^{\varphi}\right\rangle_{P^{\varphi}}$ has to be averaged over the space and sequence coordinates of the N- and C-terminal residues using the probability P^{φ} defined in the presence of an external field φ. It is important to remember that averaging is made under the assumption that the N- and C-terminus are independent. In other words, we do not take into account the restrictions imposed by the chain connectivity. So each residue can interact with itself. This is the price of the mean field approximation. Whether or not the probability P^{φ} for such a situation will tend to zero depends very much on the protein sequence. Taking into account the expression for the pairwise interactions (4), and if the interaction of the individual residues with the external fields are defined as $\Delta F \times \varphi$, the free energy of the protein conformations can be represented in the form:

$$F = F^{\varphi} + \frac{1}{2}\sum_{\substack{ij \\ i \neq j}}\sum_{\vec{r}_i \vec{r}_j}\sum_{mn} q^m\left(\vec{r}_i\right) \cdot f_{ij}\left(\vec{r}_i, \vec{r}_j\right) \cdot q^n\left(\vec{r}_j\right) - \sum_i \sum_{\vec{r}_i} \sum_m q^m\left(\vec{r}_i\right) \cdot \varphi^m\left(\vec{r}_i\right) \tag{5}$$

The summation in (5) is over the terminus of the secondary structures i, over their internal coordinates $\vec{r}_i$ and over the type of residue charges m and n. $q^m\left(\vec{r}_i\right) = \left\langle \Delta F_i^m \right\rangle_{P^{\varphi}}$ gives on average the most probable sequence regions and the corresponding residue characteristics in it, such as hydrophobicity, hydrogen-bonding propensity and so on, which under the fluctuations

are located at the internal coordinate $\vec{r}_i$ of termini i interacting with the given molecular field $\varphi^m(\vec{r}_i)$. We call $q^m(\vec{r}_i)$ the general charge by analogy with the electrostatic interactions. For the self-consistent solution from $\frac{\delta F}{\delta \varphi^m(\vec{r}_i)} = 0$ and $\frac{\delta F}{\delta q^m(\vec{r}_i)} = 0$ we obtain:

$$\frac{\delta F^{\varphi}}{\delta \varphi^m(\vec{r}_i)} = q^m(\vec{r}_i) \text{ and } \varphi^m(\vec{r}_i) = \sum_{j \neq i} \sum_{\vec{r}_j} \sum_{n} f_{ij}(\vec{r}_i, \vec{r}_j) \cdot q^n(\vec{r}_j) \tag{6}$$

where $\varphi^m(\vec{r}_i)$ plays the role of a self-consistent field for the charges of type m. The self-consistent fields $\varphi^m(\vec{r}_i)$ solve the problem of finding the appropriate external fields by which the Gibbs-Bogoliubov inequality (3) has to be minimized. Moreover they are determined in terms of some average physicochemical characteristics of the protein sequence. Minimization of free energy in (5) is carried out by an iteration process. We can begin with some arbitrary chosen molecular fields or some distribution of average charges. In each step of the iteration procedure using the equation (6) we define the molecular fields $\varphi^m(\vec{r})$ which act at each position $\vec{r}$ in the lattice. These fields depend on the distribution of mean charges obtained at the previous step. Thus at equilibrium the charge distribution obtained by the molecular fields must coincide with the charge distribution by which the molecular fields are defined. Depending on the sharpness of the minimum it is possible that the free energy will increase if the step is too long and the minimum is passed. This is overcome by introducing a parameter λ $(0 < \lambda < 1)$ in the form:

$$\varphi(\vec{r}) = \lambda \cdot \left(\varphi(\vec{r})_l - \varphi(\vec{r})_{l-1}\right) + \varphi(\vec{r})_{l-1}$$

where l is the number of iteration steps. If the free energy increases λ is decreased by half.

RESULTS

Calculations were carried out with different sets of parameters for the secondary structure formation presented in table 1. In all cases the mutations of only hydrophobic residues were considered. This is done for two main reasons. Firstly, the Φ-values of hydrophobic residues are based on packing consideration and thus correlate with the change of solvent exposure (Fersht at al., 1992; Milla et al., 1995; Burtun et al., 1997). Secondly, the change of rate constants on mutation of hydrophobic residues correlate with the changes in noncovalent bond

energies and the reorganizations in the molecular structure caused by the mutation (Fersht et al., 1992; Matouschek et al., 1995).

Table 1. Coil-helix and coil-sheet formation parameters (kcal/mol) for each of the 20 naturally occuring amino acids. All data are presented with reference to ALA residue.

	β-sheet			Helix			
Residue	Minor & Kim, 1994 site		O'Neil & DeGrado, 1990	Chakrabartty et al., 1994	Sheraga et al., 1990	Munoz & Serrano, 1995	Yang et al., 1997
	internal	edge					
ALA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CYS	-0.52	-0.08	0.54	0.83	0.05	0.60	1.17
ASP	0.94	0.10	0.62	0.90	0.27	0.59	0.78
GLU	-0.01	-0.31	0.50	0..69	0.06	0.34	0.21
PHE	-0.86	-0.16	0.36	0..93	-0.01	0.47	1.11
GLY	1.2	0.85	0.77	1.88	0.35	1.10	1.05
HIS	0.02	0.01	0.71	1.28	0.26	0.62	0.53
ILE	-1.0	-0.02	0.54	0.71	-0.04	0.35	0.38
LYS	-0.27	0.40	0.12	0.37	0.08	0.15	0.58
LEU	-0.51	0.24	0.15	0.28	-0.04	0.19	0.36
MET	-0.72	0.02	0.27	0.51	-0.07	0.21	0.52
ASN	0.08	0.24	0.70	0.90	0.19	0.60	0.85
TYR	-0.96	-0.11	0.60	0.60	0.03	0.47	1.94
PRO	>3	>4	3.77	4.26	3.35	2.72	2.56
GLN	-0.23	-0.04	0.44	0.57	0.05	0.32	0.80
ARG	-0.45	0.43	0.09	0.21	0.02	0.06	0.41
SER	-0.70	-0.63	0.42	0.79	0.20	0.50	0.76
THR	-1.1	-0.83	0.66	1.33	0.16	0.57	0.88
VAL	-0.82	-0.17	0.63	1.06	0.07	0.51	0.73
TRP	-0.54	0.17	0.32	0.91	-0.02	0.47	1.11

The mutations cover all positions in the final folded structure of the protein which are important for the stability of the structural frame around which the folding pattern and topology are formed in the TS: 1) in the hydrophobic core of the protein, 2) at the edges of the hydrophobic core and 3) at the surfaces of the secondary structures which are exposed to water.

The change upon mutation of the difference in free energy between the TS and the ground state is determined mainly by the difference in free energy of the TS (F_{TS}) between the mutant and wild-type proteins, $RT\ln\left(k_f^{wild-type}/k_f^{mutant}\right) \approx F_{TS}^{mutant} - F_{TS}^{wild-type}$. The rate constant of folding according to the TS theory of the two-state folding reaction is taken in the form, $k_f \approx e^{\left(-\frac{F_{TS}}{RT}\right)} \Big/ e^{\left(-\frac{F_{ground\ state}}{RT}\right)}$ (Zwanzig, 1997). Theoretical Φ-values ($\Phi^{calculation}$) are obtained with reference to the experimentally determined change of free energy in going from D to N states between the mutated and wild-type proteins, $\Delta F_{experiment}^{mutant} = F_N^{mutant} - F_D^{mutant}$ and $\Delta F_{experiment}^{wild-type} = F_N^{wild-type} - F_D^{wild-type}$. Hence, $\Phi^{calculated} = \dfrac{\Delta F_{TS}^{calculated}}{\Delta\Delta F_{experiment}}$, where $\Delta\Delta F_{experiment} = \Delta F_{experiment}^{mutant} - \Delta F_{experiment}^{wild-type}$ and $\Delta F_{TS}^{calculated} = F_{TS}^{mutant} - F_{TS}^{wild-type}$.

Barley chymotrypsin inhibitor 2

The three-dimensional structure of CI2 represents a globular $\alpha+\beta$ domain with 64 residues, no disulfide bridges and *cis*-peptidyl-prolyl bonds (Harpaz et al., 1994). The hydrophobic core is formed by the packing of six-stranded mixed, parallel and antiparallel, β-sheet against a single α-helix. In the β-sheet there are 8 hydrophobic residues which contribute to the hydrophobic core. The α-helix consists of three turns with 4 hydrophobic groups on its buried side, which contribute to the formation of the hydrophobic core. In addition there is a small hydrophobic pocket near one end of the reactive-site loop, which forms a right-handed crossover between the β-strands 3 and 4.

Detailed kinetic and equilibrium experiments have established that the folding and unfolding of wild-type CI2 and a range of mutants conform to a two-state model with a single rate-determining TS (Jackson & Fersht, 1991). Φ-Value analysis has shown that mutations affecting residues in the hydrophobic core have higher Φ-values than those on the surface of

the protein (Itzhaki et al., 1995). However, the Φ-values have a broad distribution and only one residue ALA16 has its full native interaction energy in the TS (indicated by a Φ-value of ~ 1.0). In general, the TS of CI2 is compact but relatively uniformly unstructured, with tertiary and secondary structure being formed in parallel.

Our calculations together with the experimental Φ-values are represented in fig.3. The calculated and the experimental Φ-values are in remarkable agreement. From the calculated Φ-values it follows that in CI2 the loop regions, which connect the α-helix with the β-sheet are flexible and conformationaly degenerated but still continue to contribute to the screening of the hydrophobic core. On the other hand, loop flexibility increases the conformational variability of the α-helix and the β-strands in the β-sheet, which makes the packing between the α-helix and the β-sheet less effective. From the calculated Φ-values it also follows that the N-terminus of the α-helix (ALA35-Φ=0.94) and the β-hairpin formed between the 5 and 4 β-strands (LEU68-Φ=.39; ILE76-Φ=0.1) are the regions which are most effectively packed in the TS. This is in full agreement with the experimental Φ-values for these regions: 1.06 for ALA/GLY35 mutation and 0.53 for LEU/ALA68 mutation (Itzhaki et al., 1995). It follows that our model predicts correctly the key role of residue ALA16 in the nucleation condensation mechanism for the folding of CI2 protein.

Interestingly, the calculations show that the Φ-values in the minicore are above the experimental Φ-values. The usual interpretation of the fractional Φ-values for the minicore suggests that the minicore is not a nucleation center for folding *via* hydrophobic clustering, nor that the tertiary interactions appear to form later than secondary interactions, but instead are formed in parallel with them (Itzhaki et al., 1995). Also because the Brönsted plot of $\ln k_F$ *versus* $\Delta F_{N-D} / RT$ is 0.3 the interpretation is that approximately 30% of the interactions in the minicore are formed in the TS state relative to the native state N. According to our model the right-handed crossover between the β-strands 3 and 4, which form the minicore, has to appear already in the D state. As a consequence some of the interactions in the minicore in the TS are approximately the same as in the denatured state. As a result the perturbation of such interactions will result in small and in some cases zero Φ-values. This follows from the fact that $F_{TS}^{mutant} - F_{TS}^{wild-type} \approx F_D^{mutant} - F_D^{wild-type}$.

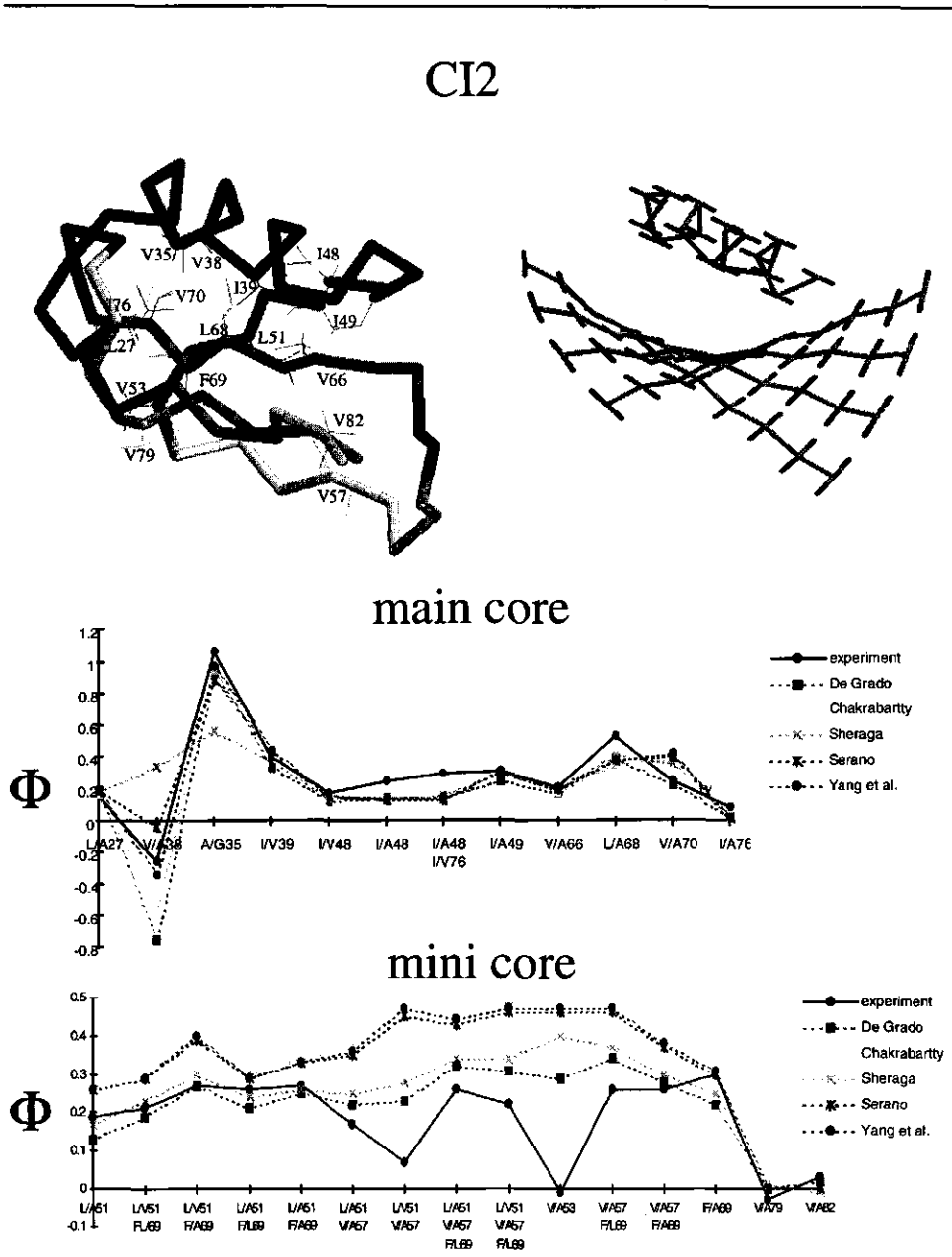


Figure.3. Experimental Φ-values, from refolding kinetic experiments, are compared with the calculated Φ-values for the barley chymotrypsin inhibitor 2 (CI2). The Φ-values for mutation in the core and in the minicore of the CI2 are shown separately. Φ-values are represented together with the schematic representation of the secondary and tertiary structure of the CI2 as well as its lattice model and the mutated hydrophobic residues. Calculations are carried out with a set of different experimentally determined parameters for the coil-α-helix and coil-β-strand transitions, table 1.

Association of two fragments of barley chymotrypsin inhibitor -[CI2-(20-59) and CI2-(60-83)]

CI2 has been truncated in two fragments CI2-(20-59) and CI2-(60-83) (de Prat Gay et al., 1994). The TS for the association of these fragments has been analyzed by protein engineering methods. It has been shown that the two fragments associate in a second-order and that the TS for the association reaction is very similar to the TS for the folding reaction of the intact protein.

According to our model the topology plays a critical role in the rate limiting step of protein folding. Therefore, the experimentally observed similarity between the TS of the intact and truncated proteins is very intriguing taking into account that for the folding reaction of the truncated protein there are not any topological preliminary requirements for folding. The calculated and the experimental Φ-values are represented in fig.4. The agreement between the two sets of data is reasonable. According to our model the fragments should not form native-like structures prior to their association in full agreement with the experimental data. More importantly the agreement with the experimental data strongly supports the mechanism in which the folding reaction for the truncated protein also proceeds through the formation of a high free energy nucleus at the TS level. Thus, the main result from the comparison of the theoretical and experimental Φ-values is that the truncated and the intact proteins have the same nucleation sites as indicated by their high Φ-values. Calculated Φ-values are shifted down relative to the experimental Φ-values. This is a result of the fact that in our model we do not take into account the volume effects and van der Waals interactions. On the other hand the shift clearly shows that the TS for the truncated protein is nearer to the native state than the TS for the intact protein.

Association of the two domains of the Arc repressor of phage P22

The Arc repressor of phage P22 consist of two polypeptide chains each 53 residues long (Breg et al., 1990; Milla et al., 1995; Milla et al., 1994). The protein has a well-defined dimeric structure which consists of an intersubunit, antiparalell β-sheet and a four α-helices packed around a single hydrophobic core. The association reaction of the two polypeptide chains occurs reversibly and without detectable intermediates in both equilibrium and kinetic studies.

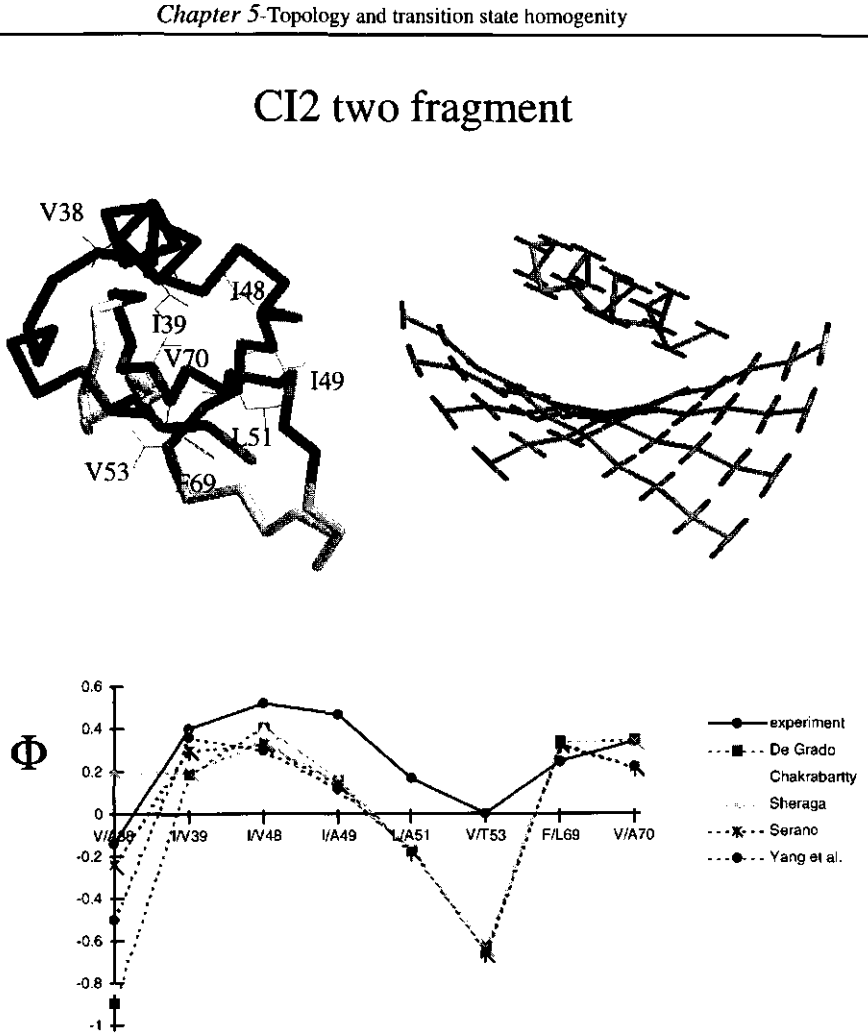


Figure.4. Experimental Φ-values, from refolding kinetic experiments, are compared with the calculated Φ-values for the association of two fragments of barley chymotrypsin inhibitor -[CI2-(20-59) and CI2-(60-83)]. Φ-values are represented together with the schematic representation of the secondary and tertiary structure of the CI2 (because of its structural similarity with its truncated form) as well as its lattice model. Mutated hydrophobic residues are also shown. Calculations are carried out with a set of different experimentally determined parameters for the coil-α-helix and coil-β-strand transitions, table 1.

This indicates that the TS for the association reaction, similarly to the association reaction of the truncated CI2 protein, is represented by a single folding nucleus. The experimental and theoretical Φ-values are represented in fig.5. The agreement between the two sets of data is good. All nucleation sites, indicated by their high Φ-values, are located in the hydrophobic

core. The calculated Φ-values, similar to the association reaction of the truncated form of CI2 protein, are shifted down relative to the experimental data. This is because the size of the nucleus is approximately the same for the Arc repressor, CI2 and the truncated form of CI2 proteins, the position of the TS along the reaction coordinate is mainly determined by the total exposure of the loop regions. Thus, our prediction is that the TS states for the Arc repressor and the truncated form of CI2 protein are closer to the native state than the TS of the CI2 protein.

Activation domain of human procarboxypeptidase

The three-dimensional structure of Ada2H represents a globular open-sandwich $\alpha+\beta$ domain with 80 residues, no disulfide bridges and 4 *trans*-prolyl peptide bonds (Garcia-Saez et al., unpublished data; Vendrell et al., 1991). The hydrophobic core is formed by the packing of a four-stranded antiparallel β-sheet against two α-helices. The two helices are oriented almost exactly antiparallel to each other, are all on the same side of the β-sheet, and the helix axes form an angle of ~ 45° relative to the direction of the β-strands. The loops linking the secondary structures are significantly less well ordered than the rest of the molecule. Equilibrium denaturation by urea or temperature is fully reversible at pH 7.0 and fits to a two-state transition (Villegas et al., 1995).

Kinetics of unfolding and refolding followed by fluorescence does not show the presence of any kinetic intermediates accumulating in the folding reaction (Villegas et al., 1995). All these data indicate that the folding of this domain is consistent with a nucleation condensation mechanism, where the folding pathway has to proceed through the formation of a high free energy nucleus (Fersht, 1995).

The data for the free energy of denaturation, extrapolated to zero concentration of denaturant, are represented in Table 2 (unpublished data of L. Serrano EMBL, Heidelberg). The calculated Φ-values together with the schematic representation of the secondary and tertiary structure of Ada2H domain are represented in fig.6. As follows from our calculations over the different sets of experimentally determined parameters for the coil-α-helix and coil-β-strand transitions, the residues, which are involved in the folding nucleus of the Ada2H domain, are ILE23, ILE15, LEU26, ALA52 and VAL54. Their Φ-values, as calculated based on O'Neil and DeGrado parameters which seems to be more consistent with the experimental

Φ-values, are 2.60 for the ILE/VAL23 mutation, 1.0 for the ILE/VAL15 mutation, 0.59 for the VAL/ALA54 mutation, 0.41 for the LEU/VAL26 mutation and 0.35 for the ALA/GLY52 mutation.

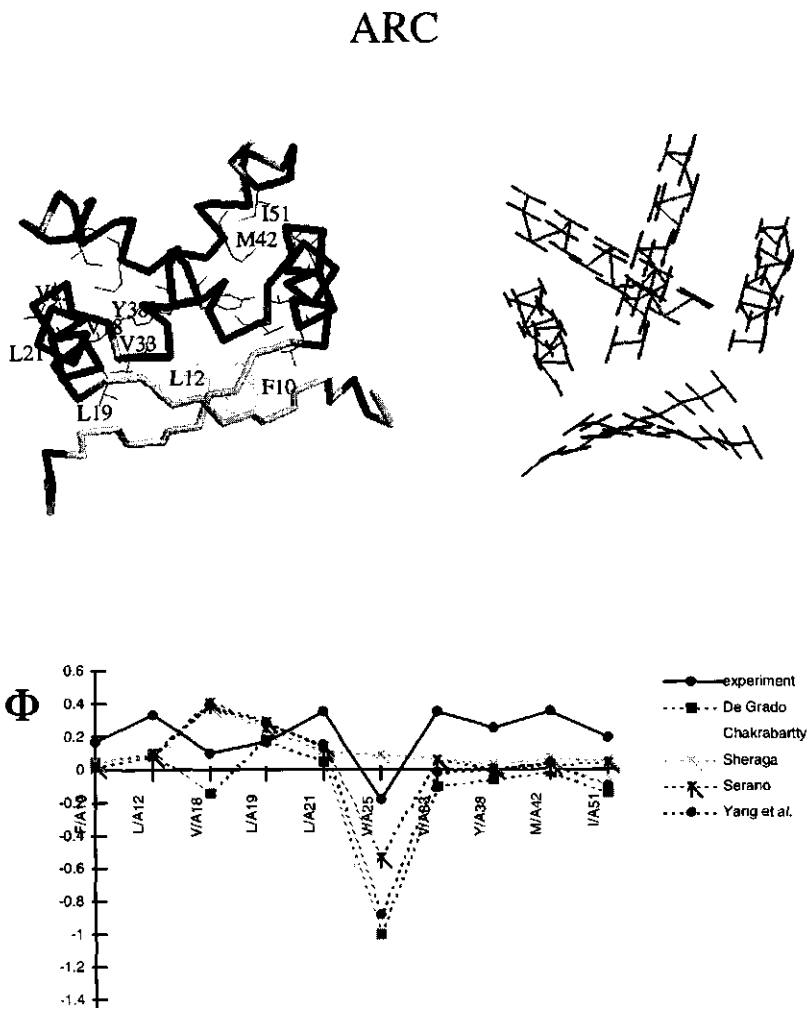


Figure.5. Experimental Φ-values, from refolding kinetic experiments, are compared with the calculated Φ-values for the association of the two domains of the Arc repressor of phage P22. Φ-values are represented together with the schematic representation of the secondary and tertiary structure of the Arc repressor and the mutated hydrophobic residues. The lattice model of Arc repressor is also shown. Calculations are carried out with a set of different experimentally determined parameters for the coil-α-helix and coil-β-strand transitions, table 1.

Table 2. Free energies of the wild-type and mutant proteins on denaturation, extrapolated to zero concentrations of denaturant.

Protein	free energy of denaturation (kcal/mol)
wild-type	-4.4±0.1
VAL/ALA12	-2.6±0.2
ILE/VAL15	-3.9±0.1
ILE/VAL23	-4.3±0.1
LEU/VAL26	-3.1±0.1
PHE/LEU37	-2.3±0.2
ALA/GLY52	-2.7±0.1
VAL/ALA54	-3.3±0.2
PHE/ALA67	-2.8±0.1
ILE/VAL73	-2.9±0.1
ILE/ALA77	-2.5±0.1

Thus, our prediction is that ILE23 and ILE15 are the key residues for the folding of the Ada2H domain. It is connected with the stability of the N-terminus of the α-helix 2 and its packing against the β-sheet. In particular ILE15 from the β-strand 1 and ALA52 and VAL54 from β-strand 4 form a small hydrophobic pocket, that packs against the ALA26, while the ALA23 is screened by the short proline rich connection between β-strand 3 and β-strand 4, fig.7.

DISCUSSION

Kinetic and experimental studies have shown that small monomeric proteins follow a two-state folding mechanism in which only the denatured state and the native states are significantly populated (Viguera et al., 1996; Alexander et al., 1992; Schindler et al., 1995; Kragelund et al., 1995; Jackson & Fersht, 1991; Burton et al., 1997; Villegas et al., 1995). This means that the denatured state and the native state are separated by a free energy barrier.

Experimental and theoretical studies represent strong evidence that at the top of the free energy barrier the folding toward the native conformation is based on the nucleation growth mechanism (Abkevich et al., 1994; Fersht, 1995; López-Hernández, Serrano, 1996). Thus, the folding problem is reduced to the question of how the folding nucleus at the top of the free energy barrier is formed from the ensemble of partly structured conformations in the denatured state.

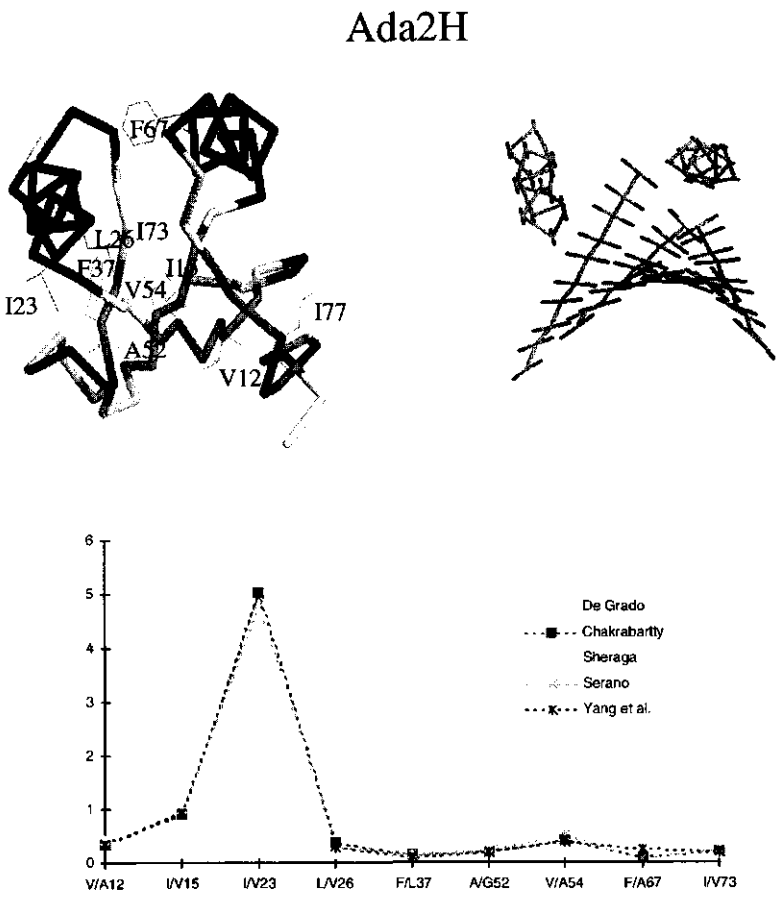


Figure.6. Representation of the calculated Φ-values, based on experimental data for the free energy of denaturation of the Ada2H domain. Φ-values are represented together with the schematic representation of the secondary and tertiary structure of the Ada2H domain and the mutated hydrophobic residues. The lattice model of the Ada2H domain is also shown. Experimental data for the free energy of denaturation together with the sequence and coordinates for the Ada2H domain as well as the hydrophobic residues, which had been mutated, was kindly provided to us by L. Serrano EMBL, Heidelberg (unpublished data).

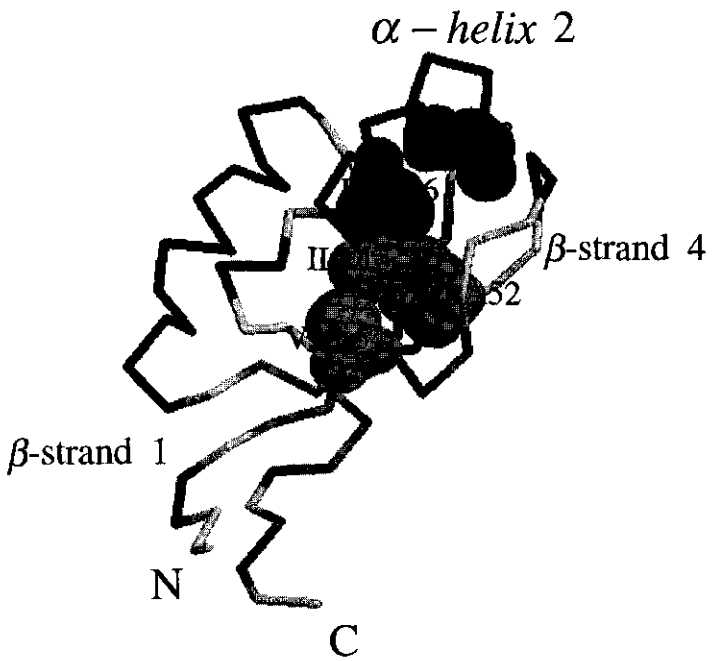


Figure.7. Schematic representation of the stabilization of the N-terminal region of the α -helix 2 from the Ada2H domain in the TS by interactions of the side chains of ILE23 and LEU26 with residues that will form the hydrophobic core. The field spheres are drawn with the full van der Waal's radii. The N-terminal region of the α -helix 2 and the residues ILE15 from β-strand 1 and ALA52 and VAL54 from β-strand 4 constitute the proposed nucleation site.

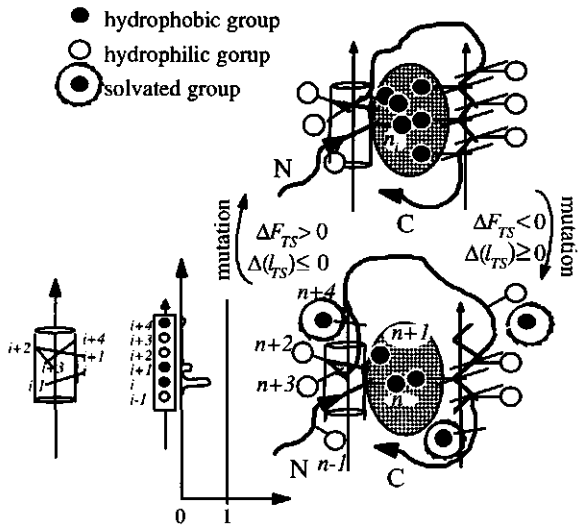
In this study we address this question by focusing on the role of topological restrictions on the formation of the folding nucleus. The important role of such restrictions is shown in recent characterization of the refolding properties of a number of small monomeric proteins. It has been shown that there is a very strong correlation between the distribution of the sequence distance between all pairs of contacting residues and the rate at which the proteins fold (Plaxco et al., 1998). Moreover, there is direct experimental evidence that both sequence and topological restrictions, operate not only at the TS level but in the denatured state as well (Martinez et al., 1998; Grantcharova, 1998). This leads to certain obligatory steps in the folding pathway. For example, it has been shown that optimum sequences for particular α-helices or β-turns along the polypeptide chain significantly stabilize the protein, and in many cases produce an acceleration of the folding reaction (Viguera et al., 1997).

Thus, while in the denatured state the nucleation process is favored by certain α-helices or β-hairpins along the sequence, in the TS the nucleation process is favored by their packing

together with some other portions of the polypeptide chain. It is because of the requirement for maximum screening of the hydrophobic groups, minimum screening of the hydrophilic ones and the maximum saturation of the hydrogen backbones, that the packing of the polypeptide chain in the TS leads to the formation of the folding nucleus with native-like topology and approximately correctly formed secondary structures. All other folding alternatives which include both changes in topology and secondary structures are separated by a high free energy gap. The folding nucleus is represented by the packing of α- and(or) β-secondary structures, the type and mutual contacts of which determines the folding pattern of the nucleus. Whereas, the order in which loop regions with different degrees of order connect the secondary structures in the folding pattern determines the topology of the nucleus.

Loop regions play a critical role in the nucleation process (Grantcharova et al., 1998; Martinez et al., 1998; Gruebele & Wolynes, 1998). For simplicity we assume two main types of loop regions in our model: 1) loop regions involved in crossover and 2) loop regions located in the periphery (for example β-hairpins). Along the polypeptide chain loops are located predominantly in regions rich in polar residues, whereas α- and(or) β-secondary structures are found predominantly in regions rich in hydrophobic residues. These sequence restrictions combined with the topology determine the order of the mutual packing of the secondary structures leading to the formation of the folding nucleus. In general the loop regions located at the periphery are shorter, solvent exposed and with less tertiary contacts with the rest of the protein than the loop regions involved in crossovers. As a first approximation, we take into account only the tertiary interactions of the crossovers.

Because of the flexibility of the crossovers, secondary structures connected to these loops, have relatively independent conformational freedom. Therefore, the free energy of the secondary structures is mainly determined by the free energy resulting from the interactions of residues within and between the secondary structures and the entropy associated with their conformational freedom. From these considerations it is clear that changes in both topology and secondary structures of the folding nucleus result in redistribution of the secondary structures along the sequence and the order of their mutual packing. As a result of this redistribution the terms contributing to the free energy of the secondary structures and the loop regions change. This would lead to the change of the relative contributions of the folding nucleus and the loop regions to the free energy of the protein molecule at the TS level, thus changing the Φ-values, fig.8.



Fraction of TS conformations in which residue n along the sequence is located at each of the coresponding α-helix positions

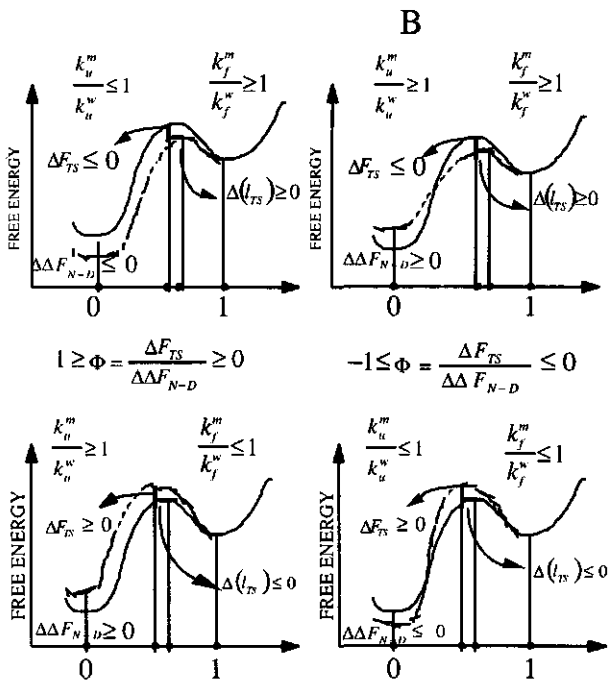


Figure.8. Schematic representation of the redistribution of the lengths between the secondary structures in the nucleus and the loop regions, as seen through the change of the free energy profile of the mutant relative to the wild-type protein. Reaction coordinate is represented by the solvent exposures of the D state (l_D), TS (l_{TS})

and the N state (l_N). It is assumed that the mutations do not change the difference of solvent exposure between the D and N states, $\Delta(l_D - l_N) \approx 0$, while for the TS $\Delta(l_{TS} - l_N) \approx \Delta(l_{TS}) \neq 0$. The assumption that $\Delta(l_D - l_N) \approx 0$ is based on the fact that for two-state proteins only the D and N state are significantly populated. In the D state there are reconfigurations which include both topological changes and local hydrogen bond breaking and formation, while in the N state there are only local reconfigurations around the site of the mutation. Hence the effect of the mutations on the change in solvent exposure in going from N to D state is negligible. The interpretation of the experimental data is based on the fact that the TS ensemble is dominated by conformations in which the contacts between a few residues, which serve as a nucleation sites, are highly populated, fig.8A. This follows from the requirement of hydrophobic core formation. Thus α- and(or) β-secondary structures, which are packed around the core, have one hydrophobic (in fig.8A this corresponds to residues n, $n+1$ and $n+4$ from the α-helix) and one hydrophilic surface (residues $n-1$, $n+2$ and $n+3$ from the α-helix). Along the sequence they are separated by flexible loops rich in polar residues. All other conformations, in which hydrophobic residues are exposed to the water (for example conformations in which residue n from position i along the α-helix is translated to position $i+2$) or hydrophilic residues are buried in the core (for example conformations in which residue n +3 from position i +3 along the α-helix is translated to position i), are separated by a free energy barrier. The protein structure around the nucleus has the most conserved geometry, while other protein portions, such as loop regions, are in process of being formed, so they are more flexible. The loop regions are characterized not only with degenerate conformations, but also with the fluctuations of their lengths. Increasing or decreasing of the stability of particular secondary structure, leads to an increase or a decrease in solvent exposure of some of the hydrophobic groups located on the edges of the hydrophobic core. This change of solvent exposure is a mesure of the redistribution of the lengths between the secondary structures and loop regions and corresponds to the movement of the TS toward the denatured D or the native N states, respectively. At the same time, because of the general inequalities $1 \geq \Phi = \frac{\Delta F_{TS}}{\Delta\Delta F_{N-D}} \geq 0$ and $-1 \leq \Phi = \frac{\Delta F_{TS}}{\Delta\Delta F_{N-D}} \leq 0$, it follows that $0 \leq \Delta F_{TS} \leq \Delta\Delta F_{N-D}$ or $0 \geq \Delta F_{TS} \geq \Delta\Delta F_{N-D}$ and $\Delta F_{TS} \geq 0 \geq \Delta\Delta F_{N-D}$ or $\Delta\Delta F_{N-D} \geq 0 \geq \Delta F_{TS}$, respectively. This leads to an increase or a decrease of the rate constants of unfolding k_u and folding k_f for the mutant protein relative to the wild-type, as is shown in fig.8B.

In the case of CI2 the folding nucleus is located in the hydrophobic core formed by the packing of a single α-helix against a six-stranded mixed, parallel and antiparalell, β-sheet. The crossover loops are those between the α-helix and the corners of the β-sheet as well as the right-handed crossover between β-strands 3 and 4. There is only one relatively short loop

connection which is located in the middle of the β-sheet. The symmetry of the lattice, the symmetry of the location of the crossover loops and the location of the β-hairpin in the middle of the β-sheet force the residues with the highest Φ-values to be located in the hydrophobic core. Moreover, the location of the residues with highest Φ-values is not changed if the crossover between the 3 and 4 β-strands is cut. Arc repressor, a predominantly α-protein, consists of two different polypeptide chains whose association results in a structure in which the secondary structures and the loops are symmetrically distributed around a single hydrophobic core. Similar to CI2, in the Arc repressor, the residues with highest Φ-values are also located in the hydrophobic core.

Our results together with the experimental data are in contrast to the funnel scenario for protein folding where the TS is represented by an ensemble of structurally different conformations with similar energies (Bryngelson, 1995). According to this approach, the folding occurs down a funnel along different parallel pathways. This implies the existence of different regions along the sequence and in the final folded structure that serve as nucleation sites. One expects that by breaking loops it is possible to favour certain pathways, rather than others. This means that the Φ-values for certain nucleation sites could be drastically reduced or even nullified. However, this is in contrast to the experimental results reported for CI2 where actually one of the crossovers was cut, as described above, but the experimental Φ-values did not change significantly. Our calculated Φ-values for this modified protein show the same distribution as the experimental Φ-values, but with the actual Φ-values shifted down (fig.4). According to our model we interpreted these results as a shift of the position of the TS along the reaction coordinate toward the native state. This is due to increasing of the tertiary contacts between the fragments of CI2 and decreasing of their entropic contributions. This means that both the wild-type CI2 and the CI2 with the active side loop cut have the same common nucleation sites. This strongly supports our assumption that the TS at its lowest free energy state is populated principally by conformations with native-like topology and approximately correctly formed secondary structures.

Because of lack of symmetry essential for the localization of the nucleation sites fully in the hydrophobic core, in some proteins these nucleation sites might in part be situated at the periphery. This is the case for example for the Ada2H domain whose nucleation sites are partly located at the periphery according to our calculations. In the Ada2H domain the only β-hairpin is located at the periphery of the β-sheet, not in the middle as in the CI2. Other β-strands are

connected with crossovers with the two α-helices which are packed from the same side against the β-sheet. According to our calculations the nucleation sites are located between the first α-helix and the β-hairpin formed between the 3 and 4 β-strands. This is similar to the experimental results which have been reported for the SH3 domain family, where the nucleation sites are located at the periphery on a single distal loop β-hairpin (Grantcharova et al., 1998; Martinez et al., 1998). It has been shown that the formation of this β-hairpin is an obligatory and rate-limiting step in the folding reaction (Martinez et al., 1998). This strongly suggests that the formation of the β-hairpin between the 3 and 4 β-strands is also an obligatory step in the folding pathway of the Ada2H domain.

Our model has some limitations. The geometrical representation of the loop regions does not include knowledge of their space coordinates. Therefore, it is not possible to account for the detailed contacts between residues located in the loop regions and the rest of the protein. Nevertheless, some contributions are taken into account. These include: 1) the free energy of their bending; 2) connectivity restrictions between the secondary structures and 3) non specific hydrophobic interactions in the case of crossover loops. We have to maintain that the detailed knowledge of the loop coordinates are mainly needed for the electrostatic interactions, for the short turns (as in β-hairpins) and for the final packing of the loops with the rest of the protein body in the native state. In the TS long loops are disrupted from the rest of the protein and as a result their long-range interactions are strongly weakened, whereas the short connections have negligible long-range contacts.

Conclusions

A theoretical model for the folding of a two-state small monomeric proteins is proposed. The folding problem is reduced to the question of how the folding nucleus in the TS is formed from the ensemble of rapidly interconverting partly structured conformations in the denatured state. It is proposed that in the denatured state the folding is energetically favored by certain highly fluctuating nucleation regions (α-helices and(or) β-hairpins), which in the experiments based on site directed mutagenesis are revealed by their high Φ-values. In the TS the folding is favored by the packing of these nucleation regions together with other portions of the polypeptide chain, thus leading to a broad distribution of the Φ-values. As a result, in the TS folding nucleus with native-like topology, approximately correctly formed secondary structures

and loops with different degrees of order are favored by the existence of a high free energy gap. A self-consistent molecular mean field theory and a lattice model of the folding nucleus are presented. The lattice model is based on packing of idealized α- and(or) β-secondary structures. A statistical mechanics theory of a linear cooperative system is used to 'inscribe' a given sequence onto its corresponding lattice model. Molecular mean fields are used to approximate the non-local interactions between residues which are far apart in the sequence but close in space. The local interactions between residues, which are close in sequence - as in the α-, β- or loop regions, are accounted for in an explicit form based on experimental parameters. The distribution of the α- and (or) β-regions along the sequence and in the lattice, together with the molecular fields, are self-consistently optimized - threading the sequence in the initial fields, determining the distribution of the secondary structures along the sequence and in the lattice, optimizing the fields to this distribution etc.

Acknowledgments

We thank Dr. Luis Serrano and co-workers from EMBL, Heidelberg for generously providing the sequence, space coordinates, mutated residues and the experimental data for the free energy of denaturation of mutated and wild-type proteins of Ada2H domain (unpublished data). We also thank Dr. Luis Serrano for the possibility to represent some preliminary results of our theoretical approach on a seminar in his laboratory, for reading the manuscript and for the helpful discussions.

REFERENCES

Abkevich, V. I., Gutin A. M., & Shakhnovich E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.

Alexander, P., Orban, J., & Bryan, P. (1992). Kinetic analysis of folding and unfolding the 56-amino acid IgG-binding domain of streptococcal protein-G. *Biochemistry* **31**, 7243-7248.

Breg, J. N., Opheusden, J. H. J., Burgering, M. J. M., Boelens, R., & Kaptein, R. (1990). Structure of Arc repressor in solution: evidence for a family of β-sheet DNA - binding proteins. *Nature* **346**, 586-589.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins Struct. Funct. Genet.* **21**, 167-195.

Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L., & Oas, T. G. (1997). The energy landscape of a fast-folding protein mapped by Ala-Gly Substitutions. *Nat. Struct. Biol.* **4**, 305-310.

Callen, H. B., (1985). Thermodynamics and an introduction to thermostatistics. John Wiley & Sons, New York.

Chakrabartty, A., Kortemme, T., & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing interactions. *Protein Sci.* **3**, 843-852.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E., & Fersht, A. R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430-440.

Dalal, S., Balasubramanian, S., & Regan, L. (1997). Changing β-sheet into α-helix. *Nat. Struct. Biol.* **4**, 548-552.

Dalby, P. A., Clarke, J., Johnson, C. M., & Fersht, A. R. (1998). Folding Intermediates of Wild-type and Mutants of Barnase. I. Use of Φ-Values Analysis and m-Values to Probe the Cooperative Nature of the Folding Pre-equilibrium. *J. Mol. Biol.* **276**, 625-646.

Dalby, P. A., Clarke, J., Johnson, C. M., & Fersht, A. R. (1998). Folding Intermediates of Wild-type and Mutants of Barnase. II. Correlation of Changes in Equilibrium Amide Exchange Kinetics with the Population of the Folding Intermediate. *J. Mol. Biol.* **276**, 647-656.

de Gennes, P. G. (1979). Scaling concepts in polymer physics. Ithaca, New York: Cornell University Press.

De Prat Gay, G., Ruiz-Sanz, J., Davis, B., & Fersht, A. R. (1994). The structure of the transition state for the association of two fragments of the barley chymotrypsin 2 to generate native-like protein: Implications for mechanisms of protein folding. *Proc. Natl. Acad. Sci. USA* **91**, 10943-10946.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D., & Chan, H. S. (1995). Principles of protein folding- aperspectives from simple exact models. *Protein Sci.* **4**, 561-602.

Dimitrov, R. A., & Crichton, R. R. (1997). Self-Consistent Field Approach to Protein Structure and Stability. I. pH Dependence of Electrostatic Contribution. *Proteins Struct. Funct. Genet.* **27**, 576-596.

Dimitrov, R. A. & Crichton, R. R. (1996) Tertiary fold prediction of globular proteins: A molecular field approach. 5th international conference: Perspectives on protein engineering, "From folds to functions", Montpellier, France, 2-6 march.

Dimitrov, R. A., Crichton, R. R. & Vervoort, J. (1998). From Fold Prediction to Protein Design Using Self-Consistent Field Approach. Twenty-third annual Lorne conference on protein structure and function. Australia, 8-12 February.

Felix, B. S., & Brooks, III C. L. (1998). Molecular picture of folding of a small α/β protein. *Proc. Natl. Acad. Sci. USA* **95**, 1562-1567.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

Fersht, A. R. (1995). Optimization of rates of protein folding: The nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA* **92**, 10869-10873.

Fersht, A. R., Matouschek, A., & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771-782.

Gibbs, J. W. (1902). Elementary Principles in Statistical Mechanics New Haven.

Gooley, P. R., Keniry, M. A., Dimitrov, R. A., Marsh, D. E., Keizer, D. W., Gayler, K. R., & Grant, B. R. (1998). The NMR solution structure and characterization of pH dependent chemical shifts of the beta-elicitin, cryptogein. *J. Biom. NMR* in press.

Grantcharova, V. P., Riddle, D. S, Santiago, J. V., & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol.* **4**, 715-721.

Gruebele, M., & Wolynes, P. G. (1998). Satisfying turns in folding transitions. *Nat. Struct. Biol.* **5**, 663-665.

Itzhaki, L. S., Otzen, D. E., & Fersht, A. R. (1995). The structure of the transition state for protein folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.

Jackson, S. E., & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for two-state transition. *Biochemistry* **30**, 10428-10435.

Jonsson, T., Waldburger, C. D., & Sauer, T. (1996). Nonlinear free energy relationships in arc repressor unfolding imply the existence of unstable, native-like folding intermediates. *Biochemistry* **35**, 4795-4802.

Kragelund, B. B., Robinson, C. V., Knudesn, J., Dobson, C. M., & Poulsen, F. M. (1995). Folding of four-helix bundle: studies of acyl-coenzyme A binding protein. *Biochemistry* **34**, 7217-7224.

Li, A., & Daggett, V. (1994). Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA* **91**, 10430-10434.

López-Hernández, E., & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Folding and Design* **1**, 43-55.

Martinez, J. C., Tereza Pisaborro, M., & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **5**, 721-729.

Matouschek, A., Otzen, D. E., Itzhaki, L. S., Jackson, S. E, & Fersht, A. R. (1995). Movement of the Position of the transition State in Protein Folding. *Biochemistry* **34**, 13656-13662.

Milla, M. E., Brown, B. M., Waldburger, C. D., & Sauer, R. T. (1995). P22 Arc Repressor: Transition State Properties Inferred from Mutational Effects on the Rates of Protein Unfolding and Refolding. *Biochemistry* **34**, 13914-13919.

Milla, M. E., Brown, B. M., Waldburger, C. D., & Sauer, R. T. (1995). Protein stability effects of a compete set of alanine substitutions in Arc repressor. *Nat. Struct. Biol.* **1**, 518-523.

Minor, D. L., & Kim, P. S. (1994). Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367**, 660-663.

Minor, D. L., & Kim, P. S. (1994). Context is a major determinant of β-sheet propensity. *Nature* **371**, 264-267.

Munoz, V., & Serrano, L. (1995). Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. II. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides. *J. Mol. Biol.* **245**, 275-296.

Myers, J. K., Pace, C. N., Scholtz, J. M. (1995). Denaturant *m* values and heat capacity changes: Relation to changes in surface areas of protein unfolding. *Protein Sci.* **4**, 2138-2148.

Oliveberg, M., Fersht, A. R. (1996). Formation of electrostatic interactions on the protein-folding pathway. *Biochemistry* **35**, 2726-2737.

O'Neil, K. T., & DeGrado, W. F. (1990). A Thermodynamic Scale for the Helix-Forming Tendencies of the Commonly Occurring Amino Acids. *Science* **250**, 646-651.

Onuchic, J. N., Socci, N. D., Zaida Luthey-Schulten, & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Folding and Design* **1**, 441-450.

Onuchic, J. N., Zaida Luthey-Schulten, & Wolynes, P. G. (1997). Theory of Protein Folding: The Energy Landscape Perspective. *Ann. Rev. Phys. Chem.* **48**, 545-600.

Orengo, C. A., Flores, T. P., Taylor, W. R., & Thornton, J. M. (1993). Identification and clasification of protein fold families. *Protein Engineering* **6**, 485-500.

Plaxco, K. W., Simons, K. T., & Baker, D. (1998). Contact order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *J. Mol. Biol.* **277**, 985-994.

Privalov, P. L. (1979). Stability of proteins. *Advan. Protein Chem.* **33**, 167-236.

Ptitsyn, O. B. (1994). Kinetic and equilibrium intermediates in protein folding. *Protein Eng.* **7**, 593-596.

Schindler, T., & Schmid, F. X. (1996). Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry.* **35**, 16833-16842.

Schindler, T., Herrler, M., Marahiel, M. A., & Schmid, F. X. (1995). Extremely rapid protein folding in the absence of intermediates. *Nat. Struct. Biol.* **2**, 663-672.

Segawa, S., & Sugihara, M. (1984). Characterization of the Transition State of Lysozyme Unfolding. I. Effect of Protein-Solvent Interactions on the Transition State. *Biopolymers* **23**, 2473-2488.

Serrano, L., Matouschek, A., & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for folding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805-818.

Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

Shakhnovich, E. I., Abkevich, V. I., & Ptitsyn, O. B. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98.

Socci, N. D., Onuchic, J. N., & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5861-5868.

Tan, Y-J., Oliveberg, M., & Fersht, A. R. (1996). Titration Properties and Thermodynamics of the Transition State for Folding: Comparison of Two-state and Multi-state Folding Pathways. *J. Mol. Biol.* **264**, 377-389.

Vendrell, J., Billeter, M., Wider, G., Avilés, F. X., & Wüthrich, K. (1991). The NMR structure of the activation domain isolated from porcine procarboxypeptidase B. *EMBO J.* **10**, 11-15.

Viguera, A. R., Serrano, L., & Wilmanns, M. (1996). Different folding transition-states may result in the same native structure. *Nat. Struct. Biol.* **3**, 874-880.

Viguera, A. R., Serrano, L. (1997). Loop length, intramolecular diffusion and protein folding. *Nat. Struct. Biol.* **4**, 939-946.

Viguera, A. R., Villegas, V., Aviles, F. X., Serrano, L. (1997). Favorable native-like helical local interactions can accelerate protein folding. *Folding Design* **2**, 23-33.

Villegas, V., Azaga, A., Catasus, L. I., Reverter, D., Mateo, P. L., Aviles, F. X., & Serrano L. (1995). Evidence for two-state Transition in the Folding Process of the Activation Domain of Human Procarboxypeptidase A2. *Biochemistry* **34**, 15105-15110.

Wójic J., Altmann K-H., & Scheraga H. A. 1990. Helix-coil Stability Constants for the Naturally Occuring Amino Acids in Water. XXIV. Half-cystine Parameters from Random Poly(hydroxybutylglutamine-co-S-methylthio-L-cysteine). *Biopolymers 30*:121-134.

Yang, J., Spek, E. J., Gong, Y., Zhou, H., Kallenbach, N. R. (1997). The role of context on α-helix stability: Host-guest analysis in a mixed background peptide model. *Protein Sci.* **6**, 1264-1272.

Zwanzig, R. (1997). Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. USA* **95**, 148-150.

6 FOLD PREDICTION OF α, β, α/β AND α+β PROTEIN ARCHITECTURES

Roumen A. Dimitrov, Colja Laane, Jacques Vervoort and Robert R. Crichton

Submitted to: *Proteins: Structure, Function and Genetics*

SUMMARY

Recent experimental and theoretical studies strongly support the fact that at the transition state (TS) level the folding of two-state small monomeric proteins, toward the native conformation, is based on the nucleation growth mechanism. The activation barrier of two-state folding reaction is predicted to have a strong uphill slope with small (narrow activation barrier) or large (broad activation barrier) movements along the reaction coordinate. On the bases of our previous results (Dimitrov, R. A., Laane, C., Vervoort, J., & Crichton, R. R. (1998). Topological requirement for the nucleus formation of a two-state folding reaction.

Implications for Φ-values calculations. *Protein Science*, submitted) as well as the present results we will show that the high-energy nucleation for two-state small monomeric proteins is most likely to propagate approximatelly isoenergetically at the TS level, where local and long-range interactions are cooperatively consolidated trough the overall native fold of the protein. All other folding alternatives, which include both changes in topology and secondary structures, are separated by a high free energy gaps. Our strategy can be formulated as follows. Firstly, using the library of different packing patterns (represented as a set of most favorable packing of α- and (or) β-secondary regions) and a limited set of thermodynamically most favorable topologies associated with them, a statistical mechanics theory of a linear cooperative system is used to 'inscribe' a given sequence onto these packing patterns in order to recognize which topology and packing pattern fits the sequence best. Secondly, the main features of the packing patterns are determined by molecular fields. Thirdly, the topologies and the molecular fields are self-consistently optimized by threading the sequence in a initial field, by determining the topology in this field, by optimizing the field to this topology, etc. The free energies of the native packing patterns and topologies are separated by a gap from their alternatives. However, the free energies of the protein topologies from the same packing pattern are rather close.

INTRODUCTION

From known X-ray protein structures, it is now well established that proteins are very similarly organized, if one restricts the comparison of protein structures to the protein chain tertiary folds without reference to the exact geometry, chain length and amino acid sequence. In such a presentation the tertiary fold is fully determined by the sequence of rigid α- and (or) β-secondary structures, by their mutual position and orientation in space, and by the irregular pattern of their connections.[1-11] Detailed analysis and comparison of different primary and tertiary protein structures have shown that in general protein sequences adopt the same fold and that the folds fall into a limited set of families.[12-15] Although we still cannot say how many protein folds there are, some preliminary estimations show that natural proteins have likely descended from approximately 1000 different ancestors, and that the number of distinct protein folds is likely to be between 400 and 700.[13,16-19]

The limited set of protein folds led to the suggestion that the problem of protein structure prediction can be solved by accumulating sufficient structural data and deriving corresponding methods for sequence-structure recognition[18-36]. As a consequence known 3D structures are used to model their folds. The usual procedure can be represented in a few steps. Firstly, the known 3D protein structures from the PDB bank are filtered according to criteria such as: resolution, sequence homology and structural match. Secondly, on the bases of compatibility scoring functions and appropriate algorithms, the sequence of the tested protein molecule is aligned to each target fold. Lastly, scores derived from the sequence-structure alignments are ranked to find the most probable target fold. However, such a intuitively clear approach has its limitations. The reason is that there is no general mathematical prove for the existence of a simplified scoring function with native-like structures as global minima. Nevertheless, the searching for universal scoring functions have lead to the formulation of general physical rules, which govern the hierarchy of the organization of protein structures.[37] These rules impose the geometrical constraints with respect to the packing and to the mutual orientation of secondary structures, as well as the restrictions on the arrangement of irregular connections between the secondary structures. As a consequence the possible protein folds are limited.[6,38-41]

The objective of this study is to show that the problem of protein structure prediction can be reduced to find suitable compact packing patterns built by α - and (or) β-regular regions and the limited set of protein topologies associated with them, followed by the detailed atom-atom energy calculations to predict the atomic coordinates within this framework of the thermodynamically most favorable packing pattern. This approach is supported by the results of recent experimental and theoretical studies on the structural properties of the transition state for the protein folding of two-state small monomeric proteins. [42-46]

Thermodynamical experiments have shown that in most cases protein folding and unfolding involves only two thermodynamically distinct states or phases.[47] The folded state consists of a spatially narrow distribution of structures, whereas the unfolded state contains a much larger number of distinct configurations with broader spatial distribution. The state at the maximum in the free-energy profile is called the 'transition state' for protein folding. In general the double-well free energy profile suggests two relaxation times: one for the motion of a protein chain in the free energy minimum corresponds to the unfolded state and is very fast. The other relaxation time is determined by the free energy difference between the folded and unfolded states and is rate-limiting for the protein folding reaction. In the fast folding part there is a

reconfiguration via a collapsed 'burst-phase' intermediate, but this strongly depends on the conditions. Thus at higher temperatures (another way to control the presence of intermediate states is to lower the pH or to increase the concentration of denaturants) at which the entropy contribution is more pronounced, partly structured intermediates are favorable. In contrast, at lower temperatures, in which the energy contribution to the free energy barrier becomes dominant, low energy intermediates are disfavored. Finally, the cooperativity suggest that kinetically the protein folding transition follows a nucleation mechanism. The nucleus is the lowest of all the free energy barriers in the transition state ensemble of the conformations. It is represented by certain nonlocal native contacts, which lead to the formation of a critical fragment, after which the subsequent dynamics lead unidirectionally to the native state.

It is the aim of the present work to show that the high-energy nucleus is most likely to be formed by the (crude) characteristics of the overall native fold of the protein molecule. After the nucleation process is started, the protein propagates further, approximately isoenergetically, at transition-state level where local and long-range interactions are cooperatively consolidated through its overall fold. The competing folds are discriminated by an energy gap.[45]

Our strategy can be formulated as follows. Firstly, protein conformations at the TS are described using a lattice model which is based on the packing of α- and(or) β-secondary structures (packing patterns). Steric restrictions between the secondary structures are taken into account by averaging over the backbone coordinates of known 3D-protein structures. A set of packing patterns is constructed which has to be consistent with the length of the tested protein chains. The computational procedure is not restricted to the idealized geometric characteristics of the secondary structures but in this article we will use only such simplified structures. They are represented by hypothetical cylindrical surfaces on which C_α-atoms of the polypeptide backbone form a right handed spiral (fig.1). The side groups of amino acid residues are treated as spheres, the center of which represents the average displacement of the side groups from the C_α-atoms. The lines which connect the C_α-atoms and the center of the side groups are perpendicular to the axes of the spirals. Lattice conformations are described by the conformational freedom of the loops and by the fluctuations of the lengths and locations of the regular α- and (or) β-regions along the sequence and in the lattice. Secondly, to each packing pattern, as a consequence of thermodynamic restrictions, a limited set of topologies is associated. Thirdly, calculations are carried out to determine the free energy of the protein

over all available sets of packing patterns and topologies associated with them. Fourthly, the packing pattern and topology with the minimal free energy is expected to be the same or at least very similar to the native fold. As a final result one obtains the distribution and corresponding fluctuations of the secondary regions along the sequence and their contacts in space.

THEORY

Free energy of protein conformations at TS

Previously[45] it has been shown that the free energy of the TS transient conformations can be presented in the form:

$$E(Q) = \sum_i E_i(\vec{r}_i, n_i) + \sum_i E_{i,i+1}(\vec{r}_i, n_i; \vec{r}_{i+1}, n_{i+1}) + \frac{1}{2} \times \sum_i \sum_{\substack{j \\ i \neq j}} E_{i,j}(\vec{r}_i, n_i; \vec{r}_j, n_j) \qquad (1)$$

where $Q(\{\vec{r}_i\}, \{n_i\})$ determines the locations of the secondary structures along the sequence and in the lattice, as well as the topological connections between their N- and C-termini. E_i is the internal energy associated with each terminus (for example the sum of free energies of elongation for α- or β-secondary structures for the individual amino acid residues which contribute to the given termini). $E_{i,i+1}$-is the energy of interaction between nearest neighbors termini along the sequence (for example the bending energy of loop connections between the N- and C- terminus of adjacent secondary structures). $E_{i,j}$-is the non-local energy of interaction between residues which are far apart in the sequence but close in space- for example hydrophobic and hydrogen bonding interactions. Following equation (1) for each protein conformation in the TS we can ascribe a statistical weight:

$$\exp\left(-\frac{E^{\varphi}(Q)}{RT}\right)$$

The free energy of the TS can be obtained by taking the logarithm of the sum over all possible weights for the available conformational states in the TS. [45,49,50]

Calculation of free energy

The free energy of the protein fold can be calculated on the basis of minimization rules.[45,48-50] For our purpose, as a starting point for the minimization procedure, the most appropriate approach is the well known classical statistical mechanics Gibbs-Bogoliubov[51,52] inequality which is quite general and does not depend on the characteristics of the investigated system. It states that the free energy of the system of interest is less than, or equal to, the free energy of the system in the presence of an external field, which approximates part or all of its internal interactions. If we represent the free energy and the energy levels of the system with or without the presence of an external field Φ by F ,$\{E_i\}$ and F^{Φ},$\{E_i^{\Phi}\}$ correspondingly the Gibbs-Bogoliubov inequality takes the form:

$$F \le F^{\Phi} + \langle E - E^{\Phi} \rangle_{P^{\Phi}} \tag{3}$$

Therefore, the free energy of the system of interest can be obtained as a minimum of the right side of the inequality (3) over the external field:

$$F = \min_{E^{\Phi}} \{ F^{\Phi} + \langle E - E^{\Phi} \rangle_{P^{\Phi}} \}$$

where F^{Φ} and P^{Φ} are determined in (3). It is important to note that inequality (3) gives only the upper limit to the free energy. As a result, the choice of the adjustable parameters and the corresponding potential functions for the pairwise residue interactions is restricted by the requirement that the limit on the right of inequality (3) is as small as possible.

Lattice model of the packing patterns and protein lattice conformations.

Thermodynamically favorable packing patterns must fulfill the natural requirement of maximum screening of hydrophobic groups, minimum screening of hydrophilic ones and the maximum saturation of hydrogen backbones between β-regions.[37] Thus β-strands are grouped in sheets which form layer structures with other sheets or helices packed on their hydrophobic faces. When the twist of the β-strands and their curl is varied, the curl of the β-sheet is also varied and this makes it possible to pass from a sandwich or α/β structure to a barrel structure. α-Helices are either stacked around a central core, or form their own layered structures. Steric restrictions between structural blocks can be taken into account by averaging over the

backbone coordinates of known 3D-protein structures and by using some simple symmetrical considerations for the mutual positions of the structural regions.

Lattice conformations of a given protein chain are characterized by: [45] 1) type of packing pattern; 2) number of occupied segments in this pattern-k; 3) $N \to C$ direction of the chain relative to the direction of the occupied segment-τ_i; 4) position of N and C ends of secondary regions along the chain and along the segments of the packing pattern-(n,r). So each end i is characterized by: $n_i^{\min} \le n \le n_i^{\max}, r_i^{\min} \le r \le r_i^{\max}$ and $\tau_i = 1$ when the $N \to C$ direction of the chain and that of the segment, through which the chain, passes is the same. In the opposite case $\tau_i = -1$(fig.2).

$$\left(r_i^{\min}, r_i^{\max}\right) = \begin{cases} \left(r_1^p, r_2^p\right), & \tau_i = 1 \quad 'N' \\ \left(r_3^p, r_4^p\right), & \tau_i = 1 \quad 'C' \end{cases}, \quad \left(r_i^{\min}, r_i^{\max}\right) = \begin{cases} \left(r_3^p, r_4^p\right), & \tau_i = -1 \quad 'N' \\ \left(r_1^p, r_2^p\right), & \tau_i = -1 \quad 'C' \end{cases}$$

where p is the number of one of the packing segments and the following relations must be fulfilled:

$$1 \le p \le k,\ 1 \le i \le 2k,\ r_1^p \le r_2^p \le r_o^p \text{ and } r_o^p < r_3^p \le r_4^p.$$

To each end of the secondary regions we ascribe only a part of the chain with internal coordinates described by:

$$r_i \le r \le r_3^p - 1,\ \tau_i = 1 \quad 'N' \qquad r_3^p \le r \le r_i,\ \qquad \tau_i = -1 \quad 'N'$$

$$r_3^p \le r \le r_i,\ \qquad \tau_i = 1 \quad 'C' \qquad r_i \le r \le r_3^p - 1,\ \tau_i = -1 \quad 'C'$$

where for the distribution of $n = n_i - \tau_p \cdot (r_i - r)$ along the chain we have:

$$0 < \Delta n \le N - (2k-1) \cdot L$$

$$n_i^{\min} = (i-1) \cdot \left(\frac{N - \Delta n}{2k-1}\right) + 1$$

$$n_i^{\max} = n_i^{\min} + \Delta n$$

Here L is a parameter, the meaning of which is the minimal length of globular structural elements (α-or β-regions and loops).

Contribution of the closest residues along the chain

Recently, using site-directed mutagenisis, it was shown for the β-structural propensities that they vary depending on both the protein and the site of mutation in the protein.[53]

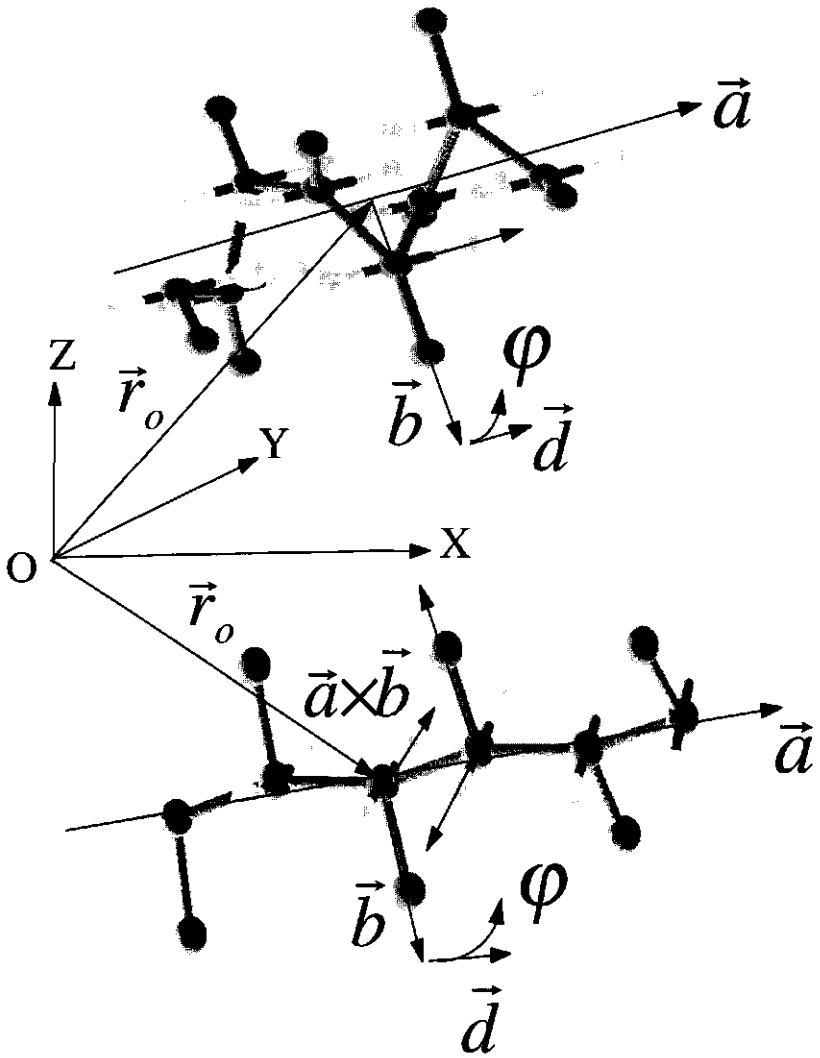


Fig.1 Geometrical representation of the idealized α- and β- secondary structures. β-Strands and α-helix are represented by the right-handed spiral, the main characteristics of which are: distance between the C_α-atoms along the helix axis-d; helix radius-R; rotation angle between two closest C_α-atoms along the helix axis-φ. For β-structures we have: d=3.3Å, R=0.25Å, φ=189.15°. For α-structures we have: d=1.5Å, R=2Å, φ=100°. The direction of the strand is marked by $\vec{a}$, and $\vec{b}$ is the direction of the side group placed at the central position $\vec{r}_0$ of the strand (see also in the text). The effective directions of $'C_\alpha \rightarrow H'$ and $'C_\alpha \rightarrow O'$ hydrogen bonds are also shown. For α-structures effective hydrogen bonds are directed toward the helix axis but it is important to note that they are not involved in intramolecular long-range interactions. Both for α- and β-structures the length of $'C_\alpha \rightarrow O'$ and $'C_\alpha \rightarrow H'$ is 2Å. The average displacement of side groups from C_α-atoms is 3Å.

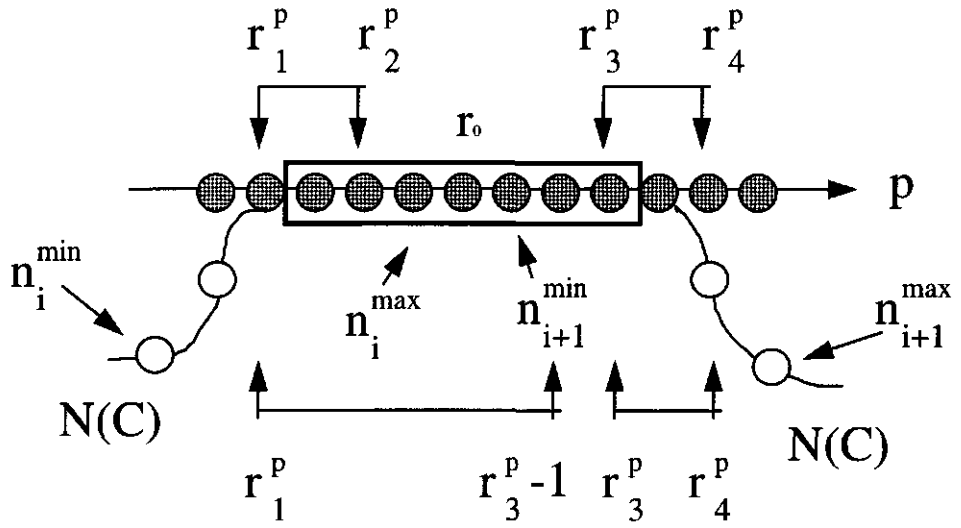


Fig.2 Schematic representation of allowed positions for 'N' and 'C' ends of structural segments in packing patterns. Each segment has 12 or 16 positions for C_α-atoms for β- or α-structure, respectively. The first and last position in each secondary structure are not taken into account they only mark the direction of entering and exiting of the chain, when it goes through some segment. The boundaries marked at the top of fig.2 represent the domains of vibration and the boundaries marked at the bottom represent the separation of secondary structures in two parts which are ascribed to 'N' or 'C' ends. These regions participate in energy interactions between the ends of structural regions.

Also, as a consequence of the comparison of helix propensities derived from small model peptides (which are mainly determined by the local interactions) and those derived from a protein, it was shown that there is a link between variation in helix propensity among amino acids and helix-tertiary interactions.[54] To avoid these difficulties we make the following approximations: firstly, in the unfolded state of the protein chain, which is taken as a zero free energy level, each amino acid residue freely explores all its conformational space. In other words, amino acid residues along the chain are considered as energetically uncoupled; secondly, the loop regions between the structural blocks in the frame of the packing patterns are considered as random-coiled like parts of the protein chain with uncoupled residues along it; thirdly, the free energy of the residue secondary structure propensities does not depend on their position along the structural regions and the location of the structural regions along the protein chain. This follows from the assumption that the tertiary interactions are much stronger determinants for residue preferences to be in a α- or β-state than the short-range interactions between residues inside the α- or β-structural regions. Further, in equilibrium the structural

regions in the frame of the packing patterns can be seen as quasi-independent systems separated by regions with decoupling residues which destroy the linear memory of chain conformations.

As a consequence of the above approximations the free energy of the structural regions is described in the form:

$$E_{i-1,i} = \Delta F_{i-1}^{N}\left(a_{n_{i-1}}\right) + E_{i-1,i}\left(\vec{r}_{i-1}, n_{i-1}; \vec{r}_i, n_i\right) + \Delta F_i^{C}\left(a_{n_i}\right) = E_{i-1}\left(\vec{r}_{i-1}, n_{i-1}\right) + E_i\left(\vec{r}_i, n_i\right)$$

$$E_{i-1}\left(\vec{r}_{i-1}, n_{i-1}\right) = \sum_{n=1}^{n_{i-1}-1} \Delta F_n^{coil-regular}\left(a_n\right) + \Delta F_{i-1}^{N}\left(a_{n_{i-1}}\right)$$

$$E_i\left(\vec{r}_i, n_i\right) = \sum_{n=1}^{n_i} \Delta F_n^{coil-regular}\left(a_n\right) + \Delta F_i^{C}\left(a_{n_i}\right)$$

where $\Delta F_n^{coil-regular}\left(a_n\right)$ is the free energy of elongation; a_n stands for one of the natural amino acids and n is its position along the protein chain; $\Delta F_{i-1}^{N}\left(a_{n_{i-1}}\right)$ and $\Delta F_i^{C}\left(a_{n_i}\right)$ are the free energy of initiation and termination, respectively. For the free energy of α- and β-structural propensities we have used the experimental values determined by O'Neil and DeGrado[55] and Milnor and Kim.[56,57] The free energy of initiation and termination are determined mainly by the fixation of the loop regions between the corresponding structural regions along the chain and are discussed in detail below. For α-helical regions the initiation and termination of free energy contributions include also the fixation of residues at the N- and C-termini of the helical regions itself. In the most simple case, when the residue specificity is not taken into account, the value of 4kcal/mol is taken from the experimental work of Platzer et al..[58] The only exception is the Pro residue for which a value $-RT \approx -0.6$ kcal/mol is added when it occupies the first (N-terminal) positions of α- and β-structural regions.

As a first approximation the contribution from the loop regions is separated into two main parts. In the first part residues are evaluated by their tendency to be exposed to solvent. The loop term has a form:

$$E_i\left(n_i, r_i\right) = -\sum_{n=1}^{n_i} \Delta F_{n,i}^{loop}\left(a_n\right)$$

$$E_{i+1}\left(n_{i+1}, r_{i+1}\right) = \sum_{n=1}^{n_{i+1}-1} \Delta F_{n,i+1}^{loop}\left(a_n\right)$$

where $\Delta F_{n,i}^{loop} = -RT\ln\left(1 + w_i \exp\left(-\frac{\Delta F_n^{hphob}}{RT}\right)\right)$ represent the loop propensities for the residue groups[36] and ΔF_i^{hphob} is the difference in hydrophobic free energy when a polar residue is transferred to a nonpolar medium.[59] Only interlayer connections are taken into account for which $w_i = 0.1$; otherwise $w_i = 0$. This approximation follows from the requirement that the hydrophobic core of protein molecules has to be well isolated from the water environment. The approximation is also consistent with the distribution of hydrophilic residues along the chain which are mainly concentrated in the loop regions.

The second loop term is connected with the free energy of loop bending. This term approximates the expression for the free energy of bend fluctuations of long polymer molecules:[60]

$$\Delta F^{bend} \approx \frac{RT\sigma}{2L} \times (\Delta\theta)^2$$

where σ is the persistent length, L is the chain length, R is the gas constant, T is the temperature and $\Delta\theta$ is the total bending angle between the ends of the chain. The rigidity of the loop regions imposes restrictions on the chain pathway. For example, antiparallel association between two β-strands closest along the chain is more favorable than a parallel ones.[37] For α-helices the rigidity of the loop regions does not impose such strong restrictions. This follows from the fact that in a typical protein chain the α-helical regions are usually directed perpendicular to the helix axis. The most simple way to take into account both α- and β-structural regions is to approximate the total bending angle between the ends of the loop regions by the bending angles $\theta_i(\vec{r}_i)$ and $\theta_{i+1}(\vec{r}_{i+1})$ at the ends of the structural regions $\vec{r}_i$ and $\vec{r}_{i+1}$ and the shortest possible loop between them-$|\vec{r}_{i+1} - \vec{r}_i|$. The length of the loop can be taken into account by the normalization constant. As a consequence, the free energy ΔF^{bend} of loop bending between two closest structural regions along the chain is applied in the form:

$$E_{i,i+1}(\vec{r}_i, n_i; \vec{r}_{i+1}, n_{i+1}) = A \cdot \frac{\left(\theta_i^2(\vec{r}_i) + \theta_{i+1}^2(\vec{r}_{i+1})\right)}{2\left(\frac{\pi}{2}\right)^2}$$

where $\vec{r}_i$ and $\vec{r}_{i+1}$ are the spatial coordinates of the terminal residues; A is the free energy for the total bending angle π which corresponds to an antiparallel association of two β-strands closest along the chain. When the loop is attached to an α-helical region the protein chain

does not change its direction relative to the shortest loop and the corresponding bending angles are closest to zero. We distinguish two main contributions to the free energy of chain bending: firstly, the bending of irregular chain portions between closest structural regions along the chain.[37] In this case, the normalization constant A is taken to be ≅3kcal/mol; secondly, the bending of structural regions itself. This term is mainly connected with the β-strands and gives the possibility to account for the free energy of a β-sheet curl. In this case A is taken to be ≅8 kcal/mol.

Free energy of long-range interactions

1. Hydrogen bonding interactions

The main role of hydrogen bonding interactions in long-range interactions is stabilization of β-strands arranged in parallel, antiparallel or mixed β-sheets in the protein globule. For an α-helix the majority of hydrogen bonds are already saturated and they cannot be involved in these sheets and should form their own layers (or clusters). The stability of β-sheets and their dimension (lengths and number of β-strands) is mainly determined by the sum of free energies of individual residues during the β-structure-coil transition, as well as by the loss of free energy of the bend formation and by the excess free energy at their edges. The free energy of bends or loops is already described above. So each amino acid residue involved in a β-sheet structure can be characterized by two states, when it is placed in the middle or at the edge of a β-sheet. A residue in the middle of a β-sheet is stabilized by hydrogen bonds on both sides. From one side the residue is stabilized by its own hydrogen bonds, from the other side it is stabilized by the hydrogen bonds of previous and subsequent residues along the chain. At the edges β-sheets residues are stabilized only from one side by their own hydrogen bonds or by the hydrogen bonds of their neighbours. To account for these two states it is appropriate to involve two points $'O'$ and $'H'$ which are placed on both sides of each C_α-atom (fig.3). These points must lie on the plane of the β-sheet. The expression for the hydrogen bonding interactions has the form:

$$\varepsilon_{ij}^{hydr} = q_i^{hydr} \times f_{ij}\left(\vec{r}_i, \vec{r}_j\right) \times q_j^{hydr}$$

where the hydrogen bonding charges are: $q_i^{hydr} = -\Delta F_i^{edge}$ and $q_j^{hydr} = 1$. The geometrical factor f_{ij} is determined by $\vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right)$, where $R = 4\,\text{Å}$. It is seen that hydrogen bonding interactions are not symmetrical: $\varepsilon_{ij} \neq \varepsilon_{ji}$.

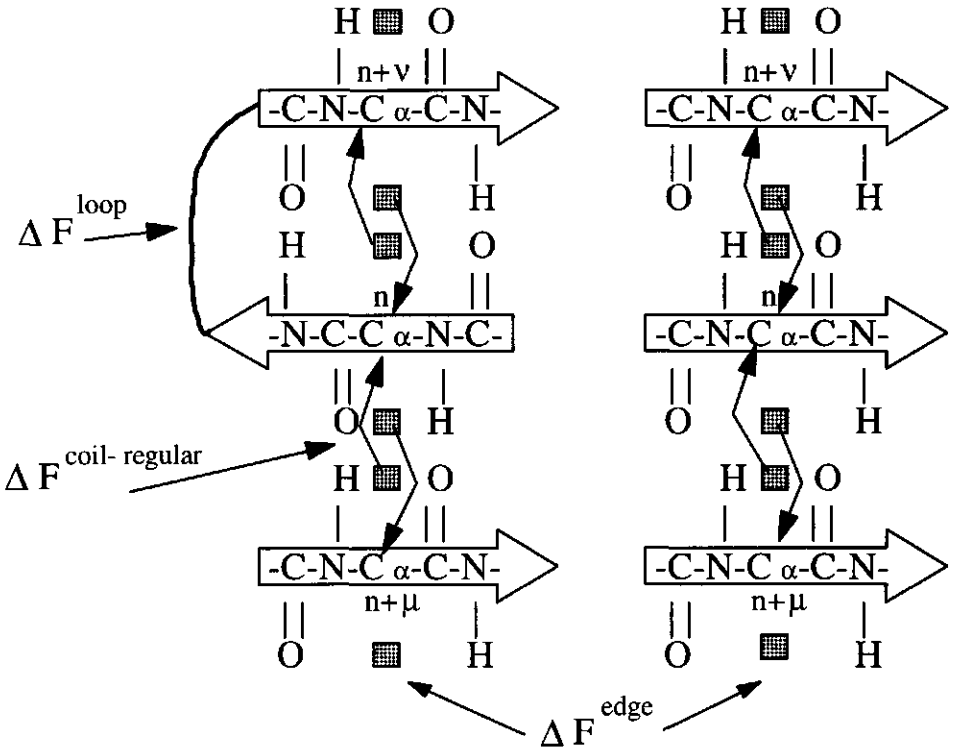


Fig.3 Representation of a model for long-range hydrogen-bonding interactions inside β-sheets. Effective hydrogen bonds are formed between $H^n \rightarrow C_\alpha^{n+\nu}$ and $O^n \rightarrow C_\alpha^{n+\mu}$ atoms. $C_\alpha^{n+\nu}$- and $C_\alpha^{n+\mu}$-atoms are located on both sides of C_α^n-atom, ν and μ are their separation along the chain relative to the C_α^n-atom. To each effective hydrogen bond we ascribe an energy term $\Delta F_n^{hd} = -\Delta F_n^{edge}$, where $\Delta F_n^{edge} > 0$ is an additional free-energy of a residue n at a β-sheet edge.[57] For the intrinsic free energy of residues, in order to compensate the additional edge terms, we put $\Delta F_\beta^{internal} = \Delta F_\beta^{coil-regular} + 2 \times \Delta F_\beta^{edge}$, where $\Delta F_\beta^{coil-regular}$ is free-energy difference between the interior of β-sheet and coil.[56]

2. Hydrophobic interactions

α-Helix-α-helix, α-helix-β-sheet or β-sheet-β-sheet contacts are stabilized mainly by hydrophobic interactions. Using the local density as defined by the residue volume distribution around each position in the lattice,[48] we represent the hydrophobic free energy of the individual residues as:

$$\varepsilon_i^{hphob} = \Delta F_i^{hphob} \times \rho(\vec{r}_i),\ \rho(\vec{r}_i) = \sum_j \frac{v_j}{V} \times \vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right) \times f\left(\left|\vec{r}_i - \vec{r}_j\right|, R\right)$$

where ΔF_i^{hphob} is the difference in hydrophobic free energy when a polar residue is transfered to a nonpolar medium and $\rho_i(\vec{r}_i)$ is the local density.

Therefore, for the pairwise interactions, it follows:

$$\varepsilon_{ij}^{hphob} = q_i^{hphob} \times f_{ij}(\vec{r}_i, \vec{r}_j) \times q_j^{hphob}$$

where the hydrophobic charges are $q_i^{hphob} = \Delta F_i^{hphob}$ and $q_j^{hphob} = \frac{v_j}{V}$; v_j is the side chain volume of residue j along the chain and V is the volume of the interacting sphere $V = \int 4\pi\Delta\vec{r}^2 \cdot f(\Delta\vec{r}, R)d\Delta\vec{r}$ for each position along the packing segments.[48] The geometrical factor f_{ij} is determined by $\vartheta\left(R - \left|\vec{r}_i - \vec{r}_j\right|\right) \times f\left(\left|\vec{r}_i - \vec{r}_j\right|, R\right)$. The radius R of the interacting sphere is 8 Å. As in the case of hydrogen bonding interactions, hydrophobic interactions are also asymmetrical: $\varepsilon_{ij} \neq \varepsilon_{ji}$.

RESULTS AND DISCUSSIONS

In order to simplify the test of the theory developed in the present paper, we restrict ourselves to small two-state monomeric proteins with less than 80 residues which nevertheless cover all basic protein architectures: engrailed homeodomain (1hdd.pdb file) -a 3α-helix protein; alpha-*amylase inhibitor HOE-467*A (1hoe.pdb file) -a 6 stranded β-sandwich; lz(sash)*112 50 S ribosomal protein (1ctf.pdb file) -an α/β-sandwich with 3α-helixes and 3β-strands; protein G (B1 domain, 2gb1.pdb file) -an α/β-sandwich with 1-helix and 4β-strands; major cold shock protein (1csp.pdb file) -a 6-stranded β-barrel.

A key question is to what extent, from a thermodynamic point of view, the protein architectures are determined in the TS on the rough level of the overall chain fold, through the secondary structure packing and loop connection constraints. Our previous results have shown that at its lowest free energy minimum of the TS, the protein molecule propagates toward its native state approximately isoenergetically through nucleation-growth mechanism. [45] The nucleus is characterized by a native like-topology, by approximately correctly formed secondary structures and by loop regions with different degrees of order. In this study, the fold recognition strategy is a consequence of the fact that the native-like nucleus is separated from all other folding alternatives by a high free energy barrier.

As a first approximation, we do not take into account the internal degrees of freedom for the secondary structures including main chain as well as side chain bonds of residue groups. Hence localization of the secondary structures in the packing patterns is controlled by two main factors: optimal orientation of residue side groups towards or apart from the core of the packing patterns with respect to their hydrophobic or hydrophilic character, and maximum saturation of hydrogen bonds between the β-strands. As a consequence, the secondary structures of the packing patterns can only be moved and rotated as a whole or twist and curl for the β-strands. In this paper, we do not take into account the steric restrictions coming from the packing of residues with different volumes of their side chain groups in the core of the packing patterns. It is well known[61] that high packing densities are readily attainable among clusters of the naturally occurring hydrophobic amino acids. Therefore, in this study we address the question to what extent in the TS the protein sequences can select their native fold from competing folds with high core hydrophobicity but different packing organization. Geometrical characteristics for the packing patterns are taken from the statistical observation of known or from theoretical considerations of simplified idealized protein structures. In fact the packing analysis of α-helices, β-sheets and α-helices with β-sheets has shown that this packing is determined mainly by the properties of the polypeptide backbone.[8,9,37] In α structures α-helices are packed at the angles of $-50°$ and $+20°$; in α/β or $\alpha+\beta$ structures α-helices are usually packed parallel to β-strands and the right-handed twist of a β-sheet leads to an angle of $40°$ between two α-helices neighboring in space; in the aligned class of β structures, where β-sheets pack face to face, the mean twist angle is $17°$, while in the

α/β structure this angle is 19° and 32° in the barrel structure. The packing patterns of protein molecules on which the theory presented here is tested are shown in fig.4.

We have examined all possible topologies for the packing patterns of the tested protein molecules. The usual consideration is based on the fact that for a given packing pattern the full set of available topologies is represented by $2^M \cdot M!$, where M is the number of the secondary structures of the packing pattern.[6] As a consequence of the assumption that the connections between the secondary structures neither cross each other, make knots or pass across the outside of the sheets or helix layers, the number of allowable protein topologies for each packing pattern is reduced to the sets shown in fig.5. Connections between the secondary structures which pass across the outside of the sheets or helix layers are not taken into account because they would essentially create additional layers.

The free energy of the protein at its lowest free energy minimum at the TS is obtained by an iteration process. [45,48] Calculations begin within a given topology and packing pattern. The protein topology is determined by: 1) the distribution of N- and C-terminus of the secondary structures along the sequence and in the packing pattern $n_i^{\min} \le n \le n_i^{\max}, r_i^{\min} \le r \le r_i^{\max}$ (see fig.2); 2) a secondary structure type associated with each terminus along the sequence; and 3) the N$\rightarrow$C direction of the sequence relative to the direction of the secondary structures in the packing pattern. Starting conditions include one of the following possibilities: 1) some arbitrary fixed protein conformation; 2) distribution of mean charges $\left\{q_i^m(\vec{r})\right\}$, where m marks the type of charges; 3) external fields $\left\{\Phi_i^m\left(\vec{r}\right)\right\}$; and 4) probabilities $P_i\left(n_i, \vec{r}_i\right)$ for the distribution of the terminus of the secondary structures along the sequence and in the packing pattern. The free energy of the protein topology is minimized with reference to the protein fluctuations in the TS. The results concerning the distribution of protein topologies, with respect to the free energy over the whole available set of topologies and their corresponding packing patterns, is shown in fig.6. The results show that hydrophobic and hydrogen bonding interactions are accurate enough, as a first approximation, for the prediction of the rough protein fold of the folding nucleus. However, there are exceptions. One small protein, for which the lowest free energy of the nucleus was not attained at its own native packing pattern and native topology, is HOE.

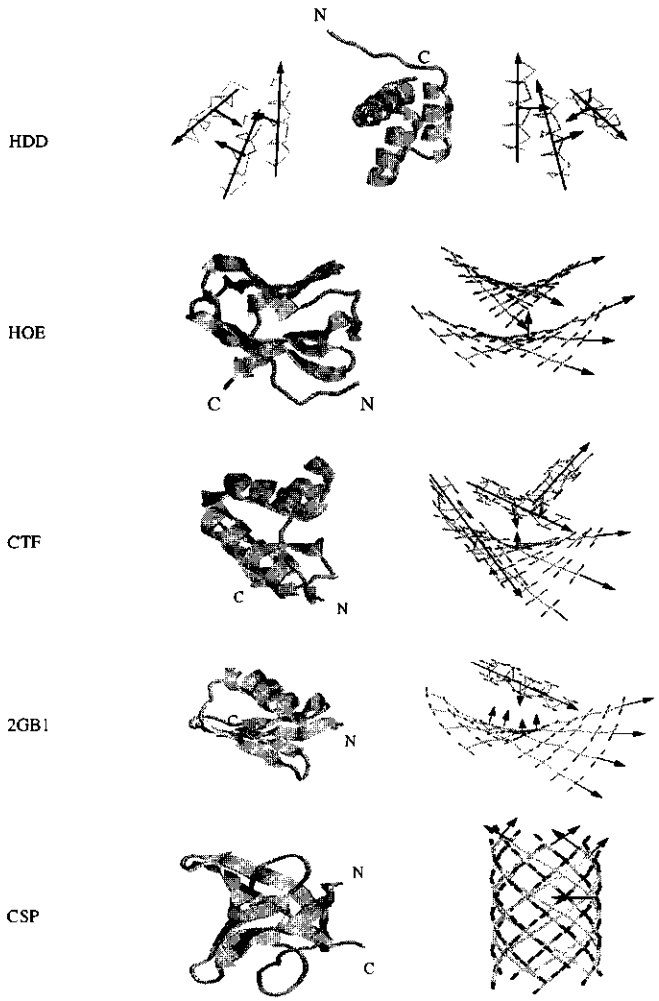
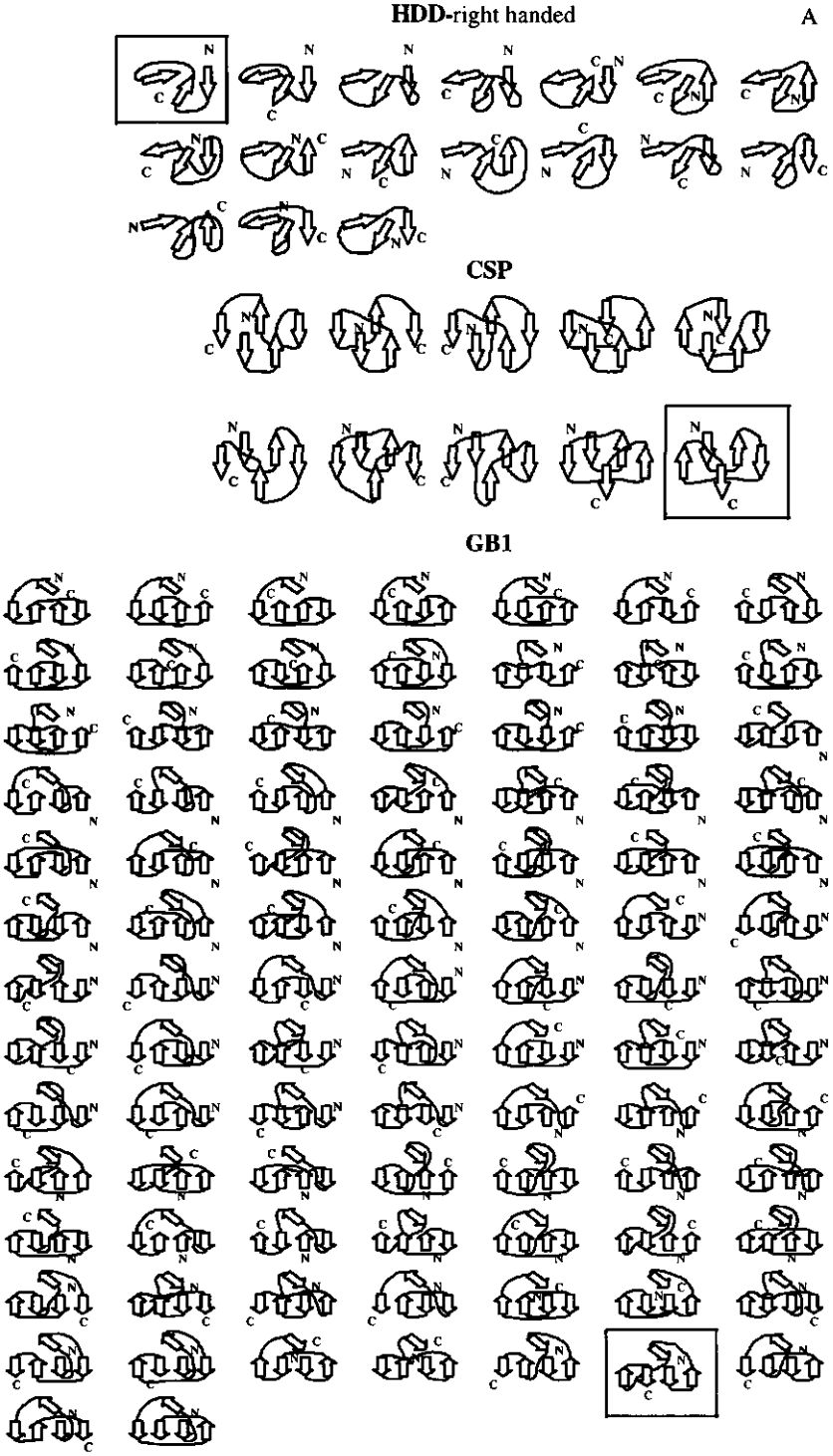


Fig.4 Representation of lattice models for several protein architectures. The distance between β-strands in the central region of each β-sheet is 4.7Å. The twist of the β-strand $\varphi = 189.15^{o}$ gives rise to the right twist of β-sheets with an angle of 19° between the closest β-strands in the sheet. The β-sheets in their turn form an angle of 30° between each other. The distance between the central regions of β-sheets is 10Å. The distance between the α-helices and the β-sheet is also 10Å. The arrangement of α-helix in the HDD molecule is obtained after averaging over its X-ray structure. The geometrical characteristics for the β-barrel are taken from the paper of Murzin et al.[8,9] The direction of central side groups in all packing patterns is toward the center of their cores. The effective directions of $'C_{\alpha} \rightarrow H'$ and $'C_{\alpha} \rightarrow O'$ hydrogen bonds, when the $'\mathrm{N} \rightarrow \mathrm{C}'$ direction of the chain is the same as in the β-segment through which the chain passes, are $[\vec{b} \times \vec{a}]$ and -$[\vec{b} \times \vec{a}]$, respectively. When the chain direction is opposite to that of the β-strand, we have for the direction of $'C_{\alpha} \rightarrow H'$ -$[\vec{b} \times \vec{a}]$ and for $'C_{\alpha} \rightarrow O'$ $[\vec{b} \times \vec{a}]$.


HDD-right handed
A
CSP
GB1

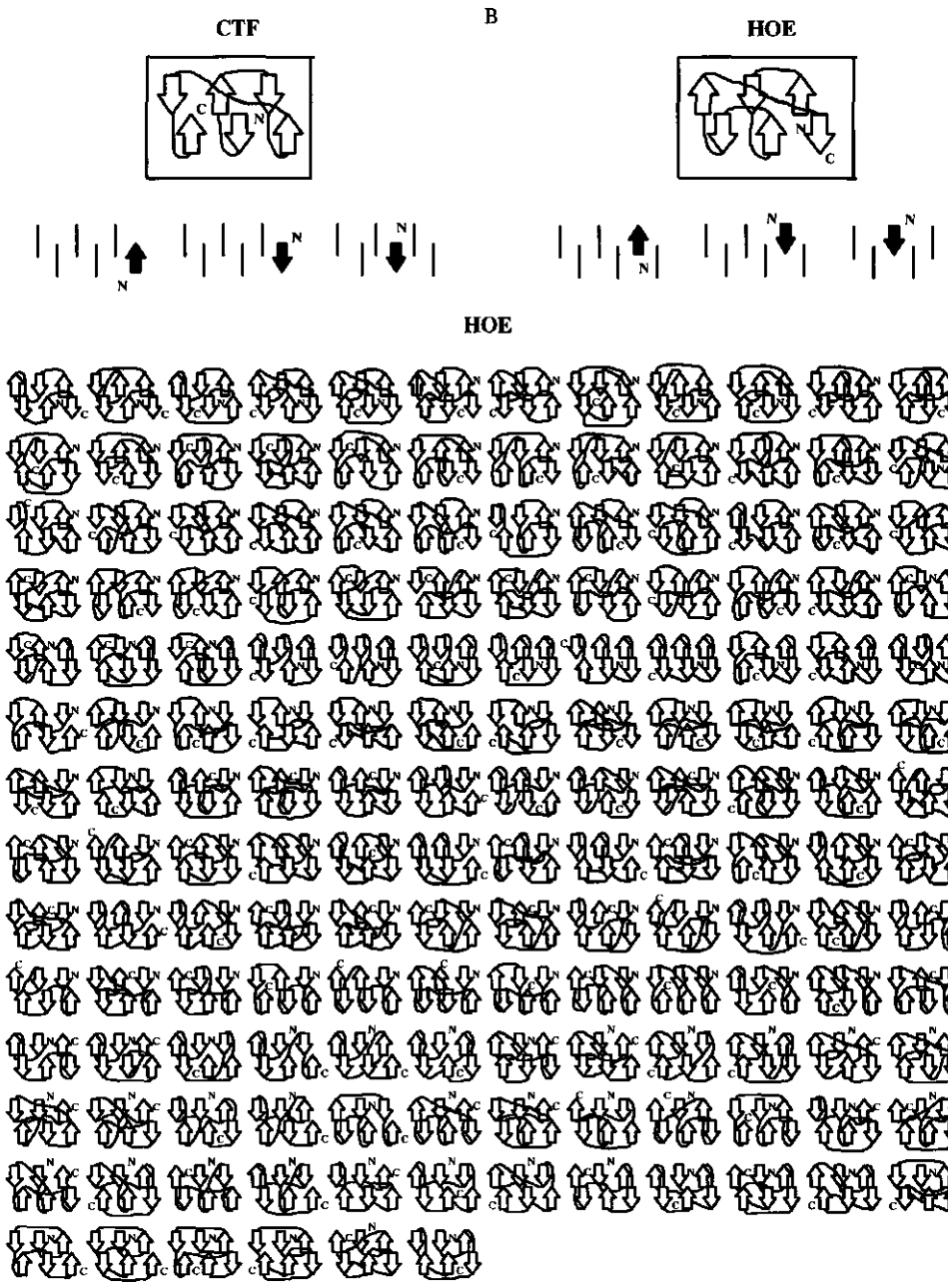


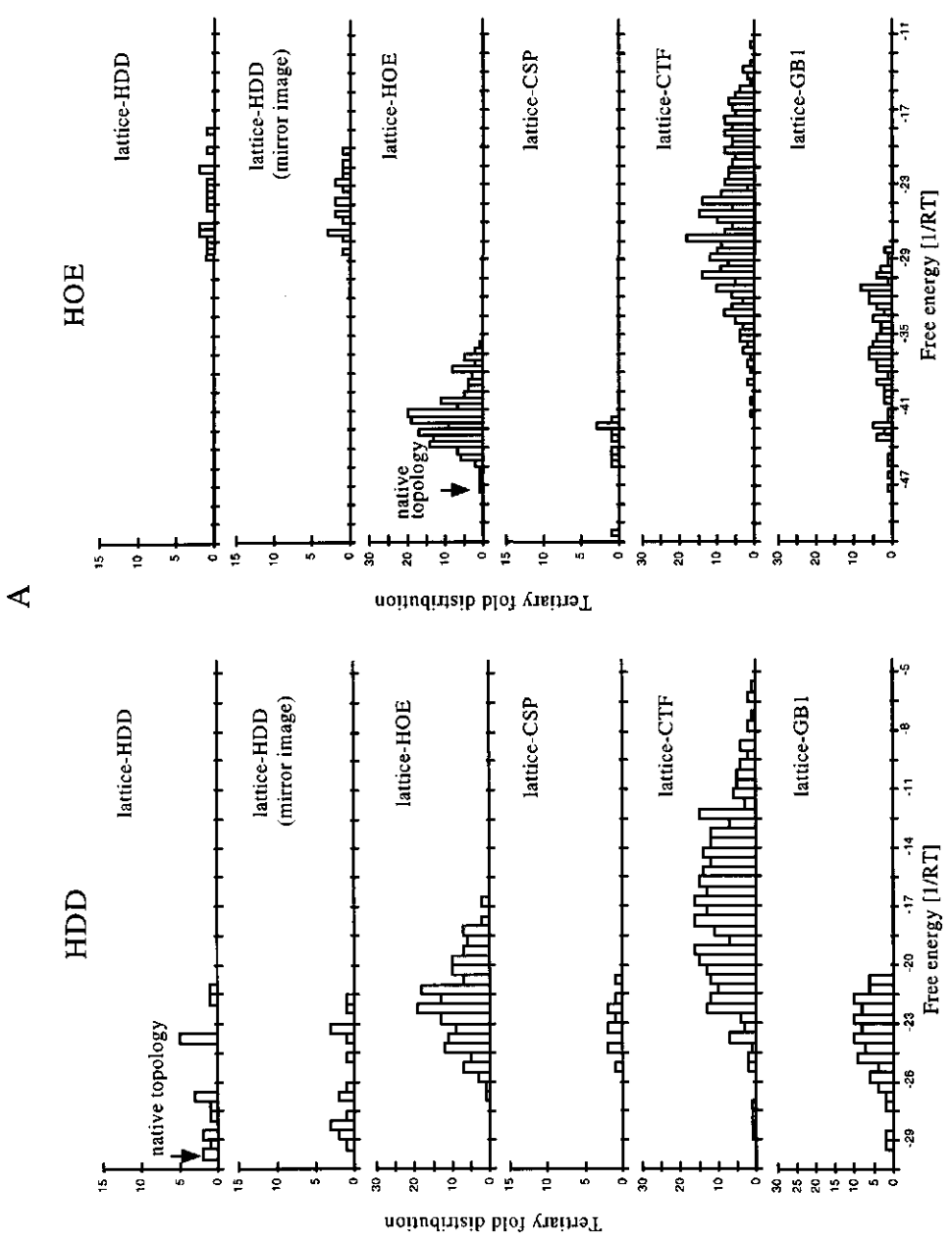
Fig.5 Most favorable set of tertiary folds for each packing pattern. Fig.5A presents the available folds for HDD, CSP and GB1 protein molecules. For HDD, in the case of left-handed α-helix bundle, there is an additional set of folds which are not shown because they are mirror-images of those shown. Fig.5B include the available folds for HOE protein. Together with their mirror-images, indicated by the position and orientation of the N-terminus of the chain, they form the full set for CTF molecule. For each protein molecule and its packing pattern the coresponding native fold is presented.

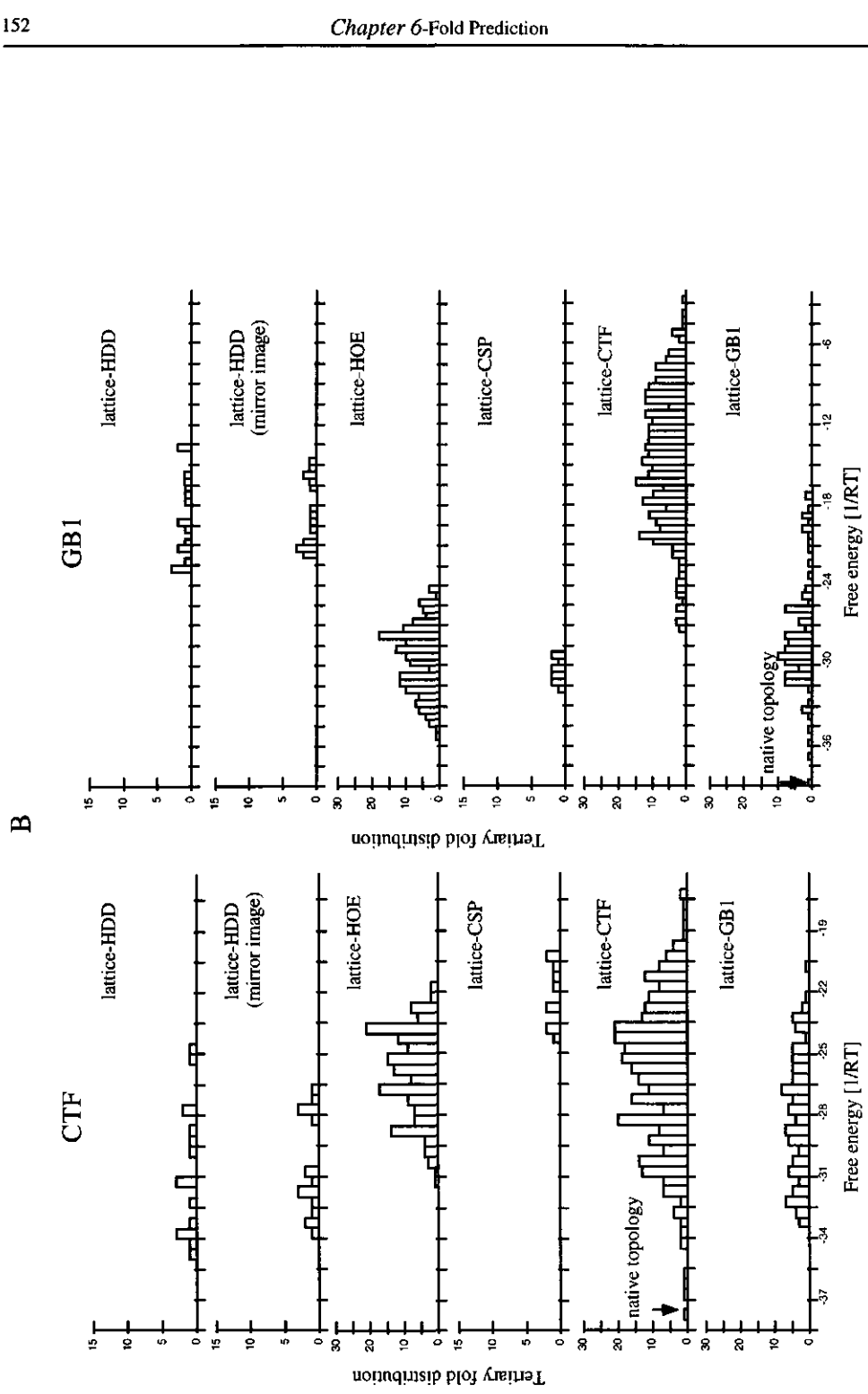
The folding nucleus for this protein molecule prefers the packing pattern of the barrel structure, rather then its own aligned packed β-sandwich structure. In the barrel structure, because of the absence of sheet edge effects, the concentration of the hydrophobic residues with bulky side chains is mainly directed towards the core of the barrel. Calculations show that the dimension of nucleus core, with a barrel packing pattern in terms of occupied core β-strand positions, is the same for all protein sequences- 5 positions on each β-strand. As a consequence the protein sequences differ from each other by the density they produce in the nucleus core. HOE is a protein molecule in which the volume of the hydrophobic nucleus core is very large and approximately one additional alanine residue greater than in the case of the CSP molecule. The other protein sequences, compared with that of the CSP molecule, cannot achieve efficient internal packing when organized into a nucleus with a barrel packing pattern.

This result, together with the experimental fact that the globular proteins are characterized with high packing densities, confirm the already existing understanding that protein conformation is linked tightly to internal packing. Thus, depending on the protein sequence, some folds (which have high core hydrophobicity because of bulky hydrophobic groups) may not be realized because of packing constraints.

We have also analyzed the iteration pathway to the equilibrium state in the frame of each packing pattern and its corresponding topologies. The calculations were carried out from random starting conditions. Results show that the protein quickly reaches a compact structure in the packing pattern and after that slowly rearranges until most of the hydrophobic regions are incorporated in the core of the compact structure already formed. The equilibrium state does not depend on the starting conditions. It is mainly due to the fact that there are no restrictions caused by volume overlap of residue side chains and as a consequence the average environment of amino acid residues in the hydrophobic core of the nucleus is uniform.

The requirement of maximum screening of hydrophobic groups, of minimum screening of hydrophilic ones and of maximum saturation of hydrogen backbones, imposes restrictions on the protein sequences and on the formation and localization of the secondary structures in the packing patterns of the nucleus.


A
HDD
HOE
lattice-HDD
lattice-HDD (mirror image)
lattice-HOE
lattice-CSP
lattice-CTF
lattice-GB1
native topology
Tertiary fold distribution
Free energy [1/RT]


B
CTF
GB1
lattice-HDD
lattice-HDD (mirror image)
lattice-HOE
lattice-CSP
lattice-CTF
lattice-GB1
native topology
Tertiary fold distribution
Free energy [1/RT]

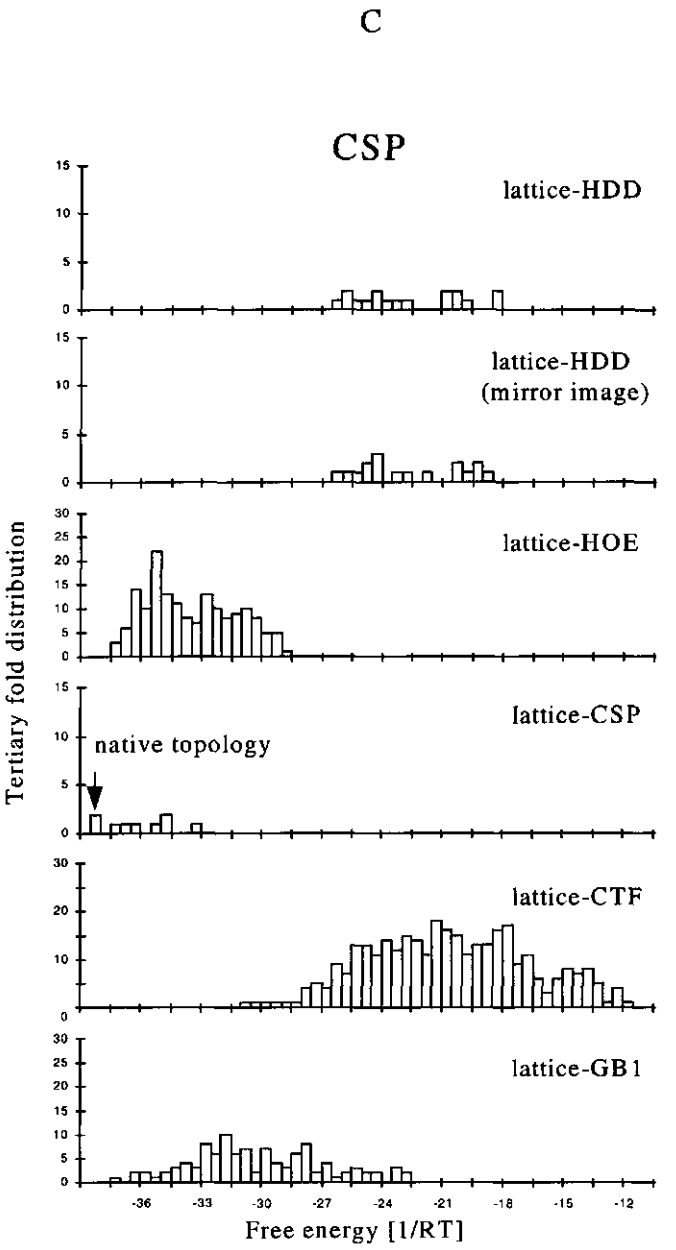


Fig.6 Representation of the distribution of tertiary folds for each protein sequence over the full set of tertiary folds and packing patterns. The free energy is presented in RT units ($T = 300^{\circ}$ K). Black arows show the positions of the native tertiary folds for the corresponding protein sequence when it is inbedded in its own native packing pattern. Each bar represent a group of tertiary folds for which the free energy difference between the coresponding folds is less or equal to 0.5 RT units.

For example, the α- or β-regions must have at least one hydrophobic surface through which they can interact with each other. In our approach such kind of restrictions are automatically taken into account. Depending on the topology and on the packing pattern of the nucleus, α- and β-secondary structures along the sequence are immersed into different potential holes (in the present paper we analyze mainly the hydrophobic and hydrogen bonding potentials). The size of the potential holes is proportional to the average size of the nucleus core around which the secondary structures are stacked. This is illustrated on fig.7B and C, where in the case of the CTF protein, the distribution of hydrophobic and hydrogen edge charges in the native nucleus of this molecule are shown. An important feature of the results presented in fig.7A is the high accuracy with which the native secondary regions along the sequence are localized in the packing pattern of the nucleus. One has to take into account that the calculations are carried out over the idealized structures, and in some cases the deviation of the native structures from the idealized is too high. For example, the average range of the angle between the β-strands and the helix axis projected onto the β-sheet or onto a plane normal to it is $-20°$ and $+10°$, or 0 and $+20°$, respectively.[37] The explanation is that the mutual orientation of the secondary structures modify only the shape of the core region, but not its average size and structure. In order to maximize the hydrophobicity of the core, the secondary structures have to be packed against each other with approximately the same sequence regions. When there is a deviation between the mutual orientation of the secondary structures in some parts of the idealized and native protein structures, this is achieved by the shift of the corresponding sequence regions relative to the secondary structures in the packing patterns. In this way the direction of the hydrophobic groups toward the core of the packing patterns is maintained. In the case of β-strands the minimal shift is two positions and in the case of α-helix the shift is one helix turn. This is exactly what one can observe in fig.7A.

Regardless of the topology and the packing pattern of the nucleus approximately the same sequence regions are involved in the α- and β-secondary structures. On the other hand, this ensures approximately the same total surface screen from water.

Taking also into account the dependence of the overall free energy of the nucleus with respect to the relative orientation of secondary structures by means of loop bending, calculations show that the fold stability of the nucleus is controlled by the following rules:

a) β-fold- Sequence regions enriched by strong hydrophobic residues are located inside the β-sheets, while weak ones are located at the edges. Amino acid residues with a small free energy loss compared with the positions located inside the β-sheets are also located at the β-sheet edges(fig.7B,C).

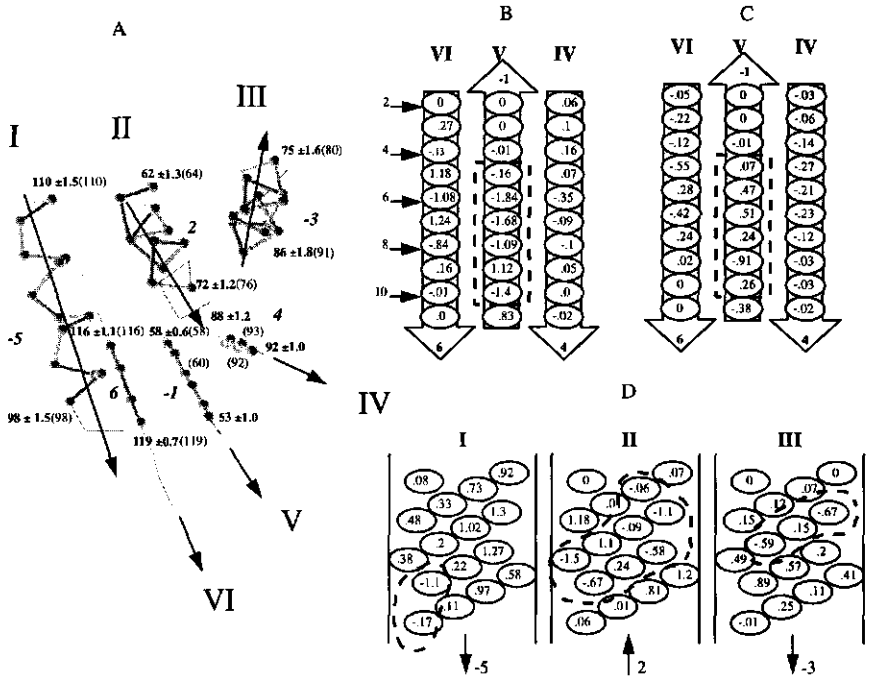


Fig.7 Representation of the predicted tertiary fold of the CTF protein. In fig.7A arrows show the direction of segments of the packing patterns. The Roman numbers mark the number of pattern segments, the Arabic numbers accompanied with sign show the position of the corresponding chain regions along the chain and their direction relative to the pattern segments through which they pass. The calculated sequences and fluctuations of the structural regions along the chain are marked by bold Arabic numbers. In parenthesis are shown the corresponding X-ray determined structural regions. Fig.7B,C,D show the distribution of the average hydrophobic (fig.7B for β and fig.7D for α pattern segments) and edge hydrogen binding charges (fig.7C for β pattern segments). In fig.7B β-strand positions, which are oriented toward the core of the packing pattern, are marked by arrows. It is seen from the figure that for the packing segments, which are situated in the middle of the α - or β - layers, the core positions as marked by dashed lines, are occupied by chain regions enriched by strongly hydrophobic residues. At the same time the β-packing segment are characterized by high positive edge free energies.

Taking into account that the hydrophobic core of the nucleus has to be well isolated from the water environment, the free energy of the nucleus has to decrease in the presence of long

interlayer connections. This is also consistent with the distribution of hydrophilic residues along the sequence which are mainly concentrated in the loop regions.

b) α-fold- The choice between the left-twist or right-twist of the α-helical regions depends on the slope between the hydrophobic paths on the surface of the α-helices and their axes. In addition the dependence of the overall free energy of the nucleus on the relative orientation of α-helical regions by means of loop bending-fig.7D has to be taken into account. The free energy of the nucleus topology at a given twist of the α-helix bundle is mainly determined by the free energy minimum of the loop bending and the maximum screening of the hydrophobic residues at the last turns of the α-helices.

c) α/β-sandwich fold- The nucleus topology is mainly determined by the competition between the maximum saturation of hydrogen backbones inside the β-sheets or their splitting into α-helical regions and a gain in decreasing the free energies of loop bending. Also the primary structures of the α-helical regions must be enriched by amino acids with a higher helix-forming propensities, but the localization of the helical regions in the packing patterns of the nucleus is determined mainly by the ability of their hydrophobic groups to be built into the core of the packing patterns.

d) β-barrel-fold- In a nucleus with a barrel packing pattern, instead of free energy lost at the edges of the β-sheets, there is a loss of free energy on the curl of the sheets. This is because the edges of the sheets are hydrogen bonded. The free energy of loop bending does not change considerably from topology to topology. Instead the main topology restriction comes from the fact that hydrophobic groups are mainly concentrated on one of the sheet sides, i.e. the one which forms the core of the barrel.

CONCLUSIONS

At the TS level all possible protein conformations can be split out into different ensembles. The crude characteristics of these ensembles are described by the limited set of thermodynamically most favorable folds. At the lowest free energy minimum of the TS the protein molecule propagates toward its native state approximately isoenergetically through a nucleation-growth mechanism. The nucleus is characterized by native like-topology, by approximately correctly formed secondary structures and by loop regions with different degrees of order. The main contributions, which stabilize the native-like fold of the nucleus in

the TS, are the long-range hydrophobic and backbone hydrogen bonding interactions, as well as free energy of loop bending and free energies of secondary structures formation. The selection of the protein fold is mainly determined by the most general characteristics coded in the protein sequence such as: the distribution of hydrophobic and hydrophilic residues along the chain, the ability of amino acids to form different secondary structures in compact chain conformations by hydrogen bonding with their spatial or protein chain neighbours and by the general rules which govern the packing of the secondary structures. In the TS the free energies of the native folds are separated by 'gaps' from all other folding alternatives. However, the difference between the free energies of the topologies from the same packing pattern are rather small. Also the accurate prediction of the native fold is not possible without taking into account steric restrictions arising from the packing of residues with different volumes of their side chain groups. The corresponding method for taking into account the steric interactions between side chains of residue groups on the basis of the self-consistent field approach will be presented shortly.

REFERENCES

1. Chothia, C & Michael, L. Structural patterns in globular proteins. *Nature* **261**, 552-558 (1976).
2. Chothia, C., Levit, M. & Richardson, D. Structure of proteins: packing of α-helices and β-sheets. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 4130-4134 (1977).
3. Chothia, C. & Janin, J. Relative orientation of close-packed β-pleated sheets in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 4146-4150 (1981).
4. Chothia, C. & Janin, J. Orthogonal packing of β-pleated sheets in proteins. *Biochemistry* **21**, 3955-3965 (1982).
5. Cohen, F. E., Sternberg, M. J. E. & Taylor, W. E. Analysis of the tertiary structure of protein β-sheet sandwiches. *J. Mol. Biol.* **148**, 253-272 (1981).
6. Cohen, F. E., Sternberg, M. J. E. & Taylor, W. R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature* **285**, 378-382 (1982).
7. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein. Chem.* **34**, 167-339 (1981).

8. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. I. A Theoretical Analysis. *J. Mol. Biol.* **236**, 1369-1381 (1994).

9. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* **236**, 1382-1400 (1994).

10. Murzin, A. G., Brenner,S. E., Hubbard, T. & Chothia, C. SCOOP: A structural clasification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540, (1995).

11. Janin, J. & Chothia, C. Packing of α-Helices onto β-Pleated Sheets and the Anatomy of α/β Proteins. *J. Mol. Biol.* **143**, **95**-128, (1995).

12. Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science* **2,** 1811-1826, (1993).

13. Orengo, C. A., Flores, T. P., Taylor, W.,R. & Thornton, J. M. Identification and clasification of protein fold families. *Protein Engineering* **6**, 485-500 (1993).

14. Hilbert, M., Böhm, G. & Jaenicke, R. Structural Reletionships of Homologous Proteins as a Fundamental Principale in Homology Modeling. *Proteins: Struct. Funct. Genet.* **17**, 138-151 (1993).

15. Sander, C. & Schneider, R. Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68 (1991).

16. Pascarella, S. & Argos, P. A data bank merging related protein structures and sequences. *Prot. Enging.* **5**, 121-137 (1992).

17. Chothia, C. One thousand protein families for the molecular biologist. *Nature* **357**, 543-544 (1992).

18. Yee, D. P., & Dill, K. A. Families and the structural relatedness among globular proteins. *Prot. Sci.* **2**, 884-899 (1993).

19. Rufino, S. D., & Blundell, T. L. Structural-based identification and clustering of protein families and superfamilies. *J. Comput. Aided Mol. Design* **8**, 5-27 (1994).

20. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thorntorn, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* (London) **323**, 347-352 (1987).

21. Sali, A. Overington J. P., Johnson, M. S. & Blundell, T. L. From Comparisons of Protein Sequences and Structures to protein Modelling and Design. *Trends Biochem. Sci.* **15**, 235-240 (1990).

22. Hubard, T. J. P., Blundell, T. L. Comparison of solvent-inaccessible cores of homologous proteins: Definitions usful for protein modelling. *Prot. Eng.* **1**, 159-171 (1987).

23. Chotia. C., Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826 (1986).

24. Chotia. C., Lesk, A. The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **LII**: 399-405 (1987).

25. Bowie, J. U., Luthy, R., & Eisenberg, B. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science* **253**, 164-172 (1991).

26. Ponder, J. M., & Richards, F. M. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791 (1987).

27. Jones, D. T., Taylor, W. R., & Thornton, J. M. A new approach to protein fold recognition. *Nature* (London) **358**, 86-89 (1992).

28. Wilmanns, M. & Eisenberg, D. Intverted protein folding by the residue pair preference profile method. *Prot. Eng.* **8**, 627-639 (1995).

29. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc. Natl. Acad. Sci. USA* **89**, 9029-9033 (1992).

30. Bryant, S. H., Lawrence, C. E. An empirical energy function for threading protein sequence through folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92-112 (1993).

31. Thomas, P. D. & Dill, K. Statistical Potentials Extracted from protein structures: how Accurate are They? *J. Mol. Biol.* **257**, 457-469 (1996).

32. Lemer, C. M.-R., Marianne, J. R. & Wodak, S. J. Protein Structure Prediction by Threading Methods: Evaluation of Current Techniques. *Proteins: Struct. Funct. Genet.* **23**, 337-355 (1995).

33. Fischer, D., Rice, D., Bowie, J. U. & Eisenberg, D. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.* **10**, 126-136 (1996).

34 Casari, G. & Sippl, M. J. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identity native folds. *J. Mol. Biol* **244**, 725-732, (1992).

35 Bahar, I. & Jernigan, R. L. Inter-residue Potentials in Globular Proteins and the Dominance of Highly Specific Hydrophilic intrractions at Close separation. *J. Mol. Biol* **266**, 195-214, (1997).

36. Finkelstein, A. V. & Reva, B. A. A search for the most stable folds of protein chains. *Nature* **351**, 497-499 (1991).

37. Chotia, C. & Finkelstein, A. V. The clasification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007-1039 (1990).

38. Richardson, J. S. β-Sheet topology and the reletednese of proteins. *Nature* **268**, 495-500 (1977).

39. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein. Chem.* **34**, 167-339 (1981).

40. Sternberg, M. J. E., & Thorntorn, J. M. On the Conformation of Proteins: Towards the Prediction of Strand Arrangements in β-Pleated Sheets. J. M. *J. Mol. Biol.* **113**, 401-418 (1977).

41. Taylor, W. R. Towards protein tertiary fold prediction using distance and motif constraints. *Prot. Eng.* **4**, 853-870 (1991).

42. Plaxco, K. W., Simons, K. T., & Baker, D. (1998). Contact order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *J. Mol. Biol.* **277**, 985-994.

43. Martinez, J. C., Tereza Pisaborro, M., & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **5**, 721-729.

44. Grantcharova, V. P., Riddle, D. S, Santiago, J. V., & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol.* **4**, 715-721.

45. Dimitrov, R. A., Laane, C., Vervoort, J., & Crichton, R. R. (1998). Topological requirement for the nucleus formation of a two-state folding reaction. Implications for Φ-values calculations. *Protein Science*, submitted,

46. Viguera, A. R., Serrano, L. (1997). Loop length, intramolecular diffusion and protein folding. *Nat. Struct. Biol.* **4**, 939-946.

47. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

48 Dimitrov, R. A., & Crichton, R. R. (1997). Self-Consistent Field Approach to Protein Structure and Stability. I. pH Dependence of Electrostatic Contribution. *Proteins Struct. Funct. Genet.* **27**, 576-596.

49 Dimitrov, R. A. & Crichton, R. R. Tertiary fold prediction of globular proteins: A molecular field approach. 5th international conference: Perspectives on protein engineering, "From folds to functions", Montpellier, France, 2-6 march 1996.

50. Dimitrov, R. A., Crichton, R. R. & Vervoort, J. (1998). From Fold Prediction to Protein Design Using Self-Consistent Field Approach. Twenty-third annual Lorne conference on protein structure and function. Australia, 8-12 February.

51 Gibbs, J. W. "Elementary Principles in Statistical Mechanics", New Haven (1902).

52 Callen, H. B. "Ttermodynamics and an introduction to thermostatistics", John Wiley & Sons, New York (1985).

53 Otzen, E. D. & Fersht, A. R. Side-Chain Determinants of β-Sheet Stability. *Biochemistry* **34**, 5718-5724 (1995).

54 Qian, H. & Chan, S. I. Interactions Between a Helical Residue and Tertiary Structures: Helix Propensities in Small Peptides and in Native Proteins. *J. Mol. Biol.* **261**, 279-288 (1996).

55 O'Neil, K. T. & DeGrado, W. F. A Thermodynamic Scale for the Helix-Forming Tendencies of the Commonly Occurring Amino Acids. *Science* **250**, 646-651 (1990)

56 Minor, D. L. Jr & Kim, P. S. Context is a major determinant of β-sheet propensity. *Nature* **371**, 264-267 (1994).

57 Minor, D. L. Jr & Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367**, 660-663 (1994).

58 Platzer, K. E. B., Ananthanarayanan, V. S., Andreatta, R. H., & Scheraga, H. A., Helix-Coil Stability Constants for Naturally Occurring Amino Acids in Water.IV. Alanine Parameters from Randam Poly(hydroxypropylglutamine-*co*-L-alanine). *Macromolecules* **5**, 177-187 (1972).

59 Fauchere, I. I. & Pliska, V. Hydrophobic parametrs π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem. Chim. Ther.* **18**, 369-375 (1983).

60 Landau, L. D. & Lifshitz, E. M. Statistical Physics, Part I. London: Pergamon (1959).

61 Behe, M. J., Lattman, E. E., & Rose, G. D. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? *Proc. Natl. Acad. Sci. USA* **88**, 4195-4199 (1991).

7 Summary

The research described in this thesis has been carried out to obtain a better understanding of the fundamental rules describing protein folding and stability. There are a few major questions around which this thesis is focused:

- What are the sequence requirements for proteins to fold rapidly and to be stable in their native conformations?
- What are the thermodynamic mechanism(s) of protein stabilization and the kinetic mechanism(s) of folding?
- Are there special native structures (packing patterns) that are more likely to correspond to the native structures of foldable proteins?
- What is the best approximation for protein-folding energetics?

The organization of the thesis is as follows:

CHAPTER 2

After a short introduction (chapter 1), a review of the basic physical principles that govern protein structure and the thermodynamics, as well as the kinetics of protein folding and unfolding reactions is presented.

CHAPTER 3

This chapter focuses on the electrostatic contribution to the protein stability. Electrostatic interactions are described on the basis of a novel approach which uses the idea of a self-consistent field adapted from statistical mechanics. This theoretical approach describes in detail properties as titration curves, protein stability and pK_a shifts. The main conclusions are: firstly, the calculated results are in excellent agreement with the experimental data, when the solution of Poisson-Boltzmann equation (PB) is based on the assumption that the ionized residues are seen as part of the high dielectric medium (rather than the interior of the protein molecule);

Secondly, the solution of PB equation outside the protein interior, depends on local characteristics, such as the packing of chain portions around ionized residues, rather than on the detailed shape of the protein molecule. Lastly, at "natural-like" conditions the contribution of electrostatic interactions to the free energy difference between the unfolded and folded states of protein molecules is close to zero. This indicates that the main driving forces for folding of protein molecules under these conditions are hydrophobic and backbone-backbone hydrogen bonding interactions.

CHAPTER 4

Chapter 4 concerns the application of the theory of electrostatic interactions to the calculation of the pKa's of the 98 residue β-elicitin, cryptogein. Unusual in this protein is the existence of four ionized groups buried in the hydrophobic core. This touches upon an important question: what are the dielectric properties of the protein interior? The NMR structure of the 98 residue β-elicitin, cryptogein was determined using ^{15}N and $^{13}C/^{15}N$ labeled protein samples. The structure was calculated with 1047 intrasubunit and 40 intersubunit NOE derived distance constraints and 236 dihedral angle constraints for each subunit using the program DYANA. The twenty best conformers were energy-minimized in OPAL to give a root-mean-square deviation to the mean structure of 0.82 Å for the backbone atoms and 1.03 Å for all heavy atoms. Using $^{1}H,^{15}N$ HSQC spectroscopy the pKa of the N- and C-termini, Tyr-12, Asp-21, Asp-30, Lys-61, Asp-72, Tyr-85 and Lys-94 were determined and the obtained results support the proposal of several stabilizing ionic interactions including a salt bridge between Asp-21 and Lys-62. The hydroxyl hydrogens of Tyr-33 and Ser-78 are clearly observed, indicating that these residues are buried and hydrogen bonded. Two other tyrosines, Tyr-47 and -87, are also not solvent accessible and show pKa's > 12. However, there is no indication that their hydroxyl atoms are hydrogen bonded. Calculation of theoretical pKa's show general agreement with the experimentally determined values and are similar for both the crystal and solution structures.

CHAPTER 5

In chapter 5 the topological requirement for the nucleus formation of a two-state folding reaction is considered. The self-consistent field approach is discussed to calculate the free energy of the folding nucleus. A self-consistent field is used to approximate the description of

the elementary long-range interactions, such as hydrogen bonding and hydrophobic interactions. Residue propensities for the corresponding α-, β- and irregular chain regions as well as the free energy of chain bending are considered in an explicit form based on experimental parameters. A theoretical model for the folding of two-state small monomeric proteins is proposed. The folding problem is reduced to the question of how the folding nucleus in the transition state (TS) is formed from the ensemble of rapidly interconverting, partly structured conformations in the denatured state. It is proposed that in the denatured state the folding is energetically favored by certain highly fluctuating nucleation regions (α-helices and(or) β-hairpins). In experiments based on site directed mutagenesis these nucleation regions are revealed by their high Φ-values. In the TS folding is favored by the packing of these nucleation regions together with other portions of the polypeptide chain thus leading to a broad distribution of the Φ-values. As a result, the folding nucleus with native-like topology and approximately correctly formed secondary structures and loops is favored over other folding nuclei.

CHAPTER 6

In chapter 6 a discussion of the problem of protein fold recognition of small monomeric proteins with less than 80 residues is represented. The fold recognition strategy is based on the fact that: firstly, at the transition state level all possible protein conformations can be split out into different ensembles of similar structures. The crude characteristics of these ensembles can be described by the limited set of thermodynamically most favorable protein folds; secondly, the folding nucleus with native-like overall fold is separated from all other folding alternatives by a high free energy barrier. As a result, at the lowest free energy minimum of the TS state the protein molecule propagates towards its native state approximately isoenergetically through an ensemble of conformations of its native fold. The main contributions, which stabilize the protein folds at the TS level, are the hydrophobic and long-range backbone hydrogen bonding interactions, as well as the free energy of chain bending, and free energies of secondary structures formation. The selection of the protein architectures is mainly determined by the most general characteristics encoded in the protein sequence, such as distribution of hydrophobic and hydrophilic residues along the chain, the ability of amino acids to form different secondary structures in compact chain conformations by hydrogen bonding with their spatial or chain neighbors, and the general rules which govern the packing of the secondary

structures. At the TS, the free energies of the native folds are separated by a 'gap' from the lowest free energies of the folds from structurally different packing patterns. However, the difference between the free energies of the native folds and the lowest free energies of the folds from the same packing patterns is rather small.

CONCLUSIONS

In conclusion, this thesis has resulted into new insights on the folding of protein molecules and how the folding reaction depends on such characteristics as environment conditions (concentration of denaturants, pH, temperature etc.) and amino acid composition. In particular the folding problem of a small two-state monomeric proteins is reduced to the question of how the folding nucleus in the TS is formed from the ensemble of rapidly interconverting partly structured conformations in the denatured state. It is shown that sequence, topological and symmetrical restrictions fully determine the folding pathway between the quasi-equilibrium pre-transition region and the TS. It is realized that the search for a unique structure involves the discrimination between different overall folded structures, and that the collapse into a globule having secondary structure does not by itself solve the problem of the search for a unique three-dimensional structure. An improvement in predicting the correct energetics of the residue-residue interactions is achieved. A lattice model, based on a more realistic representation of protein chain conformations, including secondary structure formation and side chain packing which are crucial for the formation and stabilization of the native state is presented.

Future research should elucidate the role of sequence and topological restrictions on the nucleus formation in multi-domain proteins. Steric restriction arising from the packing of residues with different volumes of their side chain groups should be taken into account also. In addition a more realistic representation for the loop regions is needed in the case of proteins with sequences greater then 80 residues.

Samenvatting

Het onderzoek in dit proefschrift is erop gericht een beter inzicht te krijgen in de factoren die bijdragen tot eiwitvouwing en eiwitstabiliteit. Op de volgende vragen wordt getracht een antwoord te verkrijgen:

-Aan welke voorwaarden moet de aminozuurvolgorde van een eiwit voldoen om snel en stabiel te vouwen tot een stabiele, natieve conformatie ?

-Wat zijn de thermodynamische karakteristieken van eiwitstabilisatie ?

-Wat zijn de kinetisch mechanismen van eiwitvouwing ?

-Zijn er speciale basisstrukturen ("pakkingspatronen") die overeenkomen met de strukturen van eiwitten ?

-Wat is de beste manier om de energie-processen van eiwitvouwing te beschrijven?

De opbouw van dit proefschrift is als volgt: Na een korte inleiding over het belang van de bestudering van vouwingsprocessen (Hoofdstuk 1), volgt een overzicht van de elementaire fysische principes die de structuur van een eiwit bepalen (Hoofdstuk 2). Naast thermodynamische grondslagen worden ook de kinetische aspecten van vouwing en ontvouwing behandeld. Hoofdstuk 3 handelt over de rol van electrostatische interacties. Om inzicht te krijgen in de rol van electrostatische interacties in eiwitten is een nieuwe fysische benadering toegepast: +Self-Consistent Field Approach-. Deze theoretische benadering kan gebruikt worden zowel voor pK_a als voor eiwitstabiliteit berekeningen. De belangrijkste conclusies uit Hoofdstuk 3 zijn:

a) De verkregen oplossingen van de Poisson-Boltzmann vergelijkingen zijn in zeer goede overeenstemming met de experimentele gegevens;

b) De oplossingen van de Poisson-Boltzmann vergelijkingen buiten de kern van een eiwit molecuul zijn afhankelijk van locale eigenschappen, zoals de pakking van de zijketens van aminozuurresiduen rond geladen groepen;

c) Tenslotte, de bijdrage van electrostatische interacties aan de vrije energie verschillen tussen gevouwen en ontvouwen eiwitten is onder natuurlijke condities zeer gering. De belangrijkste drijvende krachten bij eiwitvouwing zijn hydrofobe en +backbone-backbone- H-brug interacties.

Hoofdstuk 4: pKa's van geladen aminozuren van β-elicitine worden berekend. Hierbij wordt gebruik gemaakt van de door ons ontwikkelde theorie van electrostatische interacties. Opmerkelijk is in β-elicitine dat er vier geladen aminozuren in het hydrofobe centrum te vinden zijn. De vraag kan derhalve gesteld worden: Wat zijn de dielectrische eigenschappen in het inwendige van dit eiwit ? De drie-dimensionale struktuur van dit 98-aminozuur grote eiwit is bepaald met behulp van bekende NMR methoden. Het was noodzakelijk het eiwit te verrijken met ^{15}N en $^{13}C/^{15}N$. De 3D struktuur is berekend met het programma DYANA, gebruikmakende van 1047 intrasubunit, 40 intersubunit afstanden (verkregen uit NOE spectra) en 236 dihedrale hoeken. De 20 beste strukturen zijn energie geminimaliseerd met behulp van OPAL. Uit de NMR resultaten zijn de pKa's van Tyr-12, Asp-21, Asp-30, Lys-61, Asp-85 en Lys-94 verkregen.

De berekening van de pKa's van deze residuen, met behulp van de +Self-consistent Field Approach- uit hoofdstuk 3, laat zien dat er een goede overeenkomst is met de experimenteel verkregen waarden.

In Hoofdstuk 5 wordt de topologische vereisten voor +nucleus--vorming van een twee-toestanden vouwingsreactie bestudeerd. De +Self-consistent Field Approach- wordt gebruikt om de vrije energie van de vouwings +nucleus- uit te rekenen. Een goede berekening van de H-brug en hydrofobe interacties is hierbij essentieel. Ook worden aminozuur "propensities" ("neigingen") voor α-helices, β-sheets en "loop" gebieden, alsmede de vrije energie van een polypeptide keten "vervorming" in een expliciete vorm (dmv experimentele gegevens) meegenomen in de berekeningen. Een theoretisch model voor de vouwing van een twee-toestanden vouwingsreactie van kleine monomere eiwitten wordt door ons geponeerd. Het vouwingsprobleem wordt gereduceerd tot de vraag hoe de vouwings "nucleus" in de overgangstoestand gevormd wordt uit een ensemble van snel in elkaar overgaande, gedeeltelijk gevormde strukturen van de ontvouwen toestand. Wij stellen voor dat in de ontvouwen toestand van een eiwit de vouwing energetisch bevoordeeld wordt door snel fluctuerende "nucleation" gebieden (α-helices of β-hairpins). In experimenten van Fersht en Serrano worden

deze gebieden gekenmerkt door hoge Φ-waarden. In tegenstelling tot de ontvouwen toestand van een eiwit wordt de vouwing van de overgangstoestand bevorderd door de groepering ("packing") van deze "nucleation" gebieden met andere gedeelten van de polypeptide keten. In de overgangstoestand heeft de vouwings "nucleus" een struktuur die op de werkelijk gevouwen toestand lijkt (een "native-like topology"), met bijna correct gevormde secundaire strukturen en "loops". Het ensemble van op elkaar gelijkende strukturen in de overgangstoestand heeft een energie verdeling met een Boltzmann distributie.

In hoofstuk 6 wordt het probleem van eiwitvouwing van kleine momomere eiwitten besproken. We trachten de vraag te beantwoorden waarom een eiwit in een bepaalde "fold" terecht komt en niet in een andere "fold". Onze strategie is gebaseerd op het gegeven dat:

a) in the overgangstoestand alle mogelijke conformaties van een eiwit uitgesplitst kunnen worden in verschillende ensembles van op elkaar gelijkende strukturen. De (grove) karakteristieken hiervan kunnen beschreven worden door een beperkte set van de thermodynamisch meest waarschijnlijke "folds".

b) de vouwings +nucleus- met de "fold" die het meest lijkt op de natieve toestand is energetisch gescheiden (dwz lager in energie) van alle andere vouwingsalternatieven.

Dientengevolge, gaat het eiwit molecuul in de laagste energie toestand van de vouwings "nucleus" bijna isoenergetisch, via een ensemble van conformaties van de natieve "fold", naar de natieve gevouwen toestand. De belangijkste bijdragen aan het stabiliseren van de eiwit "folds" in de overgangstoestand zijn hydrofobe en "long-range backbone-backbone" H-brug interacties, de vrije energie van een polypeptide keten vervorming en de vrije energie van secundaire struktuur formatie. De beantwoording van de vraag waarom een eiwit in een bepaalde "fold" terecht komt en niet in een andere "fold" wordt bepaald door algemene karakteristieken, zoals de aminozuur sequentie, de distributie van hydrofiele en hydrofobe residuen, de mogelijkheid van aminozuren om verschillende secundaire strukturen aan te nemen, alsmede door algemene regels die de pakking bepalen van secundaire struktuur toestanden. In de overgangstoestand zijn de vrije energieën van de natieve "folds" lager in energie dan de laagste vrije energie niveaus van de andere "folds". Echter het verschil in energie is gering.

Заключение

Проведените в представената досторска дисертация изследвания имат за цел да разширят и уточнят съществуващите представи за фундаменталните физични принципи които описват белтъчната кинетика (БК) и термодинамична стабилност (ТС) на белтъчната молекула (БМ). Следните фундаментални въпроса стоият в основата на проведените изследвания:

-Какви отраничения трябва да се наложат върху възможните вариации в разпределението на амино киселините (АК) по дължината на белтъчната верига (БВ) за да се намали времето за което БМ преминавайки през различни конформационни изменения се трансформира в своята биологично активна конформация (БАК)? Как наложените ограничения върху разпределението на АК се отразяват на ТС на белтъчната молекула?

-Какви са физическите принсиципи и механизмът на БК и ТС на БАК?

-Съществува ли специален набор от конформации чрез който и само чрез който БМ могат да осъществяват своите биологично активни функции?

-Какви са основните взаимодействия които описват конформационните и енергетични изменения а така също направлението на БК по посока на БАК?

Представената докторска дисертация е организирана в следните глави:

ГЛАВА 2

След кратко въведение (глава 1), в глава 2 е изложен обзор на сеществуващите физични теории за фундаменталните принципи които са положени в основата на третичната структура на БМ, нейната ТС и механизмът чрез който БМ се трансформира в

права и обратна посока между своята БАК и ансамблът от конформации който се получава при нейната денатурация.

ГЛАВА 3

В тази глава е изследвана природата на електростатическите взаимодействия и техния принос в ТС на БМ. Електростатическите взаимодействия са разгледани на основата на статистическата механика, използвайки метода на самосъгласованото молекулярно поле (СМП). Методът на СМП дава възможност детайлно да се изследват такива свойства на БМ като: криви на титруване, ТС и изменения в pK на йонизуемите групи при тяхното титруване в съставът на БМ и по отделно. Получени са следните основни резултати:

-Проведените изчисления на основата на СМП и на предположението че йонизуемите групи, които са разположени на повърхността на БМ, имат висока диелектрична проницаемост са в отлично съгласие с експерименталните данни, получени при титруване на 74 йонизуеми групи принадлежащи на 6 БМ;

-Решението получено на основата на СМП зависи само от локалната плътност в разпределението на белтъчния материал около йонизуемите групи, но не от формата на БМ;

-Във физиологични условия приносът на електростатическите взаимодействия в ТС на БМ е много малък и може да се пренебрегне в сравнение с приноса на хидрофобните и водородни взаимодействия.

ГЛАВА 4

В тази глава е разгледано приложението на разработената в глава 2 теория на електростатическите взаимодействия за оценка на измененията на pK на йонизуемите групи в БМ β-clicin cryptogen. Измененията се отчитат по отношение на експериментално определените стандартни стойности на pK на отделните йонизуеми групи. Изчисленията са проведени използвайки структура на β-clicin получена на основата на ЯМР използвайки ^{15}N и $^{13}C/^{15}N$ като маркери. Третичната структура на β-clicin е получена на основата на набор от NOE разстояния-1047 между двата домена и 40 за всеки домен поотделно. От получените структури са отделени 20-те най-добри. За всяко от 20-те

структури е извършена минимизация на нейната енергия използвайки програмата OPAL. Средните квадратични отклонения на атомите от пептидните групи са 0.82. За всички останали тежки атоми средното квадратично отклонение е 1.03. Използвайки HSQC спектрофотометричните данни за ^{1}H и ^{15}N, pK на следните групи са определени: крайните N- и C-групи на БМ, Tyr-12, Asp-21, Asp-30, Lys-61, Asp-72, Tyr-85 и Lys-94. Получените резултати дават основание да се предположи наличието на множество стабилизиращи взаимодействия между йонизуемите групи в това число солева-връзка между Asp-21 Lys-62. Получените данни показват още, че Tyr-33 и Ser-78 образуват водородна връзка и са разположени в хидрофобното ядро на БМ. Tyr-47 и-87 са също недостъпни за разтвора и показват pK >12. Но получените експериментални данни за тези йонизуеми групи не потвърждават тяхното участие в образуване на водородни връзки. pK на йонизуемите групи изчислени на основата на разработената в глава 2 теория на електростатическите взаимодействия, показват много добро съгласие с експерименталните данни. Изчисленията проведени върху третичните структури на β–clicin, получени на основата на ЯМР и ренгено-струстурен анализ, показват сходни резултати.

ГЛАВА 5

В тази глава е разгледано влиянието на топологическите ограничения върху упоковката на БВ при образуването на бариерния зародиш на биологически активната третична структура. Разгледан е подробно въпросът за приложението на СМП за пресмятане на свободната енергия на зародиша. СМП се използва при оценката на взаимодействието между страничните атомни групи на АК, които от една страна са разположени близко в пространството, а от друга страна разстоянието между тях по дължината на БВ е голямо. Взаимодеиствието между съседни по дължината на БВ АК се разглежда на основата на експериментално определени параметри. Предложен е модел който обяснява какви конформационни трансформации претърпява БМ в хода на образуване на нейната БАК. Така например за БМ с дължина на БВ не по-голяма от 80 АК е показано че във физиологически условия кинетиката на образуване на БАК преминава през две основни състояния:

-денатурирано състояние с компактна но дифузна опаковка на БВ;

-нискоенергетично, кратко живеещо бариерно състояние.

В денатурираното състояние БМ извършва дифузно движение между различни конформационни състояния. Конформационните изменения са много бързи и не водят до натрупване на БМ в определени конформационни състояния. Предложен е механизъм съгласно който в определен порядък по дължината на БВ се образуват α-спирали и β-шпилки. Те служат като зародиши за упаковката на БВ. Положението на α- и β-зародишите по дължината на БВ се определя на основата на пертурбации в изменението на свободната енергия на БМ при прехода и й от денатурираното в бариерното състояние. Зародишите са разположени в тези участуци по дължината на БВ, където измененията в АК водят до значителни пертурбации в изменението на свободната енерия. Останалите участъци от БВ, които взимат участие в упаковката на БМ, са сравнително инертни по отношение на мутации. По такъв начин упаковката на БВ, в кратко живеещите нисko-енергетични бариерни конформации, води до образуването на широк спектер от пертурбации в свободната енергия в пълно съгласие с експерименталните данни. Кратко живеещите бариерни състояния се характеризират с определен порядък на вторичните струстури по дължината на БВ и тяхната взаимна опаковка в пространството. Последната е свързана с ролята на нативната топология на БМ.

ГЛАВА 6

В тази глава се разглежда въпроса за предсказване на топологията на кратко живеещите бариерни конформационни състояния. Основа за такова предсказване се явяват следните факти:

-всички бариерни конформации могат да бъдат разделени на групи, като конформациите от всяка група имат еднакво разпределение на вторичните структури по дължината на БВ и еднакъв порядък на тяхната взаимна упаковка в пространството.;

-свободната енергия на групата от конформации, която се отличава с нативна топология на упаковката на БВ, образува минимума на бариерното конформационно състояние и се отделя от спектъра на свободните енергии на останалите групи.

Основните взаимодействия които стабилизират нативната топология на бариерното състояние са: хидрофобни взаимодействия; водородните връзки между амидните и хидрохилни атомни групи на основната верига от ковалентни връзки на БМ; свободната енергия на огъване на нерегулярните участъци по дължината на БВ, които свързват вторичните струстури и свободната енергия с която отделните АК участвубат при образуването на α– и β–вторичните струстури. Селекцията на белтъчната топология се базира на най-общи закономерности закодирани в разпределението на АК по дължината на БВ: разпределението на хидрофобните и хидрофилни АК; разликата в свободните енергии с които отделните АК участвуват при образуването на вторичните структури и общите принципи които управляват упаковката на вторичните структури.

ЗАКЛЮЧЕНИЕ

В заключение, получените в представената докторска дисертация резултати водят до ново разбиране и нова формулировка на въпроса за механизмът на БК и ТС на биологично активната структура на БМ. Тази нова формулировка дава възможност за първи път да се постави и изследва въпросът за ролята на нативната топология при образуване на бариерните конформации. Така например в случая на БМ с дължина на БВ не по-голяма от 80 АК е предложен модел на БК в съгласие с който последователността от конформационни трансформации между денатурираното и бариерното състояние напълно се определя в рамките на определена 'кинетична пътечка'. 'Кинетичната пътечка' от своя страна се определя от последователността на АК по дължината на БВ, а така също от топологията и симетрията на биологически активната белтъчна структура. Един от важните резултати е ясното разбиране че упаковката на БВ в компактно бариерно състояние с участието на множество вторични структури не решава автоматично избора на нативната топология. Необходима е селекция измежду различни компактни упаковки на БВ, които се определят от различните разпределения на вторичните структури по дължината на БВ и тяхната упаковка в пространството. Интерестно е да се отбележи че правилния избор на нативна топология се съгласува и с наличието на 'кинетична пътечка' за упаковката на БВ и с

наличието на на термодинамичен минимум който съответствува на нативната топология. Предложен е нов решетъчен модел който описва бариерните конформационни състояния. Този нов решетъчен модел позволява БМ да се опише подробно даже на атомно ниво. В същото време, базирайки се на СМП, решетъчният модел позволява за няколко секунди (използвайки персонален компютър) да се изчислят с удивителна точност пертурбациите на свободната енергия на бариерното състояние, които са достъпни само в много сложни и скъпо струващи експерименти.

В заключение, бъдещи изследвания трябва да бъдат проведени бърху многодоменни БМ. Необходимо е да се изследва още влиянието на стерическите взаимодействия между страничните атомни групу на АК при упаковката на БВ. Стерическите взаимодействия имат силно селективно свойство. Интересно е да се установи дали и при многодоменните БМ също съществува 'кинетична пътечка'.

Acknowledgements

First of all, I would like to thank my promotors Robert Crichton and Colja Laane, who gave me the opportunity to work at their respective laboratories. Their critical and sharp opinions always helped me to move on with the research. I would especially like to thank Colja Laane whose excellent and critical guidance was essential to finish this dissertation.

I give my very special thanks to my co-promotor Jacques Vervoort for his scientific support and for helping me to solve my every/day problems. To begin a new life in a foreign country is exciting but always difficult. I would like to thank Jacques Vervoort, Ivonne Rietjens, Laura Ausma together with Ankie Lamberts from the Deans Office for solving my housing problem in a very desperate situation.

During my stay in Wageningen I have met very nice people. I have had a wonderful time with Hans Wassink, Adrie Westphal, Ahmed Osman, Michel Eppink, Eyke van den Ban, Walter van Dongen, Willy van den Berg, Carlo van Mierlo, Aart de Kok, Ton Visser, Martin Bouwmans and Martina Duyvis.

In Wageningen I have been collaborating with many researchers. I am very grateful to the co-authors of the chapter 4, especially to Paul Gooley from the Department of Biochemistry, University of Melbourne, Australia. I also would like to thank Luis Serrano from the EMBL, Heidelberg, Germany whose collaboration and advice made it possible to finalize my favorite chapter 5.

I also would like to thank my Belgian friends Prof. G. L. Hennebert, Cristophe Henry, Agnes and Hugo Mignet for their attention and support during my stay in Belgium. I would especially like to thank Roberta Wade for her care about my family.

Last but not least, I want to thank my family, my daughter Yasmina and especially my wife Dilnora for being patient, supportive even in the most difficult moments.

Curriculum vitae

Roumen Atanasov Dimitrov was born on July 1th 1959 in Sofia, Bulgaria. He received his high school diploma in 1977 from the "Georgi Dimitrov" High School in Sofia (Bulgaria). In 1985 he obtained his master of science degree at the Department of Theoretical Condensed Matter Physics from the "Kliment Ohridsky" University (Sofia, Bulgaria). Between 1985 and 1998 he was employed as a research scientist at the Institute of Organic Chemistry, Bulgarian Academy of Sciences. From 1990 to 1994 he had a post-graduate specialization on computational molecular biology at the Institute of Protein Research, Puschino, Moscow, Russia. In 1995 he started his Ph.D. studies at the Unité de Biochimie, l'Université de Louvain-la-Neuve, Belgique (Prof. dr. R.R. Crichton). From September 1th 1997 till February 1th 1999 he finalized his Ph.D. programme at the Department of Biochemistry, Wageningen University and Research Centre (Prof. dr. Colja Laane, dr. J. Vervoort). The research which was carried out during his Ph.D. period is described in this thesis.

List of publications

R. A. Dimitrov & R. R. Crichton. Tertiary fold prediction of globular proteins: A molecular field approach. Proceeding 5th International Conference: Perspectives on protein engineering, "From folds to functions", Montpellier, France, 2-6 march, 1996, p.

R. A. Dimitrov & R. R. Crichton. Self-Consistent Field Approach to Protein Structure and stability. I. pH Dependence of Electrostatic Contribution. *Proteins: Structure, Function and Genetics*, **4**: 576-596, 1997.

R. A. Dimitrov, R. R. Crichton & J. Vervoort. From Fold Prediction to Protein Design Using Self-Consistent Field Approach. Proceeding Twenty-third annual Lorne conference on protein structure and function. Australia, 8-12 February, 1998, p. A-23.

P. R. Gooley, M. A. Keniry, R. A. Dimitrov, D. E. Marsh, D. W. Keizer, K. R. Gayler & B. R. Grant. The NMR solution structure and characterization of pH dependent chemical shifts of the beta-elicitin, cryptogein. *J. Biom. NMR* in press, (1998).

R. A. Dimitrov, C. Laane, J. Vervoort & R. R. Crichton. Topological requirement for the nucleus formation of a two-state folding reaction. Implications for Φ-values calculations. *Protein Science*, submitted, (1998).

R. A. Dimitrov, C. Laane, J. Vervoort & R. R. Crichton. Fold prediction of α, β, α/β and α+β protein architectures. *Proteins: Structure, Function and Genetics*, submitted, (1998).