INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/775 -4

COLLEGE IN BIOPHYSICS:
EXPERIMENTAL AND THEORETICAL ASPECTS OF BIOMOLECULES

26 September - 14 October 1994

Miramare - Trieste, Italy

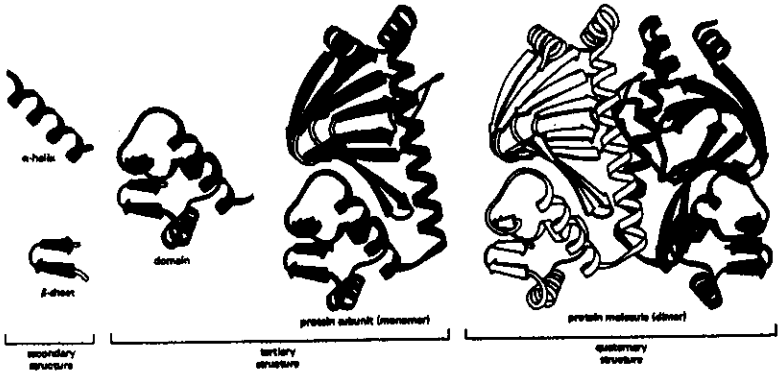
Structure and Geometry of Proteins and Membranes: A Physical View

Henrik Bohr
The Technical University of Denmark
Lyngby, Denmark

STRUCTURE AND GEOMETRY OF PROTEINS AND MEMBRANES: A PHYSICAL VIEW

Henrik G. Bohr
CBS,Department of Physical Chemistry, DTU The Technical University of Denmark, Lyngby DK-2800, Denmark.

Lecture notes about protein and lipid structure analysis. The lectures were delivered in September 1994 at the International Center for Theoretical Physics, ICTP, in Trieste, Italy for the college in biophysics. The lectures are made for an audience interested in working on mathematical models in molecular biology from a theoretical physics perspective.



CONTENT

6. Protein structure and chemical reaction kinetics

6a. Translation of contact dynamics to chemical reaction kinetics.

6b. Crossing of activation barriers.

6c. Path Integral Formalism.

7. Structure of biological membranes

7a. Phenomenology of agglomerations of lipids in bio-membranes.

8. Differential geometrical model of closed membranes

8a. Differential geometry of membranes as 2-dimensional embeded surfaces.

8b. Topological thermodynamics of closed membranes.

10. Future outlook

References after each chapter.

The following text is for many parts more or less taken from original papers written by the author and co-workers. In order to make the credit clear Chapter 1 has parts of paragraphs that we used for our introduction in the book: Protein structure by distance analysis, by H. Bohr and S. Brunak, IOS press (1994). Chapter 2 is standart text book biochemistry. Chapter 3 is partly taken from the paper: Protein fold class prediction by knowledge based systems, by M. Reczko, H. Bohr, P. Sudhakar, A. Hatzigeorgiou and S. Subramaniam (to be published in Protein Ingeneering (1994). Chapter 4 is from the paper: How many fold classes are to be found, by Per-Anker Lindgaard and H. Bohr (submitted to FEBS Letters (1994). Chapter 5 is partly from the article: Initial events in protein folding from an information processing viewpoint, by H. Bohr and P. G. Wolynes, Phys. Rev A. 46, 5242 (1992). Chapter 6 is from the paper: Domain growth in protein folding, by J. Wang, H. Bohr and P. G. Wolynes (unfinished but partly published in the book: Protein structure by distance analysis, IOS press (1994). Chapter 7 and 8 are from a paper called: Thermodynamics and Topology of clossed membranes, by H. Bohr, John Ipsen and Steen Markvorson (submitted to J. de Physique, 1993). The figure material is from the book: The structure and action of proteins, by R. Dickerson and I. Geis, W. A. Benjamin Inc. (1969).

Introduction

The field of protein structure determination contains a vast and ever increasing amount of scientific contributions due to the great importance of protein design and functionality in bio-technology, and, even more owing to the fact that prediction of accurate 3-dimensional structures of proteins from their sequence is still an unsolved problem.

In the light of this vast landscape of scientific information and achievements, aiming ultimately at fulfilling the goal of protein structure prediction from genome sequence data, the present collection of lecture notes is intended to address the more limited aspect of protein structure determination in the distance geometry approach in order to obtain a clearer picture of the state of the art for a part of the subject while avoiding more general notions of protein folding already described well elsewhere[1]. In discussing protein structure determination it is important to present both experimental as well as theoretical aspects of the subject in order to obtain a balanced presentation of facts and speculation.

The distance geometry approach to protein structure determination, which we shall focuse strongly on in these lecture notes, is in the following to be understood as protein structure analysis, experimentally as well as theoretically, carried out on the basis of exact distance measures. With respect to experimental techniques this implies that protein structures are described in time or space by means of detailed distance information within the molecule, rather than protein structure formation being described by a phenomenological study of e.g. bio-chemical reactions. The detailed experimental techniques can either be X-ray Diffraction Crystallography, Nuclear Magnetic Resonance, NMR, methods, Circular Dichroism methods, Infrared Spectroscopy, Neutron Scattering etc., the first technique being the most established, the second dealing with problems of solvents, the third having advantages in particular structure analysis.

As far as theoretical studies are concerned the limitation of distance geometry approaches implies that protein dynamics and protein structure prediction are studied under the constraints of certain given experimental distance information or under the fulfillment of certain distances within the protein in order to limit the degree of uncertainty in protein structure analysis or structure prediction.

Although the problem of protein structure prediction from sequence is greatly reduced, given knowledge about certain inter-molecular distances, one should still be aware of the complexity in generating a full and detailed 3-dimensional protein structure from often very sparse, and at best, incomplete information about distances within a protein. In fact, many experiments can only give distance inequalities rather than exact real valued distances and often in a 2-dimensional form whereby the mathematical puzzle of generating the full 3-dimensional structure is, in principle, rendered unsolvable. However, there are various approximation techniques[2] described in here in chapter 3 and 5 that can circumvent these problems mostly with the use of computer simulation techniques. For a very detailed and thorough treatment of the mathematical problems in distance geometry analysis the reader is refered to the book by C. M. Crippen and T. F. Havel: "Distance Geometry and Molecular Conformations"[3, 4].

Apart from generation of 3-dimensional structures of proteins from distance constraints the distance geometry approach to protein structure analysis has also been understood in a wider sense to encompas energy potential methods based on distances and angles in the molecules. One approach[5] is to transform the problem of protein structure prediction into the problem of minimizing an energy function for an analogous spin glass system[8]

where the spin states correspond to protein configurations. This method is in line with distance geometry approaches in the sense that such energy function optimization basically implies satisfying a great number of distance constraints and simultaneously comparing sequences corresponding to these protein configurations. Somewhat in the same spirit is comparative protein modelling, performed by satisfying a set of spatial restraints and aims at making exhaustive enumerations of protein conformations. Another use of potential function is to identify correctly formed protein structures rather than predicting new structures from sequences. Moreover a whole new self-consistent molecular field theory is used to predict 3-dimensional structures of globular proteins.

A modern theme recurrent throughout modern protein research has been the concept of general classes of protein folds rather than describing specific protein structures. It is believed that proteins appearing in organisms are based on a limited repertoire of different core structures or folding motifs. In the past it has been common to classify proteins with respect to sequence similarity for evolutionary purposes or, most commonly, to group proteins with respect to their function so that, for example, proteases go in one group, immunoglobulins in another etc. The concept of protein folds[7] is, however, related to topological characteristica so that given folds belong to the same fold class if they share the same topological structure. A fold is a distinct geometrical domain of a protein (e.g. a cluster of super secondary structures), either of the whole protein or part of it. Often a necessary requirement, albeit not a sufficient, is that protein folds belong to the same class if they have more than 50% sequence identity. Proteases are for example devided into several fold-classes. A typical example of a fold class is the Tim Barrel class. One of the many questions concerning fold classes, and addressed in this book, is the problem of being able to identify them from sequence studies[26] and from distance geometry analysis. Another problem is to find an appropriate choice of parameters to link the different classes, such as a parameter for packing of secondary structures. This question arises especially when an entirely new protein, with practically no sequence similarity to any known structure, has to fit into or establish a relationship to one of the known classes. A very relevant question is in this context to ask how the most "extreme" classes could be characterized.

Connected to these protein folds is the new idea of "threading" [10, 11, 21, 13] meaning that protein sequences are being "threaded" through various different folding motifs in order to identify misfolded structures through an empirical evaluation function that can distinguish incorrect from correct folds. For reasons of simplicity folding motifs have been represented as linear profiles of local environmental properties independent of the type of fold being considered, e.g. secondary structures, at each residue in a known protein structure. Specific sequences can be given evaluation scores depending on preferences of the alligned residues for their respective environmental categories. Instead of representing folding motifs as linear profiles they can be represented as 2-dimensional contact matrices or as distance matrices[14, 15, 16, 17, 18] in the spirit of this forum.

Predicting which fold-class a given protein belongs to on the basis of its sequence can also be of great help in predicting distance matrices and a whole plan for predicting protein structures in the distance matrix approach could be deviced, perhaps leading to higher accuracy at lower sequence similarity than has yet been achieved. According to this plan[26] neural networks are trained on proteins from each fold-class exclusively, in order to develop an ability to predict distance matrices for new proteins belonging to the fold-class of the training set. There is good reason to believe that distance matrices can be fairly correctly predicted by neural networks for proteins homologous to the ones the network has been trained on[20]. The long term hope is to be able to develop prediction schemes for protein folds and (the inverse folding problem) to understand how much changes in their sequence is required for transforming a fold from one class into another. In more direct words one could ask how many substitutions are needed to give, for example Lysozyme, the functionality of a Cytochrome.

Protein structure determination is indeed an interesting and versatile forum for scientific discussions of the methodologies of bio-technology. All considered it is fair to say, concerning the goal of generating new protein structures, that while the experimental efforts focuss on still higher accuracy in protein structure determination the theoretical counter part of prediction methodologies is rather, till the present, achieving the gross features of protein structures at low resolution.
Nondum clivum exsuperavimus[22].

References

[1] D. B. Wetlaufer, "The protein folding problem", AAAS selected Symposium, Vol. 89, Westview Publisher, Boulder, USA (1984).

[2] J. Bohr *et al., J. Mol. Biol.*, **231**, 861-869 (1993).

[3] G. M. Crippen and T F. Havel, Distance Geometry and Molecular Conformation. Wiley, New York (1988).

[4] G. M. Crippen, *J. Mathematical Chemistry*, **6**, 307-324 (1991).

[5] R. A. Goldstein, Z. Luthey-Schulten and P. G. Wolynes, *PNAS*, **89**, 4918-4922 (1992).

[6] M. S. Friedrich and P. G. Wolynes, *Science*, **246**, 371-377 (1989).

[7] T. L. Blundell and M. S. Johnson, *Protein Science*, **2**, 877-883 (1993).

[8] S. Pascarella and P. Argos, *Prot. Eng.*, **5**, 121-137 (1992).

[9] D. Jones and J. Thornton, *L. Comp. Aid. Mol. Design*, **7**, 439-456 (1993).

[10] J. Novotny, A. A. Rashin and R. E. Bruccoleri, *Proteins*, **4**, 19-30 (1988)

[11] D. Eisenberg and A. D. McLachlan, *Nature*, **319**, 199-203 (1986).

[12] J. S. Fetrow and S. H. Bryant, *Biotechnology*, **11**, 479-484 (1993).

[13] L. M. Gregoret and F. E. Cohen, *J. Mol. Biol.*, **211**, 959-974 (1990).

[14] W. Taylor, *Prot. Eng.*, **4**, 853-870 (1991).

[15] J. T. Jones, W. R. Taylor and J. M. Thornton, *Nature*, **358**, 86-89 (1992).

[16] G. M. Crippen, *Biochemistry*, **30**, 4232-4237 (1991).

[17] M. J. Sippl, *J. Mol. Biol.*, **213**, 859-883 (1990).

[18] L. Holm and C. Sander, *J. Mol. Biol.*, **233**, 123-138 (1993).

[19] M. Reczko, H. Bohr, S. Subramaniam, S. V. Pamidighantam and A. Hatzigeorgiou, "Predicting what fold-class a protein belongs to." (Prot. Eng.) (1994).

[20] H. Bohr et al., *FEBS*, **261**, 43-46 (1990).

[21] O. B. Ptitsyn, R. H Pain, G. V. Semisotnov, E. Zerovnik and O. I. Razgulyaev, *FEBS Lett.*, **262**, 20 (1990).

[22] Citation from Seneca in "Epistolae moralis" (translation: Nobody has yet reached the summit (of solving the protein folding problem)) (around 60 a.c.).

2. The basic structural elements of proteins

In this chapter we shall only briefly introduce the basic notions in the field of protein structure analysis. There are highly recommendable textbooks in molecular biology that give introductions to protein science from many different perspectives[1, 2, 3]. Concerning the build up of the protein backbone by elementary atomic constituents there are only a few rules to learn and therefore it is very easy to acquire the basic knowledge about the assembly of a realistic, plastic protein toy model. These rules are also fairly easy to derive from a little quantum chemistry. However, there are more subtle facts about the basic assembly of the peptide chain, such as the broken chiral symmetry and the topology of the backbone ribbon, that has up to now not been fully explained. It turns out that there is an interesting differential geometrical study of the one-dimensional backbone chain to be undertaken and which could be related to the physical conditions of the pre-folding era of the the protein assembly in the ribosomes.

In this chapter we shall mostly be concerned with the more trivial and fully digestable facts of protein assembly from a toy model point of view. The first sections will be concerned with the atomic building blocks and their bonding geometry. The next short section will be about the few rules that are governing the backbone geometry and the last sections are about the most well-known ordered domains or substructures seen in ordinary folded proteins.

2a. The building blocks of proteins.

Proteins are long chain polymers of amino acids. They are linear, non-branched similar to polyethylene or polystyrene but with a much more versatile nature than the latter due to the very different type of amino acids involved. The 20 different amino acids have all an amino link, $(CO-NH)$, in common but each with a different radical (the side-chain) attached to a carbon atom termed the C_α atom. The amino, or more often called the peptide, links are connected to each other in a linear fashion such that the carbonyl end of one link is connected to the amino end of the next link and so that the resulting polypeptide chain (the protein without the side-chains) has a clear orientation.

Thus a protein molecule has a fairly easy structure with respect to its atomic constituents being (see figure 1 below) first a nitrogen atom followed by a carbon atom with a side-chain (one out of 20) attached to it and then finally followed by another carbon atom with an oxygen attached to it. The remaining sites are accupied by hydrogen atoms. This peptide unit is repeated typicaly several hundred times (for an average size protein) but mostly with a different side-chain attached to the C_α atom. The link between each amino acid connecting the carbonyl end with the next amino end has a partial doubble bonded nature that makes the peptide chain fairly ridgid.

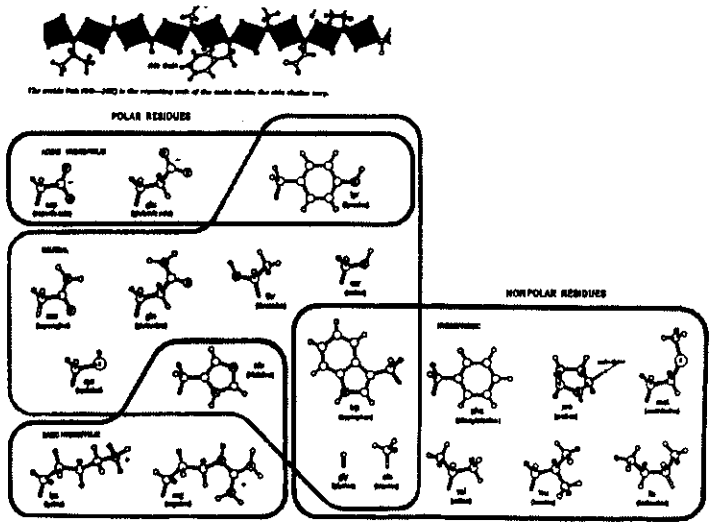


Figure 1 a,b. a: A picture of the unfolded peptide chain. b: All the 20 amino acids grouped.

The chemical activity of this polypeptide chain is for most parts controlled by the electrostatic nature of the different side-chains. These 20 common amino acids can be divided into polar and nonpolar where the polar ones can be either charged positive (basic hydrophilic) or negative (acidic hydrophilic) and neutral. The nonpolar amino acids are to a higher or lesser degree hydrophobic. The role of being hydrophilic or hydrophobic (turning towards or away from water molecules) becomes, as we shall see later, an important factor in the folding process when the protein is attaining its "native" active structure. Figure 1.b above is dipicturing all the common 20 different amino acids. These

amino acids have their side-chains sticking out from the peptide chain (often named the protein backbone) in a large variety of steric angles dictated by a complicated mixture of electrostatics and steric hindrance. A given protein with a fixed content of different amino acids will often attain a large set of different conformations, each being characterized by specific values of side-chain orientations that are important for the proteins functionality.

Before getting into the detailed geometrical structure of the protein molecules in the next subsection we shall end this paragraph with an appreciation of the enormous variety or versatility that the proteins with the building blocks described above provides. The variety of proteins is far bigger than the amount of atoms in the whole universe. Take for example an average size protein of 150 amino acids. Since there are 20 amino acid types (in common use) this gives a variety of 20^{156} configurations and if we also take into account all the different conformations each amino acid can attain we arrive at a number that is many orders bigger than the number of atoms in the universe (which could be estimated to be around 10^{80}, only counting visible matter). The size of the variety of protein configurations is relevant to the discussion of how the biological evolution can transverse such a huge state space and still come out with successfull spieces. Later we shall actually see that there is a way out of this dilemmma since we in chapter 5[20] can show that evolution of protein dynamics at certain stages, and to a certain extend also the biological evolution progresses like a neural network with an associative memory that can learn from mistakes.

2b. Chemical bonding and the implication to protein structure.

In order to understand the nature of the chemical bondings in the peptide chain it is illustrative to look at similar but simpler examples of chemical bondings in pure carbon hydrates. From the study of the molecular orbitals in methane and ethylene one can get a quite clear understanding of the possible bondings that are associated with carbon atoms. There are for example very pedagogical drawings on these molecular orbitals in the book on protein structure by Dickerson and Geis[2].

A few general facts about molecular orbits in the relevant atoms of the proteins should first be mentioned. When, for instance, carbon, nitrogen and oxygen atoms form bonds they ordinarily use their $2s, 2p_x, 2p_y, 2p_z$ atomic orbitals. Carbon has four valence electrons, nitrogen five and oxygen six beside the two electrons in the filled $1s$ orbital that does not participate in any bonding. Hydrogen has one valence electron. These valence electrons are not necessarily filling up the most straightforward orbitals but can be hybridized. In the case of methane, CH_4, the four orbitals of carbon, $2s$ and $2p$, do not combine directly with $1s$ electron orbitals in hydrogen but are observed to be tetrahedrally arranged around the carbon. In this case the carbon orbitals can be thought of as being combined (hybridized) to form four alike sp^3 atomic orbitals directed towards the corners of a tetrahedron and these then each combine with an $1s$ hydrogen orbital to build a $C-H$ bond with 2 electrons. Such σ type orbitals are cylindrically symmetrical about the $C-H$ axis with a bond energy of 99 $kcal/mol$ providing extra stability of this methane molecule compared to the situation of five isolated atoms, see figure 2a. below.

Figure 2a.: The molecular orbitals of Methane forming a Tetrahedron when hybridized.

In the case of the ethylene molecule, $CH_2 = CH_2$ we encounter another type of bond, the double bond. In this case the $2s$ and two of the $2p$ orbitals for each carbon atom hybridized to form three sp^2 orbitals, lying 120 degrees apart in a plane and resulting in four σ type $C-H$ bonds. In addition the two unused $2p$ orbitals are combined to form a different type of $C-C$ orbital, a π bond that is not cylindrically symmetrical about the $C-C$ axis.

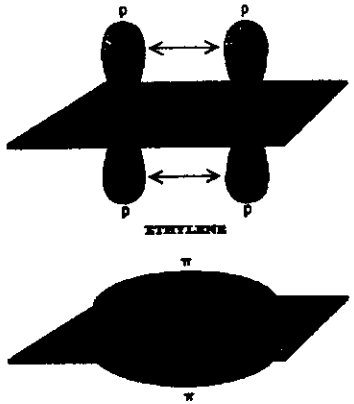


Figure 2b.: The σ and π orbitals of Ethylene.

It is something in between these two types of bonds we encounter in the protein peptide chain beside the σ bond. A simple example of that is when a carboxyl group, COO^- is ionized. Instead of the usual picture of having a double bond between the carbon atom and one of the oxygen atoms and a single bond between carbon and the negatively charged oxygen we rather have a partial double bond phenomena between the carbon atom and the two oxygens, a kind of resonance phenomena, such that the double bond electrons are being "delocalized" and the negative charge is "spread" out over the whole carboxyl group. Similar phenomena is seen in the protein peptide unit where there is a "resonance" phenomena between the $C=O$ double bond and the $C-N$ single bond with the double

bond electrons being delocalized to form a π type orbital that extend over all three atoms in the chain $O-C-N$. This provides extra stability to the peptide chain and gives this special geometry that is so characteristic for the protein backbone. This extended π bond in the peptide chain strongly limmits the number of degrees of freedom down to basically 2 variables (the dihedral angles, ϕ and ψ) for each amino acid. The energy gained from forming the peptide π bond is around $32kcal/mol$.

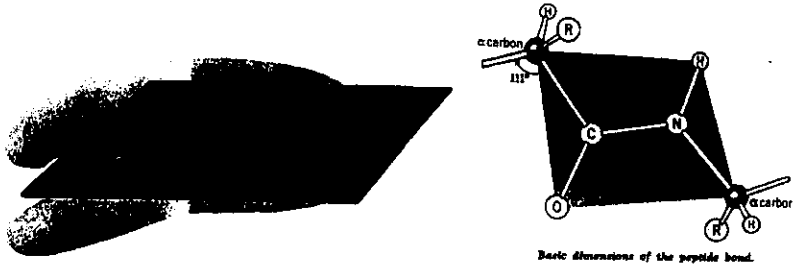


Figure 2c,d c: π bond across the peptide plane. d: The atoms in the peptide plane.

2c. Basic rules of the peptide chain.

In the last section we saw what influence the nature of the chemical bondings made on the geometry of the peptide chain or, as we shall call it from now on, the protein backbone. The extended π orbital across the nitrogen, carbon oxygen, $N-C-O$, atoms forced the repeated peptide units to lay in a plane. Since any three points will lay in a plane anyway we mean of course that this ridgit plane also includes the position of the hydrogen atom attached to the nitrogen. Also the two flanking C_α atomic positions are included in this peptide plan but due to their rotational degrees of freedom they are able to rotate around their peptide bond which define their dihedral angles (ϕ and ψ - the former at the $C_\alpha - N$ axis and the latter at the $C_\alpha - C$ axis). Apart from minor vibrational degrees of freedom these dihedral angles are the only conformational variables, two for each residue, that eventually are to be fixed by the side-chains mutual interactions and steric hindrance.

Before getting into that problem we shall first discuss the remaining reflection symmetry left over in the backbone geometry. If we look carefully at the peptide plane in between each C_α atom (see figure 2d above) we discover that even though we fix the peptide plane the C_α atom, oposite to the oxygen atom, can exchange place with the latter by a 180 degree rotation around the $C - N$ axis. We shall refere to the one depictured below as the "trans" configuration and the other to the "cis" configuration. It turns out that the "cis" configuration is slightly less favorable, probably due to a bend in the peptide chain that is caused by steric problems. Actually only in a few cases we encounter the "cis" configuration in known proteins and that is mostly associated with the Pro residue. Furthermore, if we look at the peptide chain from the CO across the C_α atom towards the amino group NH we can either have the side-chain sticking out towards the left side or the right side. The former is refered to as the left handed, or the L-form, of the amino acid and the latter to the right handed, or the R-form. In the biology we see around us we basically only find the L-form of the amino acids as if they once and for all have decided to be left handed. This apparant brake down of the reflection symmetry is strange because we on larger scales usually see a manifestation of the mirrow symmetry.

As we discussed before, the nature of the chemical bonds in the protein backbone left us with only two degrees of freedom, the didedral angles ϕ and ψ, around the C_α atoms for each residue. However, up to now we have mostly just considered the backbone geometry without the side-chains attached to each C_α atom. It turns out that if we also consider the side-chains we are ending up with a much more restrictive region of allowed values of these dihedral angles due to the various steric hindrances and mutual interactions that we have to consider for each side-chain. Included the side-chains actually makes it necessary to consider or include another dihedral angle around the $C_\beta - C_\alpha$ axis, usually denoted as the χ angle. If we plot each of the dihedral angles' alowed values for each of the residues in a protein in a 2-dimensional diagram we discover distinct features that tell us about the actual local structures that are present in the protein under consideration. For all known proteins we see in fact a universal pattern in these allowed regions that indicate the existance of common local structures, the so-called secondary structures in proteins. This brings us to the next subsection.

2d. Secondary Structures in Proteins

Much have been said about this topic in lecture notes on protein structure. We shall hence limmit ourself to only a brief introduction about the subject.

As we saw from last section there appear a universal pattern in the "local" structure of almost all proteins known up to now. The fact is that there appear distinct substructures in each protein that can be classified to be either helical, sheet like and a last category we denote as random coil. In the last class we include single loops or turns. These distinct substructures are stabilized by hydrogen bonds which in turn becomes the usual classifying criteria for these substructures. One could, however, also make the distinction of these substructures according to the dihedral angles allowed region. To see this we plot these regions in a 2-dimensional diagram where the x-axis contains the ϕ angles and the y-axis the ψ angles. Such a plot depictured below is called the Ramachandran plot and contains many features.. The white areas contain the alowed values and the dark the seldomly occuring values. Up in the right corner we encounter the sheet structures and more to the middle we find the helical structures. The most frequently occuring helical structure is the α with 3.8 residues per turn and that is mostly found to be right handed. The fractional number of residues per winding is due to the fact that it provides the helical element with maximal stability since the hydrogen bonds appear asymmetrical in that case (with respect to cylindical symmetry). In figure 3 a,b the helical and sheet

structures are depictures with detailed hydrogen bond patterns. There are proteins with only helical structures such as the four-helix bundle. The helical structure are also the only substructures in most globular protein. The other substructures, the beta sheets can occur both as parallel or anti-parallel patterns and are the dominant substructures in immuno globulin and most proteases.

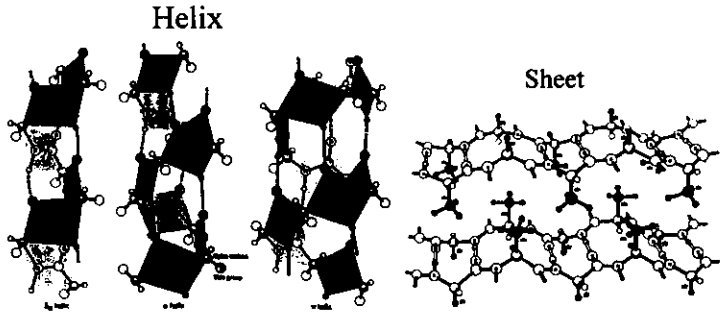


Figure 3a,b. a: 3 different helices, $3_{10}, \alpha$ and π helix.
b: Beta-sheet in Silk.

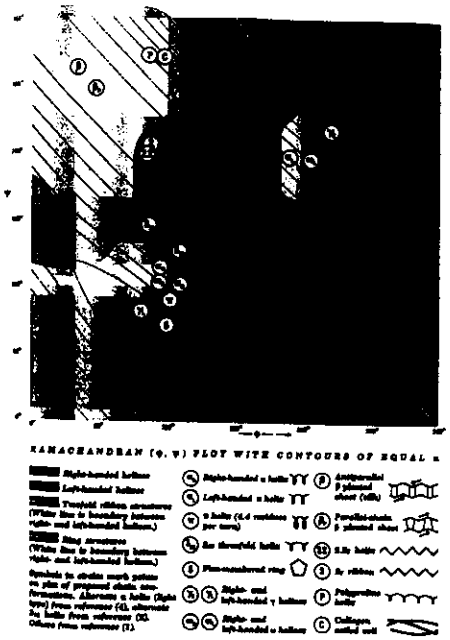


Figure 3c. c: The Ramachandran plot of substructures in the average protein.

These substructures are called the secondary structures because they occur on the second hierachical level of organisation, the first level being the sequence and the third level being the tertiary structures, the end product of the folding process. There has been an extensive effort in the field to produce prediction schemes that could determine the occurence of these structures from sequence information. These secondary structures will again arrange themselves into tertiary, or sometimes even into quartiary structures consisting of several domains of tertiary structures.

2e. Tertiary structure and distance geometry.

By the tertiary structure of a protein we mean the "native", folded, 3-dimensional structure (backbone as well as side-chains). In the case where the protein consists of ordered domains of folded subunits of secondary structures we call the 3-dimensional complete structure of the assembled domains for the quartiary structure. Later in the text we shall discuss various prediction schemes for the tertiary structure. Here we shall

like to introduce a convenient way of representing the 3-dimensional protein structures in the so-called distance geometry approach which is very much connected to the central issue in these lectures.

In the distance geometry approach we utilize predominantly the distance matrix which is defined as the 2-dimensional matrix whose elements are the actual distances between the atoms in the protein. In most cases we only include the distances between the C_α atoms and are then only concerned with the structure of the backbone. The matrix element $d_{i,j}$ is hence the distance between the position of the C_α atom of the i'th residue and that of the j'th residue. Since it is often only possible to measure distances approximately correct we often work with binary distance matrices. They are dependent on what value one choses as a threshold for defining the binary distances or (better) the distance inequalities. This means that if we chose a distance threshold of 8 $\AA$ all distances below 8 implies that the corresponding matrix element is 1 while distances above 8 make the matrix elements 0.

Below in figure 4 we show a binary distance matrix where the dark portions correspond to 1 and the light ones to 0. The amino acids are numbered along the x-axis as well as the y-axis. Since every amino acid is close to itself and its neighbours the diagonal and the next to the diagonal lines are dark. For pedagogical reasons we have made the next to the diagonal line white in order to be able to distinguish the areas close to the diagonal, i.e. the close neighbourhood around each amino acid. It is interesting and important that all the regular substructures such as the secondary structures can easily be determined from the distance matrix. For example the helical structures will be elongated dark areas (sausages) along the diagonal (extending out 4 lines from the diagonal when being alpha helices), while the anti-parallel beta-sheets are represented by bars orthogonal to the diagonal and sticking out as much as the length of the paticipating strands. The parallel beta-sheets are rods being parallel to the diagonal and detached from that.

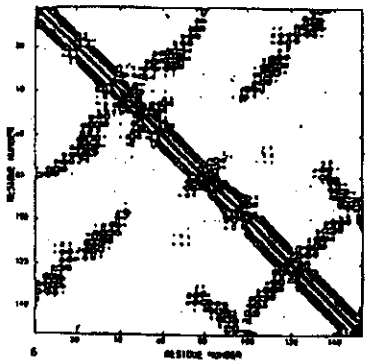


Figure 4: 2-dimensional plot of binary distance matrix of Rubredoxin, Threshold=8 Å.

References

[1] Schultz and Schirmer : Principle of protein structure, Springer Verlag (1978).

[2] R. Dickerson and I. Geis : Introduction to protein structure, W. A. Benjamin Inc. (1969).

[3] T. E. Creighton : Proteins, structure and molecular properties, Freeman (1984).

3. Structural Classification of Folded Proteins

1cm

In this chapter we shall introduce and discuss the concept of protein fold classes. Apart from mentioning the phenomenology of deviding proteins into fold classes (i.e. division with respect to appearance of structural domains) there is the quite successful story of predicting what fold class a protein belongs to just using sequence information. In the recent past the author has been involved in a project where Neural Network methodology has been used for predicting a protein fold class from the amino acid sequence. Using a hierarchical scheme of fold classification, a recurrent network was trained to construct features that characterize the membership of the fold class. At the highest level, a 4 class scheme was used and the network performed with a high accuracy of about 90%. In the case of fold classes defined by the presence of similar substructures or a certain percentage (30% - 60%) of sequence identity, the network determines for a set of 125 novel proteins the correct fold class (out of a total of 42 classes) to an accuracy of 81.6%. The prediction accuracy is well above 70% also for those test proteins with a maximal sequence identity of less than 25% amongst the training proteins, thus, establishing the robustness of the prediction. Such a scheme is very useful for assessing protein structural topology from sequence information alone and serves as a basis for further detailed homology modeling.

3a. Phenomenological look at Protein Folds.

It has recently been proposed[1, 2] that all the known 3-dimensional protein structures can be grouped into a smaller number of characteristic structural classes consisting of domains from homologous proteins with a similar topological configuration of their backbone. These structural domains or the so-called folds of the proteins were introduced in order to clarify the notion of structural similarity. Such fold classes could contain entire proteins or well-defined sub-domains of proteins. Pascarella and Argos[1] have used

topological similarity as a measure of fold class homology, while Holm and Sanders[3] have used similarity of distance matrices to determine fold class membership. Orengo et al.,[4] have reported a classification of proteins from the protein structural database into either 150 homologous folds or 112 analogous folds from structural comparison. Chothia[2] has postulated, based on known protein sequences and structures that the total number of fold classes is expected to be around 1000. While it is feasible to define membership to a fold class once the three dimensional structure of the protein is determined, efforts to predict fold classes only from sequences have met with little success. The exceptions are those where there is significant sequence homology between the protein whose structure is to be determined and one whose structure is established. Most frequently, sequences which have very little homology are known to belong to the same fold class. For example, the proteins Adenosine Deaminase(1add), Aldolase A(1ald), Aldose Reductase(1ads), the first domain of Cyclodextrin Glycosyltransferase(1cdg), Beta-Amylase(1btc), Endo-1,4-Beta-D-Glucanase(1tml), the second domain of Chloromuconate Cycloisomerase(1chr.A), second domain of Enolase(4enl), Glycolate Oxidase(1gox), Narbonin(1nar), first domain of Trimethylamine Dehydrogenase(2tmd.A), the second domain of Ribulose-1,5- Bisphosphatase(5rub.A), Triose Phosphate Isomerase(1tre.A), Tryptophan Synthase (1wsy.A) and Xylose Isomerase(6xia)[5], all belong to the "barrel" class and the sequence homology between any pair of these is insignificant.

In most definitions of fold classes, each member would have more than 50% sequence identity to each other although domains with far less sequence similarity could belong to the same class. It is important that each protein within a class would have a structure with a large topological similarity and a similar packing pattern to other members of the class. The details of the primary sequence in itself are less important. The notion of fold classes is important for predicting new protein structures using homology modeling. In homology modeling an unknown 3-dimensional protein structure is inferred from other known 3-dimensional protein structures whose amino acid sequences are similar to the sequence of the protein in question.

As we shall see later one can always make a crude classification of protein domains into what we call super fold classes by simply distinguishing them from their content of secondary structures. Such a super classification might actually also turn out to be deeply connected to the folding process and could also give rise to a measure of distance among the fold classes in the way that folds most different in secondary structure content are most far apart. We define thus four superclasses being: 1. The class of pure alpha helices (denoted α), 2. The class with only beta sheets (denoted β), 3. The class with alpha helices and beta sheets clearly separated (written $\alpha + \beta$) and finally, 4. The class of folds having alpha helices and beta sheets entangled (denoted $\alpha \cdot \beta$). These four classes are very well illustrated by the four prototypical proteins depictured below in figure 4a.

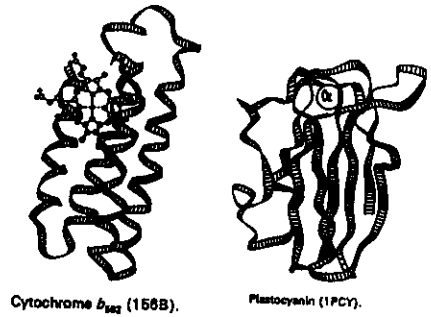


Figure 4a,b: The ribon representation of typical members from the super fold class α (left) and β (right).

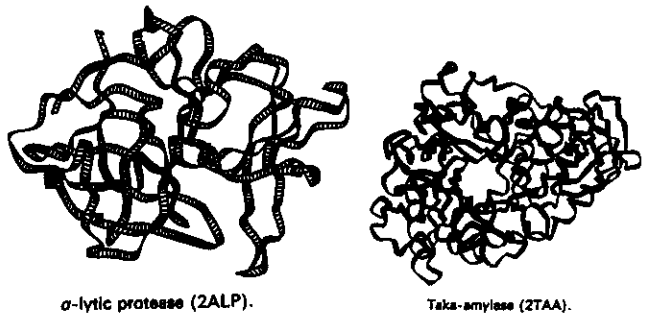


Figure 4c,d: The ribon representation of typical members from the super fold class $\alpha + \beta$ (left) and $\alpha \cdot \beta$ (right).

3b. Prediction schemes for protein fold classes.

It has been shown[6, 7, 8] that one can predict or model protein structures to high accuracy by using structural information from proteins belonging to the same fold class or family.

However, for protein sequences with very little homology to other proteins there exists no method that can predict the 3-dimensional structure to high accuracy from their

sequence data alone. On the other hand proteins with little sequence homology could be similar in structure to a whole class of other structures or domains. It is apparent that protein folding into a structure is coded by information that is not transparent from sequential similarity alone. Several techniques have been developed for inferring homology at the structural level from fold class membership. Some of these incorporate a combination of secondary structure prediction schemes, functional similarity, recognition of key structural motifs and use of machine learning methods for sequence-structure mapping[9, 10, 3, 11, 12, 13]. One method that successfully utilizes the information of the structure of homologous proteins uses artificial neural networks. The neural networks can be trained exclusively on homologous proteins as a basis for predicting a new protein structure from the corresponding sequence. Such a scheme is useful only when the protein in question has any relationship to any of the existing fold classes.

The proposed scheme, which consists of two steps, rests on the rationale that neural networks can be effectively trained to induce features from a system that characterize it. In the first step, a feed-forward neural network is used to determine the fold class of a protein from its sequence data. In the second step, the predicted fold class with its characteristic domains is used as input into a large recurrent neural network to predict the distance matrix for the protein. Such a distance matrix prediction should be accurate enough for constructing the 3-dimensional backbone structure for the protein, which can then be subsequently refined by side chain placement and molecular mechanics methods.

In the following section the neural network methodology for predicting the fold class of a protein will be discussed. In the subsequent section some results from neural network studies are presented. A hierarchy of fold classification is used in our scheme and this is shown to yield best prediction of fold classes.

3b1. Neural Network Methodology

The basic elements of an artificial neural network, the neurons, are the processing units which produce output from a characteristic non-linear function of a weighted sum of input data. A neural network is a group of such neurons and the neurons can communicate with each other through mutual interconnections. The network will gradually acquire a global information processing capacity for classifying data by being exposed (trained) to many pairs of corresponding input and output data such that new output can be generated from new input. If a set of input is denoted by $\{x_j\}$ and the corresponding output is denoted by $\{y_i\}$ the process at each neuron i in the network can be described by

$$y_i = f(\sum_j W_{ij} x_j + \eta_i) \tag{1}$$

where W_{ij} are the weights of the connections leading to the neuron i, η_i and f are the characteristics of the non-linear function for the neuron. As is obvious from the equation, such type of networks can be considered as a non-linear map between the input and output data.

The most straightforward type of neural networks employed for this study were feed-forward networks of the multi-layered perceptron type. These layers of neurons are referred as, mentioned in the consecutive order, the input layer, the hidden layers and the output layer. The reason for choosing this network among many other types is its ability to be generalizable to molecular biology data[14, 15, 16, 17]. The simple structure both with respect to processing of data and training is an additional advantage with such a network. The training was carried out using the back-propagation error algorithm[18] which is also the most commonly used. The training procedure is performed until a cost function C has reached a local minimum e.g. by a gradient descent. The cost function C is normally written as,

$$C = \frac{1}{2} \sum_{\alpha,i} (t_i^\alpha - z_i^\alpha)^2 \tag{2}$$

which is simply the squared sum of errors; t_i being the correct target value and z_i the actual value of the output neurons.

Various aspects of the use of perceptron layered nets have been studied to predict secondary structure or contacts in proteins on the basis of their sequence of amino acids. The network task has been to correlate sequence data input with the occurence of contacts between residues as output data. The input data of residue types are represented as binary numbers and the output as integers of e.g. residue contacts correlated to others. In each instance of training a vector of input values of residue types, where the size of the vector (window) represents the correlation among the residues, is to be related to a vector of output values of potential contacts corresponding to a specific residue (e.g. the one in the middle of the window) in the input vector. The network study was carried out on several types of network architectures, one being for example 60×20 (60 is the window size) input elements, 400 hidden neurons and 30 output neurons, the latter describing to which of the 30 residues preceding the residue in the middle of the input window a contact is formed.

It is important when utilizing neural networks to understand some basic facts of common knowledge about the architecture of the network in relation to the training. Firstly, the network should be dimensioned according to the training set, i.e. the number of adjustable parameters (the synaptic weights and thresholds) should not exceed the number of training examples. There is a heuristic rule that the number of training examples should be around 1.5 times larger than the number of synaptic weights. The ability to learn and recall learned data increases with the size of the hidden layer while the ability to generalize decreases with an increasing number of hidden neurons above a certain limit. This fact can clearly be understood when one considers the network as essentially a curve fitter between points depicting relations between input and output data in the training set. Therefore it is also easy to see that a network can be overtrained when the training process reaches the point where the spurious data points are memorized. Secondly, the training process and the construction of the training set is of great importance because the predictive power of the network is dependent on how clearly the training set is defined and how many patterns are exposed during the training.

The largest success in the present application was obtained with a training and construction procedure, called Cascade-Correlation[19]. This algorithm optimizes both the

weights in a feed-forward network and the number of hidden units by adding units during the training process see figure 5. The initial network contains only input and output units and is first trained using the normal delta-rule which is the special case of the back-propagation algorithm without hidden units. Thus the first phase of the training leads to the same solution that would be obtained by a perceptron and maps only those input patterns that may be separated linearly onto different output patterns. This linear part of the mapping may cover already a lot of input/output pattern pairs in the training set. To further reduce the error, one hidden unit, that is initially not connected, is added to the output layer. The weights leading into this unit are adapted by maximizing the correlation between the activity of this unit and the residual error occurring at each output unit. After this adaptation, all weights into this unit are frozen and the new hidden unit is connected to the output layer with all new weights set to 0. All weights connected to the output units are trained again to minimize the error function. The process of adding new hidden units that maximize the correlation between their activity and the remaining error at the output layer is repeated until the mapping has the desired accuracy. Since each new hidden unit is also connected to all existing hidden units, the network contains as many hidden layers as hidden units.

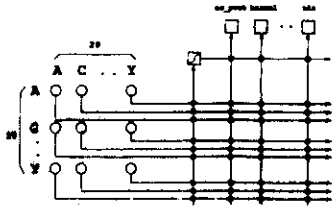


Figure 5.: A picture of the Cascade Correlation Network.

In order to evaluate the performance of the network, various statistical measures have been proposed. In the case of a dual valued output the Mathews coefficient, C_M[20, 21], was used to monitor the performance. If the two possible output values are denoted by 0 and 1 (signifying fold class membership or non membership) and if p is the number of correctly predicted examples of 1s, $\bar{p}$ the number of correctly predicted examples of 0s, q the number of examples of 1s incorrectly predicted and $\bar{q}$ is the number of examples of 0s incorrectly predicted then we define the coefficient C_M as:

$$C_M = \frac{p\bar{p} - q\bar{q}}{\sqrt{(p+q)(p+\bar{q})(\bar{p}+q)(\bar{p}+\bar{q})}} \tag{3}$$

For complete coincidence with the correct decisions (ideal performance) the measure is 1 and for complete anti-coincidence the value of C_M is -1. A poor net will give $C = 0$ indicating that it does not capture any correlation in the training set in spite the fact that it might be able to predict several correct values.

3b2. Neural Network Implementation

The actual neural networks for predicting fold classes are constructed from the SNNS (Stuttgart Neural Network Simulator) environment[22] and are of the feed-forward type. The networks are trained on a selection of proteins from each of 42 fold classes containing domain segments of proteins or often the whole proteins. The input representation for each protein domain is a 20×20 matrix containing the relative frequencies of dipeptides occurring in neighboring positions in the primary sequence of the domain. To calculate these frequencies, the number of occurrences of a dipeptide is counted in the protein sequence and divided by the total number of residues in that sequence. All protein domains are transformed this way into one input pattern of fixed size. Small insertions and deletions from the protein sequence cause only small changes in the dipeptide frequencies. The same holds true for rearrangements of larger elements in the sequence that do not change the local sequences. There are many cases where members of the same fold class differ mostly by permutations of sequence elements. Such permutations of the primary sequence lead to very similar dipeptide matrices which supports similar classification results. Each fold class is represented by one output unit which should have an activation close to 1.0 if the domain coded in the input layer is a member of that fold class. In all other cases the activity should be close to 0. When an unknown sequence is classified, the fold class corresponding to the largest activation at the output unit is assigned to the sequence. This is the usual "winner-takes-all" evaluation of the output of a classifier. In order to facilitate the interpretation of misclassifications all the fold classes were grouped into larger super-fold classes that have a natural one dimensional order inferred from physical properties of the folds. The super-fold class prediction and the fine grained classification should then assign classes that are close in this order.

As mentioned earlier, a general prediction of the 3-dimensional structure of a novel protein on the basis of its sequence of amino acids is likely to be successful by computational techniques, and especially neural networks, only when the fold class to which the protein belongs to can be determined first. A subsequent determination of the 3-dimensional structure of the protein can be obtained through a prediction of the distance matrix that represent the 3-dimensional backbone structure. The distance matrix prediction can be carried out by a neural network trained on the protein folds from the same fold class. In the next section, we describe the methods used to classify proteins into fold classes for training the network. Three distinct approaches giving a hierarchy of classification of folds are outlined.

3b3. Fold Classifications from Packing Analysis

Protein fold classifications from the literature have been used so far. At the most primitive level, we have classified proteins into large classes of alpha, beta, alpha+beta and alpha-beta proteins following Lesk and Chothia[23]. In a more detailed scheme, the classification of Pascarella and Argos[1], further enhanced by Walsh[24] has been utilized. In addition, a novel method for characterizing the fold topology of a protein is presented here. While the average density inside a protein is nearly a constant, the packing of residues is determined by the overall topology[25]. Arguably, all the information pertain-

ing to the three dimensional structure and hence the topology of the protein is contained at the most refined level in the distance matrix and at a less refined level in the packing density. We define the latter as the number of pairwise atomic contacts in the protein as a function of distance. The maxima and minima that occur in this packing density are very dependent on the nature of the overall protein fold. We have obtained this packing density for all the proteins in the database and classified them based on the similarity of the packing density features. Not surprisingly, this classification groups proteins into classes that are entirely similar to the earlier classification of Pascarella and Argos. It presents the 13 super-fold classes obtained from the packing density analysis. However, this method enables the creation of a coarse-grained set of folds that encompasses several fold class members of the Pascaralla and Argos set. This super-fold class delineation is used in training the neural networks. To our knowledge, this is the first effort to use a hierarchy of fold classifications to obtain sequence-structure correlation and prediction.

The frequency of contacts between atoms at various distances within a domain or a whole protein is plotted against the measure of distances in $\AA$ along the horizontal axis and the normalized frequency (occurrence) along the vertical axis. This results in a characteristic contact distribution for each structure of protein domains. Some structures are represented by a very broad distribution while others have a sharp delta-like distribution. The maxima in the normalized frequency of the distribution is a characteristic signature of the underlying lattice structure of the domain. For example a typical protease structure like a zig-zag lattice will have a distinct peak in the pair correlation distribution at the lattice spacing length. The position, τ, of the peak in the distribution was taken as a simple measure of the domain structure and all the domain structures were hence classified into distinct groups of folds using this criterion. Folds with the smallest values of peak positions, τ, turned out to be small peptides while intermediate ranges of τ usually could represent globular proteins. Large values of τ represented immunoglobulins and ac-proteases. Small values of τ thus signified little regularity and large values represented highly regular underlying lattice frames. The results of the performance of the neural networks using the data provided by the τ dependent fold class grouping will be presented in the following section.

3b4. Results for predicting Fold Classes

The main results in this paper are concerning the prediction of fold classes from sequence alone since that is the most novel element and distance matrix prediction from a homologous training set is well-known and is described elsewhere[17, 26]. The training set and testing set are both constructed from the data set of the 42 classes of domains used in ref.1. Roughly half of each fold class domains are used for training. The rationale for choosing the 42 classes from the Pascarella and Argos definition of folds, was to make certain that there are enough members in each class in order to perform a valid test. The fold class predictions are performed in three different levels of detail. The first classification uses the 4 super-fold classes based entirely on the secondary structure composition and arrangement in the proteins. The classifications are based on proteins containing the secondary structures, only alpha, only beta, one alpha and one beta domain and one containing a combination of alpha and beta secondary structure elements, respectively. In the second scheme, 13 fold classes each containing 3 members or more are defined by the packing density scheme described above. By using the τ measure we define a set of 13 super-fold classes that are used for prediction of the coarse fold class. In the third scheme, the full set of 42 classes is used for fine grained classification.

For the first case of 4 super-fold classes a network trained up to 97.2% accuracy and had a test score of 90.4% with an average Mathews coefficient of 0.81 which is a very high performance compared to other secondary structure content predictors. The matrix representing the actual prediction of the fold class membership is presented in Table I. The corresponding Mathews coefficients that represent the prediction accuracy is given in the last column. The analogous results where the 13 super-fold class set obtained from packing density analysis is used were presented in Table II. This fold classification gives a less accurate performance of training being up to 90% correct and the test being 65% correct which render this classification to be less useful for neural network based prediction schemes. The third case that is based on much better distributed classification yields a remarkable performance of 100% on the training set and with a test score of 78% in predicting a fold class correct on the basis of the sequence. Furthermore, adding the output of the 4 super-fold classes network to the input of the 42 class based network enhanced its performance to 81.6% on the test with an average Mathews coefficient of 0.7. The results are presented in the *permutation matrix* of Table III. In Tables III the number in row i and column j counts all cases where a test protein that is predicted to be in class j in fact belongs to class i. Optimally, all test cases should be counted on the main diagonal of the permutation matrix. For the case of the 42 fold class prediction, the relation between maximal sequence identity of a test sequence to the sequences used in the training set and prediction accuracy is given in Figure 6. The four points at 25, 50, 75 and 100% sequence identity defining the solid line give the average prediction accuracy for those test cases that have a maximal sequence similarity between 0 and 25% , 25 and 50%, 50 and 75%, 75 and 100% to the training set. The fold class prediction is still more than 71% correct for those test sequences with 0 to 25% sequence identity to the training set, which is an important property for a large scale application of this prediction method.

TABLE I

true	predicted				Corr
	α	$\alpha+\beta$	α/β	β	
α	22	1	2	2	0.807
$\alpha+\beta$	2	7	1	1	0.678
α/β	0	0	24	1	0.904
β	1	1	0	60	0.905

TABLE II

true	predicted 0	1	2	3	4	5	6	7	8	9	10	11	12
0:	0	0	0	0	0	0	0	0	0	0	0	0	1
1:	0	0	0	0	0	0	0	0	0	0	0	0	0
2:	0	0	4	0	0	0	0	1	0	0	0	0	0
3:	0	0	0	0	0	0	0	1	0	0	1	0	0
4:	0	0	0	0	0	0	0	2	0	0	0	0	0
5:	0	0	0	0	0	0	1	0	0	0	0	0	0
6:	0	0	0	0	0	0	10	2	1	0	0	0	0
7:	0	0	0	0	0	0	3	27	3	0	2	1	0
8:	0	0	0	0	0	0	1	2	6	0	0	1	0
9:	0	0	0	0	0	0	0	0	0	2	2	0	1
10:	0	0	0	0	0	0	0	0	0	1	6	1	0
11:	0	0	0	0	0	0	0	2	0	0	2	4	0
12:	0	0	0	0	0	0	0	0	0	0	0	0	0

STATISTICS (91 patterns)
right : 64.84 % (59 pattern(s))

correlation coeficcient for class 0: 0.000000
no testpattern for class 1 present.
correlation coeficcient for class 2: 0.889272
correlation coeficcient for class 3: 0.000000
correlation coeficcient for class 4: 0.000000
correlation coeficcient for class 5: 0.000000
correlation coeficcient for class 6: 0.665020
correlation coeficcient for class 7: 0.565617
correlation coeficcient for class 8: 0.550617
correlation coeficcient for class 9: 0.495656
correlation coeficcient for class 10: 0.538666
correlation coeficcient for class 11: 0.492922
no testpattern for class 12 present.

TABLE III

	true folderclass	predicted folderclass: α												α+β					α/β							β												Corr
1	helix_bndl	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	
2	cytc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
3	hmr	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.718
4	wrp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
5	globin	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
6	lsm	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.787
7	crp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.704
8	ca_bind	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.397
9	tln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.701
10	cts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.008
11	pap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
12	cra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
13	cpp	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
14	wga	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
15	sns	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.704
16	plipase	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
17	gap	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.704
18	inhibit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
19	xia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
20	kinase	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
21	binding	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
22	barrel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
23	eglin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
24	pgk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
25	dfr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
26	abt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
27	a_prot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.000
28	cpa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.000
29	ferredox	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.704
30	fxc	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.704
31	pti	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.704
32	rdx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.704
33	virus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1.000
34	virus_prot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.866
35	gcr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1.000
36	igb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.862
37	il	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.870
38	ac_prot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.558
39	tox	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.922
40	plasto	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1.000
41	lln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.813
42	hoe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.573

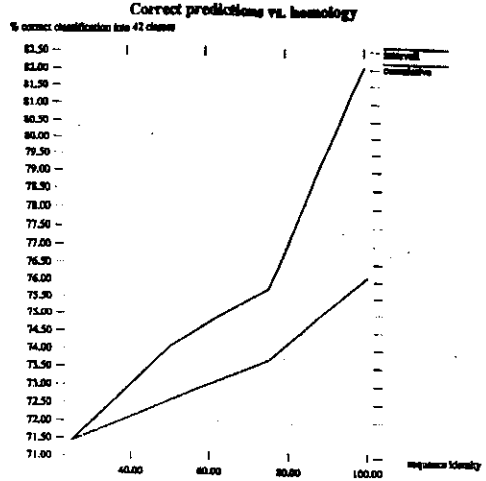


Figure 6.: This figure shows the correctness in the prediction versus homology measured in
intervals (upper curve) or accumulatively (lower curve).

3b5. Discussion

An artificial neural network system has been constructed to classify 3-dimensional protein structures by predicting what fold class they belong to on the basis of their sequence alone. Once that is decided one may predict the corresponding distance matrix e.g. by recurrent neural networks that are trained on proteins from the chosen fold class and subsequently construct a 3-dimensional structure for the test protein by a minimization procedure. The networks appear to train surprisingly well (81.2% correct and an average Mathews coefficient of 0.7) on the task of predicting fold classification, even for test proteins with a maximal sequence identity of less than 25% to all training proteins.

The best results for training and predicting fold class membership was obtained using the 4 class scheme. Amongst all the proteins tested 90% prediction accuracy was achieved. Most surprisingly, beta stranded domains and proteins were predicted with high accuracy. Interestingly, it seems that neural networks are able to achieve greater than 80% accuracy in predicting the fold classes as compared to their prediction of the secondary structures of peptides[27]. One explanation for that may be due to the postulate that around 70% of the secondary structures found in the native structure are formed at an early stage (i.e. msec) of protein folding and thus without training the network on intermediate structures the performance will never surpass the 70%. The determination of the folds is similar to the determination of the topology of the protein backbone and that, on the other hand, depends only on the overall packing of secondary structural elements. Furthermore the new classification of folds that we proposed is partially dependent on the content of secondary structures. Low values of the τ parameter represent alpha-rich fold classes and high values of τ represent beta-rich fold classes.

References

[1] Pascarella, S., Argos P. (1992) Protein Engng., 5, 121-137.

[2] Chothia, C. (1992) Nature, 357, 543-544.

[3] Holm, L., Sander, C. (1993) J. Mol. Biol., 233, 123.

[4] Orengo, C. A., Flores, T.P., Taylor, W. R., Thornton, J. M. (1993) Protein Engng., 6, 485-500.

[5] Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., Weng, J. (1987) Protein Data Bank. In Crystallographic Databases - Information Content, Software Systems, Scientific Applications., Allen, F. H., Bergerhof, G., Sievers, R., Eds., 108-132, Data Commission of the international Union of Crystallography, Bonn/Cambridge/Chester.

Bernstein, F.C., Koetzle, T. F., Williams, J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977) J. Mol. Biol., 112, 535-542.

[6] Bassolino-klimas, D., Bruccoleri, R. E., Subramaniam, S. (1992) Protein Science, 1, 1465-1476.

[7] Viswanathan, M., Anchin, J. M., Droupadi, P. R., Mandal, C., Linthicum, D. S., Subramaniam, S. (1994) Molecular Biophysics Technical Report UIUC-BI-MB-94-02.

[8] Goldstein, R. A., Luthey-Schulten, Z. A., Wolynes, P. G. (1992) Proc. Natl. Acad. Sci. USA, 89, 9029-9033.

[9] Ioerger, T. R., Rendell, L. A., Subramaniam, S. (1993) In Proc. First International conference on intelligent systems for molecular biolog y, AAAI press, Menlo Park, CA, 198-206.

[10] Bryant, S. H., Lawrence, C. E. (1993) Proteins: Struct. Func. Genetics, 16, 92-112.

[11] Sippl, M. J. (1990) J. Mol. Biol., 213, 859-883.

[12] Johnson, M. S., Overington, J. P., Blundell, T. L. (1993) J. Mol. Biol., 231, 735-752.

[13] Jones, D., Thornton, J. (1993) J. Comput. Aided Mol. Design, 7, 439-456.

[14] Qian, N., Sejnowski, T. J. (1988) J. Mol. Biol., 202, 865-884.

[15] Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, L., Olsen, O. H., Petersen, S. B. (1988) FEBS Lett., 241, 223-228.

[16] Holley, L. H., Karplus, M. (1989) Proc. Natl. Acad. Sci. USA, 86, 152-156.

[17] Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B., Petersen, S. B. (1990) FEBS Lett., 261, 43-46.

[18] Rumelhart, D. E., McClelland, J. L. (eds.) (1986) Parallel Distributed Processing, MIT Press, Cambridge, MA.

[19] Fahlman, S. E., Lebiere, C. (1990) In "Advances in Neural Information Processing systems II", D.S. Touretzky, (Ed.) Los Altos, CA: Morgan Kaufmann, 524-532.

[20] Mathews, B. W. (1975) Biochem. Biophys. Acta, 405, 442.

[21] Stolorz, P., Lapedes, A., Xia, Y. (1991) "Predicting Protein Secondary Structure Using Neural Net and Statistical Methods", Los Alamos Preprint LA-UR-91-15.

[22] Zell, A., Mache, N., Sommer, T., Korb, T. (1991) In Proc. Applications of Neural Networks Conf., SPIE, Aerospace Sensing Intl. Symposium, Orlando Florida, 1469, 708-719.

[23] Lesk, A. M., "Protein Architechture A Practical Approach" (1991) Oxford University Press. Oxford.

[24] Walsh, L. L. (1992) Protein Science 1:5, Diskette Appendix. Walsh, L. L. (1994) personal communication.

[25] Kauzmann, W., Moore, K., Schultz, D. (1974) Nature, 248, 447-449.

[26] Reczko, M., Bohr, H. In "Protein Structure by Distance Analysis" H. Bohr and S. Brunak (eds.) (1994), IOS Press, Amsterdam, 87-97.

[27] Bohr, H., Goldstein, R., Wolynes, P. G. (1992) AMSE Periodicals, Modelling, measurement and control, C, 31, 55.

Table Captions

Table I. Neural network Prediction of the four super-fold classes based on the secondary structure content alone. The matrix elements represent the number of correctly predicted protein domains in each fold class. The last column is the Mathews coefficient (see text).

Table II. The matrix representing the number of protein domains, belonging to a fold class defined according to the packing densities, that are predicted correctly. The last column represents the Mathews coefficients (see text) for the predictions. The dashed entries indicate non-availability of test set proteins.

Table III. Neural network prediction of the fold classes from Pascarella and Argos's set of 42 fold classes. The matrix elements represent the number of correctly predicted protein domains in each fold class. The last column is the Mathews coefficient (see text).

4. A Statistical Mechanical model for formation of protein fold classes.

In these sections we shall discuss ways of constructing fold classes from purely theoretical means by applying statistical mechanical spin models and with that technique try to estimate the total number of possible fold classes. We shall in details present a model for the dynamics in the early stages of protein folding leading to a structural classification of protein folds. The overall goal is to determine what possible topological fold-classes a protein can adapt in the beginning of the folding process. The model turns out to be a relative simple spin system with angular variables and simulations results in the order of thousands fold-classes.

4a. A simple Hamiltonian for distinguishing protein fold classes.

Protein folding is an intriguing problem. How can nature fold a macromolecule containing thousands of atoms into unique compact structures without testing the whole phase space of configurations?

Recently it has been appearant that there is a finite set of distinct fold-classes containing protein domains (a whole or part of a protein) with a distinct topological structure. Such a set has been estimated to consists of around one or two thousands of fold classes of which around one hundred structures of protein domains already have been determined. We shall here try to understand how such distinct topological fold classes can arise from an appropriate physical theory of the dynamics of stages of protein folding.

The dynamics in the early stages of protein folding is not in this context meant to include interactions from a specific surrounding media or other proteins such as chaperones. The information necessary to set-up such a dynamical framework is hence supposed to be predominantly sequence data of the primary protein structure leading to a specific structure and an important feature is to explain the extension of the dynamics of a 1-dim. start to a particular 3-dim. configuration.

Before continuing we shall like to refer to some recent experimental investigations on folding of simpler proteins, such as lysozyme, by NMR-techniques carded out by Dobson et al. In the picture arising from those experiments and which we shall base our analysis on, the topological folds are being formed early in the protein folding processes when many of the secondary structure elements, and especially the alpha-helices are present. In later stages the more complicated structure of especially the parallel beta-sheets are formed and more accurate docking and detailed side-chain conformations are getting in place. Such a scenario which consists of a rapid formation of the topology and roughly determines the fold class type is essential for our analysis and in fact also makes sense evolutionary wise since it costs too much to end up in a wrong topology and unravel the wrong knots.

We shall here investigate a new approach in which we consider the weakest forces first - or at an early stage. This sounds counterintuitive. However, there are several examples in physics, where it is the weak forces which determine the gross structure and the strong forces which determine the details. A well know example, which in fact will be close to the present approach, is the Heisenberg ferromagnet. It represents a system described by spin vectors $\mathbf{S} = (S_x, S_y, S_z) or (S, \theta, \phi)$, which are completely isotropic in space. The spins interact with and isotropic interaction, i.e. it is invariant for any rotation of the reference frame

$$\mathcal{H} = -J \sum_{ij} \mathbf{S}_i \cdot \mathbf{S}_j \; - \; \lim_{h \to 0} \; h \sum_i \hat{\mathbf{e}}_z \cdot \mathbf{S}_i, \tag{4}$$

where $J > 0$ is the important, large interaction parameter which dictates that the grounds that, i.e. the lowest energy state has all spins parallel. However, it cannot determine the direction in which they point. The full rotational symmetry is broken by the infinitesimal field $\mathbf{h} = h\hat{\mathbf{e}}_z = (0, 0, h)$. Below a certain temperature T_c the strong interactions causes a further break in the symmetry between spin states pointing up or down the $z-$axis, and domains thereof are formed.

In the following we shall construct a minimal model for protein folding in order to establish a vocabulary and a language in which these can be described and subsequently classified. We shall start by assuming that the proteins form a discrete or quasi continuous tape, which although being flexible, can transfer information about angular twists along itself to some distance. The protein is characterized by a linear information in terms of the twenty letters in the amino acid alphabet. This information is sufficient for nature to determine the folding. Let us suppose that there are no *long range* forces, i.e. unless parts of the tape are very close in space there is no interaction; however if that happens the strong *short range* interactions get in to operation, for example the Hydrogen-bonds, Van der Walls and/or other chemical bonds. Imagine we start the protein in a fully extended state. This is not necessarily linear, but simply such that no part is close to any other. There are a very large number of such states, compared to the unique close folding which we know is the 'ground state'. The extended state therefore corresponds to a high temperature configuration of the system. If the tape has equal surface tension on both sides, the tape will be approximately flat. Now suppose the short range forces along the tape change the surfaces tension of either side, locally. This will provide a bending force on the tape. At a given temperature a section with uniform, differing surface tension will curl up as a spiral. We shall understand this as the so-called α-helix. This *secondary structure* is more stiff and rod like than the original tape. It is important to notice that this curling up can be done without any global turning or twisting of the whole tape. It just give rise to a contraction of the overall extent of the tape. This does not result in the formation of any new crossings, but rather a straightening of the remaining tape. In respect to the analogy to the Heisenberg model eq.(4) it corresponds to the formation of small ferromagnetic domains. At the considered temperature let us assume that the extended tape is sub-structured according to the underlying amino-acid letter code into two groups of secondary structures. One, which we denote by large letters $A, B, C, ...$, representing the described α-helix and also potential pieces for the formation of some β-sheets. The latter cannot be described at this temperature since they require the short range forces between different parts of the tape and not just forces along the tape. The second group which we denote by small letters $a, b, c, ...$, consists of the remaining connecting pieces of the tape, the unconnected strands and the turns. All elements are assumed to be approximately linear. Each element is connected by a 'hinge' which is characterized by a direction in space, perpendicular to the plane in which the two joining elements can rotate. Using a spin $\mathcal{S}_i$ for this description we can both define the direction and the sense of the bend between the two elements. We then make the crucial assumption that each element is sufficiently rigid to define the relative optimum direction of the spins attached to its ends. We then symbolize the protein as the sequence of secondary structures with preferential bending forces acting between them

$$a\, \mathcal{S}_1\, A\, \mathcal{S}_2\, b\, \mathcal{S}_3\, B\, \mathcal{S}_4\, c\, \mathcal{S}_5\, C\, \mathcal{S}_6\, d\, \cdots . \tag{5}$$

It is at this level we shall attempt to classify the various protein foldings. We assume that the underlying linear information defines the subdivision into the secondary structures and the preferred angle between the elements. We are now ready to formalize the model in order to be able to make computer simulations and predictions of folding classes. We remark that the description is independent of the lengths of the elements. It is also independent of the position in space and of interactions between the elements. This is not simply a lattice model, but in principle it can be made much more general with arbitrary angles and lengths. At a later stage we will include such interactions between the elements of the first group $A, B, C, ...$, in particular the potential β-sheet elements. The final evaluation if a folding is energetically favorable will then be judged on the basis of the strong, short range forces. However, we expect the angular forces to determine the general dynamical features early in the folding, which is the focus of this study, leaving the optimalization to the strong forces for later stages.

The angular model

The model described above is still too complicated to be practical. At a first level it is probably not important to allow continuous variations in the possible angles so we assume only one allowed angle, and the value of the angle is not essential for the argument in the first stage. For ease of representation we therefore choose this as 90°, perhaps also including the value 0°. Let us traverse the protein represented by eq.(5) from left to right. Each element $P = A, B, C, ...$ has then a direction $\hat{\mathbf{e}}^P_\alpha$ in a artesian coordinate system with $\alpha = x, y, z$. Similarly each element $p = a, b, c, ...$ is characterized by $\hat{\mathbf{e}}^p_\alpha$. The structure is given by the sequence of spin vectors $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4 \cdots$. The spins have unit lengths and may each point in either of the six directions $\pm x, \pm y, \pm z$. If we consider only the 90° and 0° turns a unique description for the orientation between two elements is given by

$$\begin{aligned} \hat{\mathbf{e}}^A_\alpha &= \hat{\mathbf{e}}^a_\alpha \times \mathcal{S}_1 + (\mathbf{e}^a_\alpha \cdot \mathcal{S}_1)\mathbf{e}^a_\alpha, \\ \hat{\mathbf{e}}^b_\alpha &= \hat{\mathbf{e}}^A_\alpha \times \mathcal{S}_2 + (\mathbf{e}^A_\alpha \cdot \mathcal{S}_2)\mathbf{e}^A_\alpha, \quad \text{etc.} \end{aligned} \tag{6}$$

The cross product takes care of the 90° turns and the dot-product of the possibility of straight continuation and the rather unlikely return-on-it-self possibility, corresponding to the turn angle 0° and 180°. Since all angles are either ±90° or 0° (and 180° which we neglect) there is no overlap from the terms in eq. (6). It is now clear that the folding is uniquely described by the sequence and state of the 'hinge' variables, the spins $\mathcal{S}_i$ and the element variables $\hat{\mathbf{e}}^P_\alpha$ and $\hat{\mathbf{e}}^p_\alpha$. A given sequence of spins $\mathcal{S}_i$ and start direction $\hat{\mathbf{e}}^a_\alpha$ is a rigid building prescription, by which any later element direction $\hat{\mathbf{e}}^i_\alpha$ is exactly determined. (If we give length information on each element, the precise position in space is in fact given). However, this is too strict we want just to give building guide lines. For an element of

group one, optimally surrounded by parallel spins ($\uparrow A \uparrow$), let us say it gains an energy J if it is the case, gains nothing if they are perpendicular ($\uparrow A \rightarrow$) and pays an energy $-J$ if the spins are anti-parallel ($\uparrow A \downarrow$). If the spins should have a right twist we would give an energy gain K for the right twist, 0 for parallel or anti-parallel and $-K$ for the wrong, left-twist. We can define similar energy conditions for elements of group two, with possibly different, and lower energy values j, k. We then form a linear chain of these energy variables, describing the preferred state of its surrounding spins f.ex.

$$j \Updownarrow K \Updownarrow k \Updownarrow (-K) \Updownarrow j \Updownarrow J \Updownarrow (-j) \Updownarrow \cdots, \tag{7}$$

where $\Updownarrow$ represents any of the possible six spin directions for the 'hinge' spins. We notice this is a more flexible description than eq. (5). The structure is now determined by the interaction constants sequence $\mathcal{J}_1, \mathcal{J}_2, \mathcal{J}_3, \mathcal{J}_4, \cdots$, given in eq. (7), as an example, as $j, K, -k, -K, j, -j, \cdots$. This gives a unique best set of the spin variables $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4 \cdots$. ¿From those the ground state can be constructed from 6. If that is all we want we could just as well take all constants equal in magnitude, say equal to one, leaving just the signs. This would be kind of interaction 'spin' variables $\mathcal{J}_i$. However we could also consider 'wrong' turns and then it would be nice to have different energy parameters here to give us the energy cost for that. A change in a spin ($\mathcal{S}_i$) direction at a junction p has the dramatic consequence of rotating the entire remaining pieces of the protein around this junction. We shall assume that there is no inertia and no steric hindrance in doing so. Expressed in an other way we do not care how the system has arrived at any state which we can measure the energy for. The energy of any state of the spins is given by an interaction Hamiltonian

$$\begin{aligned}\mathcal{H} = &- \sum_{P=2n+1}^{2N-1} (J_P \mathcal{S}_P \cdot \mathcal{S}_{P+1} + K_P \mathcal{S}_P \times \mathcal{S}_{P+1} \cdot \hat{\mathbf{e}}_\alpha^P) \\ &- \sum_{p=2n}^{2N} (j_p \mathcal{S}_p \cdot \mathcal{S}_{p+1} + k_p \mathcal{S}_p \times \mathcal{S}_{p+1} \cdot \hat{\mathbf{e}}_\alpha^p).\end{aligned} \tag{8}$$

This now looks like $\mathcal{S}_0, \mathcal{J}_0, \mathcal{S}_1, \mathcal{J}_1, \mathcal{S}_2, \mathcal{J}_2, \mathcal{S}_3, \mathcal{J}_3, \cdots, \mathcal{J}_{2n+1}, \mathcal{S}_{2N+2}$. One may start by fixing $\mathcal{S}_0 = z$ f.ex and $\mathbf{e}_\alpha^0 = x$, the rest should then follow from eq. (6). In eq. (8) the index n is the summation index running from $n = 0$ to N, where $2N + 3$ is the total number of spins (the two in the ends can be disregarded (should find nicer formulation)) The constant J_P determines the energy for having the spins at the ends of a group one element P to have parallel or anti parallel spins in the x, y or z- direction. This could be added in a more general treatment. The constant K_P determines the energy for having the spins perpendicular or 'anti'perpendicular to each other. We have here disregarded the cases with angle 0°, and cases with the spins along the element direction. For the α-helix it is rather clear that the interaction between the spins will be simply related to the number of amino acids which the helix is formed by. So the ground state is given by the sequence $\mathcal{J}_0, \mathcal{J}_1, \cdots, \mathcal{J}_{2n+1}$. Each have four possibilities $\pm J, \pm K$ or $\pm j, \pm k$. That gives, I think, 4^{2N-1} possibilities for a chain of $2N + 1$ elements. One could plot out all the states and discard the most open ones. That would leave us with the most probable cases (classes). The information is the same if we specify the spin in eq. (6) from the outset. However, the $\mathcal{J}_0, \mathcal{J}_1, \cdots, \mathcal{J}_{2n+1}$ sequence is more directly connected to the amino acid chain information. We can also judge energy differences between good and bad foldings for the same sequence. We need a simple 'compactness' measure. One could try the following (at first):

Assume all lengths are equal (to 1, say; i.e. the same as as lattice model at first). Find the site coordinates of the spins $\mathcal{S}$. $\mathcal{S}_0 = (0, 0, 0)$, $\mathcal{S}_i = (x_i, y_i, z_i)$. 1) find the center of gravity from the coordinates

$$\frac{1}{2N+3} \sum_i^{2N+3} \alpha_i = < \alpha >, \alpha = x, y, z. \tag{9}$$

2) find a spread factor, for example one could calculate the moments around the center of gravity

$$F = \frac{1}{2N+3} \sum_i^{2N+3} [(x_i - < x >)^n + (y_i - < y >)^n + (z_i - < z >)^n]. \tag{10}$$

Let us assume small F is good and selects a desirable set of the interaction variables $\mathcal{J}_0, \mathcal{J}_1, \cdots, \mathcal{J}_{2n+1}$. One could now choose a 'temperature', and do normal Monte Carlo simulation of $\mathcal{J}_i$ to find the states at a given 'temperature'. The smaller the more compact configurations. The F energy and the turn energies could be put into play simultaneously (allowing wrong turns at an energy cost (later)). We can now further chose different lengths. First, f.ex one for group one and another (shorter probably) for group two. Further, one could consider the robustness of the classes under a distribution of lengths around these values. Finally it should be possible to relax the right angle conditions as well.

A certain set of amino acid 'words' $LMN \cdots RST$ give an interaction constant f.ex. $+J$ for parallel spins. It is clear that the reverse order $TSR \cdots NML$ also gives $+J$, i.e. belongs to the same class. This may not be favorable. It is a well know problem in neural net work theory where a Hamiltonian description normally requires that the interactions are independent of the bond direction $J_{i \rightarrow j} = J_{i \leftarrow j}$. Let us first make the assumption that this is unimportant; subsequently one can consider when it is important. This a typical classification indicator. The description eq. (6) contains the direction information and so do the cross-product terms in eq. (8), however this information is lost in the dot-product terms. It might be possible to change J to K by moving the 'hinge' one or more unit, and change even their signs. It corresponds to choosing a neighboring amino acid word for example $LMN \cdots RSTU$ followed by an other shorter one e.g. $MN \cdots RST$. This will be important in the final optimization process, where the strong forces come into play. The corresponding energy should be attached to the element variable as will be further discussed below. The second sum in eq.(8), similarity, represents the turn energy for the spins surrounding an element of group two. Here we have treated all group elements equally. In principle there could be many more parameters than the four introduced. However, the main purpose here is to schematize the problem, still retaining the main physics. The constants J, K, j, k can, as we have seen, in principle be related to the amino acid lettering sequence (ref to Liebmann). The introduction of the 'hinge' variables thus enables us to fairly explicitly write down the folding energy.

However, it is more convenient to introduce new variables attached to each element describing the state of the spins surrounding it, mainly in preparation for the final optimization process. Consider element of the first group, say A. If the spins $\mathcal{S}_1$ and $\mathcal{S}_2$ are perpendicular we take the new element variable to be $\mathbf{S}^A = (\mathcal{S}_1 + \mathcal{S}_2)/\sqrt{2}$. By this we simplify the phase space by a factor of two, since we only say the element is twisted but we do not keep the distinction between f.ex $(\mathcal{S}_1, \mathcal{S}_2) = (\hat{\mathbf{x}}_1, \hat{\mathbf{y}}_2)$ and $(\hat{\mathbf{y}}_1, \hat{\mathbf{x}}_2)$ since the sum

in both cases is a vector in the $(1,1)$-direction. The element spin thus gets four states representing different twists. The element direction $\hat{\mathbf{e}}_\alpha$ is implicitly given by the new state variable, namely perpendicular to the plane containing $\mathcal{S}_1$ and $\mathcal{S}_2$; again the sense of the direction is lost. It can probably be determined from the sequence of the elements, though. The case with the parallel and the anti-parallel spins at the ends is more complicated. In this case $\hat{\mathbf{e}}_\alpha$ is four times degenerate and only determined to be perpendicular to the direction of $\mathcal{S}_1$. Let us define the transformation so that $\mathbf{S}^A = (\mathcal{S}_1 \cdot \mathcal{S}_2)\sum_\alpha |\mathcal{S}_1 \cdot \hat{\mathbf{e}}_\alpha|\hat{\mathbf{e}}_\alpha$. Again the transformation looses a factor of two in information content, since $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2)$ and $(-\hat{\mathbf{x}}_1, -\hat{\mathbf{x}}_2)$ give the same $\mathbf{S}^A = \hat{\mathbf{e}}_\alpha$, and the anti parallel case, the same for $(\hat{\mathbf{x}}_1, -\hat{\mathbf{x}}_2)$ and $(-\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2)$ giving $\mathbf{S}^A = -\hat{\mathbf{e}}_\alpha$. This transformation is furthermore non-linear, but it is unique. Next we proceed the same way to determine the element variables $\mathbf{s}^a$ for the group two elements. What we have done here is to design rather complicated names to a number of building blocks. In fact several are identical by simple rotations. Consider a case with the spins surrounding an element are parallel. The attached elements therefore are perpendicular to the element tape as $\sqcup$. We have given three different names to this according to whether the legs point in the x, y or z-direction without sign. Normally one would just consider one building block, which could be turned in all possible (six) directions. It is convenient to think of the building blocks as a kind of electrical plugs. The one described has two holes and two pins on the same face $\boxed{\circ\ \circ\ \ J\ \ \circ\ \circ}$. This we could for the big letter elements $\pm Jand \pm K$ call a Ferro-plug. Similarly a plug with hole and pins on opposite faces one bound call an Anti-ferro-plug, and one with holes in the yz-face and pins in xz-face a Right-twist-plug and in the $-xz$-face a Left-twist-plug. Similarly for the small letter elements $\pm jand \pm k$ one can define four different plugs with holes and pins in the ends. We need two different kinds of holes for example $\square$ $\diamond$ and corresponding pins. Instead of the two different building blocks we could also have just one kind with holes on the sides and pins in the end. For the 90° case considered there are only these two times four distinct plugs or building blocks. The plugs can be twisted at an energy cost as described above according to the internal word $LMN\cdots RST$. As described until now the plugs are independent of the order $TSR\cdots NML$, but considered as electrical plugs they have a current direction built in. The element spin variables have twelve states $\pm x, \pm y, \pm z, \pm xy, \pm yz, \pm zx$ and the directions still have a direct physical meaning relative to the element directions. However, we could at this stage just consider the variables as Potts variables with twelve states $(1, 2, 3\cdots 12)$. It is therefore easy to generalize the situation to one in which we keep the complete description using Potts variables with 24 states $(1, 2, 3\cdots 24)$. This eliminates the information reduction in going from eq. (6) to (11). Now it should be possible to determine the interaction in terms of the constants J, K, j, k. This is done by considering a 'hinge' at position i joining the elements p and P. Let us fix the 'hinge' spin at i $\mathcal{S}_i$ in one direction $\hat{\mathbf{e}}_\alpha$. Then we write down the energy for the cases where $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i+1}$ are rotated in all possible directions (this requires rotation of large protein pieces). The possibilities can be exactly identified as pairs of element spins $\mathbf{s}_i^p\mathbf{S}_{i+1}^P$, where we have given the 'hinge' number i to the group two element according to eq. (5). By repeating this for all possible directions for $\mathcal{S}_i$ we can construct the full interaction matrix for the element variables. There is a large reduction in complexity in going to the element variables, the 'plugs', since the energy of their relative states is only related to the joint at i and we neglect the complicated global rotations of the rest of the protein.

We have therefore reduced the problem to an interacting chain of spin variables representing the elements of group one and two, with interaction constants which can (in principle) be derived from the original amino acid code. The Hamiltonian then reads.

$$\mathcal{H} = -\sum_{i=1}^{N\approx 20}\sum_{\alpha=1,\beta=1}^{12,12} \mathcal{J}_{i,i+1}^{\alpha,\beta}(J,K,j,k)s_{a,i}^{\beta}S_{A,i+1}^{\alpha} + \text{c.c} \qquad (11)$$

The interaction matrix $\mathcal{J}_{i,i+1}^{\alpha,\beta}$ can probably be simplified considerable by using the high symmetry of the problem as demonstrated with the 'plug' concept, such that only a few independent interaction constants need to be considered - probably four corresponding to the given J, K, j, k set.

We are now ready to start simulations assuming a certain chain of spin interactions with the Hamiltonian eq.(refeq5). We start at high temperatures. Although it is impossible to retrieve the original folding state because of the decimation of information in the step going to the element variables from eq.(8) to (11) we should be able to find sets of $\mathcal{J}_{i,i+1}^{\alpha,\beta}$ which give open structures, which are not desirable and close structures for which one probably can analyst the states corresponding to the more detailed Hamiltonian eq.(8). Then one could plot out the results, see how they look, add the short range interaction variables and determine the good foldings.

Now comes the question, how can we make classes. The result lies in the possible sequences of $\mathcal{J}_{i,i+1}^{\alpha,\beta}$ in the dividing assumptions made as if it depends on the direction of the word $LMN\cdots RSTU$ or not, and perhaps in the number of elements we allow for. Some real runs will probably make the situation more evident. We have established a unique language in which the foldings can be described with a desired degree of accuracy, with possibility for neglecting what we may guess are less important details.

The question of similarity between different foldings should be well defined in our procedure (the question is if it is useful of course). We have constructed a frame in which the distances do not play a role until late. The exact length of α-helixes do not play a role. The feature that β-sheets can be formed at a later stage is included. Temperature can be used as a folding 'instrument'. An energy function has been derived based on real physical principles.

4b. How many fold classes are there in total?

Along with the rapid expansion of sequence data from the worlds genome projects it is interesting to asses how many radically new protein structures which are yet to be discovered. This has remained an important and open question in micro biology. Remarkably one can obtain an upper bound of ≈ 4000 fold classes, from simple physical arguments.

Recently Chothia[1] addressed the question of how many protein families or fold classes there might be from a very different point of view. Based on the pace of discovery and the presently known number Chothia estimated a total of one thousand families. More interesting yet would be if that number was contained in the information provided by the amino acid sequence of the proteins themselves. A simple model for super structures of secondary protein structures is here shown to give approximately that number just from the linear sequence information and the constraint that the useful proteins are densely packed. A fold means[2, 3, 4, 5, 6, 7, 8] a particular structural topology that a folded

protein, or part of it, can assume in its native state. The new paradigm is to classify proteins by their structural topology rather than their sequence or, as usual, their function.

Proteins appear to belong to families, like plants, with specific characteristics. The families contain many variants. Linné[9] in the 18th century succeeded in the field of botany to identify the important classification parameters. It gives a systematic, although not natural classification. The dense fold patterns for proteins may be such characteristics and we shall identify a class of similar folds with families in agreement with Chothia (and with the qualifications mentioned that the fold classes need not be the natural families, a problem already encountered by Linné in his classification). Chothia gave a good review of the present knowledge of the protein families, which need not be repeated here.

A major puzzle has been how nature can fold a long protein into its dense form in a short time without trying all possibilities. This can be elucidated by well-known problems in condensed matter physics. In many cases in physics of highly degenerate problems it is the weak symmetry breaking forces which determine the major structure, whereas it is the strong forces which determine the details. An example is the Heisenberg magnet for which an infinitesimal anisotropy field will determine the overall direction of the ordering vector, whereas all other properties are determined by the exchange interaction. Consider now an extended protein, already composed of its secondary structures, with elements such as (1) α-helixes [10], β-strands for potential β-sheets and (2) the intervening strands (e.g. representing turns). Let us then suppose that the linear amino acid information provides a very weak preference for a particular bending at each junction. This would break the symmetry, which gives rise to the enormous "random walk" degeneracy, and thus determines essentially the final folds, the fold classes, and to a certain extent removes the Levinthal paradox[11]. He pointed out, that the space of possible configuration, i.e. the number of local minima that a medium size protein can attain, is enormous ($\sim 20^{100}$) judged just from the linear sequence information. Our hypothesis is that nature solves the paradox and reduces the number of possibilities drastically by breaking the proteins up into building blocks, secondary structures, domains etc., with specific rules of connections.

Let us introduce a highly simplified model. Suppose the above mentioned elements are essentially linear of various lengths, the direction is given by a unit vector $\hat{e}$. To describe the junctions consider for example at site P a vertical α-helix ($\hat{e}_z^P$) with a strand entering at the bottom in the direction north. Suppose the α-helix can transfer the information that the strand going out at the top is approximately in the same direction (≈north), opposite (≈south), or ≈east, or ≈west. The situation can be described by the hinge variables $\mathcal{S}_p$ as spins giving the hinge direction and sense of turn. The information along the element is then reduced to a set of interaction constants J_p, K_p within a total of four values $(\pm J, \pm K)$ which favors one of the four relative spin directions. The same can be done for the other elements (J_p, K_p for element (1) and j_p, k_p for element (2)). Thus one could characterize a given fold configuration (here a four-helix-bundle) by a linear string of e.g. the following content: $\Updownarrow J \Updownarrow j \Updownarrow K \Updownarrow j \Updownarrow -K \Updownarrow j \Updownarrow -J \Updownarrow$,
where $\Updownarrow$ represents one of the of symbols is given in the figure below, together with a more realistic representation of the four-helix bundle 1hmq. The reduced information giving the spin directions can be furnished by many amino acid sequences. This provides in fact the basis for the classification, i.e. many variants having the same fold. In order to be able to describe the energy cost for violating the optimum fold we write the argument as a Hamiltonian

$$\mathcal{H} = -\sum_{P=2n+1} (J_P \mathcal{S}_P \cdot \mathcal{S}_{P+1} \;+\; K_P \mathcal{S}_P \times \mathcal{S}_{P+1} \cdot \hat{\mathbf{e}}_\alpha^P) - \sum_{p=2n} (j_p \mathcal{S}_p \cdot \mathcal{S}_{p+1} \;+\; k_p \mathcal{S}_p \times \mathcal{S}_{p+1} \cdot \hat{\mathbf{e}}_\alpha^p). \tag{12}$$

This simple model is more general than a lattice model because we are allowing for both a variable length of the elements and for a flexibility in angles. From a specific sequence of j_p, k_p and J_p, K_p we can find the corresponding spin directions and construct the protein structure given the lengths of the elements. By Monte Carlo computer simulations the model can be shown to exhibit known

folds amongst a wealth of other structures such as non-compact and loosely packed structures and structures that are too densely entangled in one another.

After randomly varying the linear representation and permuting the coupling constants one can search for the minima of the energy function given above. In that way one can obtain members from different folding classes and in principle traversing the whole space of possible folds. A phase diagram can easily be extracted from such an analysis based on the energy function $\mathcal{H}$ and we already know from previous analysis [12, 13, 14] that this type of energy function exhibits an ordered phase at low temperature corresponding to folded structures of proteins and disordered phases at higher temperature corresponding to unfolded or misfolded patterns.

Here we shall demonstrate that the model is amenable to a simple estimate of the number of dense protein fold classes. For a random sequence of interaction parameters it is most likely that the structure is extended. The number of such structures is very large and can be estimated as $\mathcal{N}_{random} \propto z^N = 4^N \approx 10^{12}$, where $N = 20$ (taken as the average number of structural elements in proteins times two, counting the intervening strands too) is a typical number of junctions (spins) and $z = 4$ the turn possibilities. The problem is to find those sequences which result in dense structures. This problem is closely related to that of self avoiding random walks (SAW), see e.g. de Gennes[15]. The end-to-end square distance R_f^2 can be shown to scale with N as $R_f \propto N^\nu$, $\nu = 3/5$, for dimension $d = 3$. For the protein to be dense, R_f must be small, and that reduces the number of possibilities drastically to $\mathcal{N}_{SAW,close} \propto \bar{z}^N / R_f^d$. For self avoiding walks $\bar{z} = z - 1 - \epsilon$, where the -1 corrects for the direct return and where $\epsilon = 0.32$ for $d = 3$ is a small correction for possible later crossings. By construction of the model we have no direct returns and thus $\bar{z} = 4$ or better $\bar{z} \approx 4 - \epsilon$. However, we want to require that the proteins are still denser. The root-mean-square radius scales as $< r^2 > \propto N$. Therefore the number of proteins with extent being of the order of the length of the elements is further reduced by a factor $N^{\frac{1}{2}}$. Hence one obtains the following estimate of the number of proteins which are dense and self avoiding

$$\mathcal{N}_{SAW,dense} \propto \bar{z}^N / N^{d(\nu+\frac{1}{2})} = \bar{z}^N / N^{3.3}. \tag{13}$$

For a typical protein consisting of 10 α-helixes or β-strands there are $N = 18$ junctions; and the number of distinct structural folds is computed from (2) to be $\mathcal{N} = 10^6$. This number is vastly reduced from that of the free random walk estimation ($\approx 10^{11}$). It must further be reduced considerably by requiring that the potential β-strands are close together in space in order to form β-sheets. Assuming n α-helixes and m β-strands the

restriction gives a reduction factor $\propto n!m!/(n+m)!$, which for $n \approx m$ gives a reduction factor 250 and brings the estimated number down to

$$\mathcal{N}_{foldclasses} \approx 4000 \tag{14}$$

for the typical proteins. This is very close to the estimate by Chothia, who obtained his estimate in a totally different way. It is still a bit higher and allows therefore both for more possible findings than foreseen by Chothia, and for the possibility that nature in the course of the evolution has not used all the statistically possible options, or rather that some folded protein configuration were discarded because of lack of functionality. Clearly a further reduction is arising if the structures must in addition fulfill certain functional demands.

The argument given here has been to estimate the number, or upper bound, (≤ 4000) of the final structures. In these the strong forces between the elements can equally well be included, probably distorting the structures somewhat. However, the more important aspect in our model is that the proteins already in their linear amino acid combination have the fold class impregnated within them, allowing the fold to be much more well determined and self-organized than that conceivable from a totally random trial and error method. It is clear from the argument that a given fold cannot be used to determine the amino acid sequence, whereas the reverse is possible. In our model the folds are coded in a decimating code. We have only discussed relatively small proteins, but as discussed by Chothia, the larger proteins produced at the recent stages of evolution are generally combinations of the elementary ones. This will not increase the number of possibilities drastically!

References

[1] C. Chothia, Nature, **357**, 543 (1992).

[2] J. W. Ponder and F. M. Richards, J. Mol. Biol. **193**, 775 (1987).

[3] S. T. Rao and M. G. Rossmann, J. Mol. Biol. **76**, 241 (1973).

[4] J. S. Richardson, Adv. Protein Chem. **34**, 167 (1981).

[5] D. J. Jones, W. R. Taylor and J. M. Thornton, Nature **358**, 86 (1992).

[6] T. L. Blundell and M. S. Johnson, Protein Science, **2**, 877 (1993).

[7] S. Pascarella and P. Argos, Protein Engineering **5**,

[8] C. Sander and L. Holm, J. Mol. Biol., **225**, 93 (1992). 121 (1992).

[9] C. von Linné *Fundamenta Botanica* (1738), *Species Plantarum* (1753).

[10] Nature only uses α-helices with one handednes, therefore we only assign one element to an α-helix.

[11] C. J. Levinthal, Chem. Phys. **65**, 99 (1968).

[12] T. Castan and P. A. Lindgaard, Phys. Rev. B, **40**, 5069 (1989).

[13] M. Sasai and P. G. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).

[14] H. Bohr and P. G. Wolynes, Phys. Rev. A, **46**, 5242 (1992).

[15] P. G. de Gennes, *Scaling Concept in Polymer Physics*, Cornell University Press, Ithaca (1979).

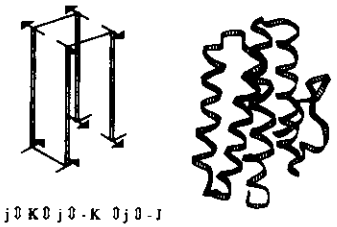


Figure: Chain model representation of the four helix bundle protein (left side) and (to the right) the real protein in a ribbon representation

Large5. Formation of fold patterns in the early stages of protein folding

In chapter 5 a whole new formalism is presented for analysing quantitatively the early stages of protein folding from a structural viewpoint. Briefly the main ideas of such an approach is the following. The processes of protein folding can be described by neural networks using a description of the protein configuration contacts. The initial fast energetically down-hill process from the random coil structure to a collapsed state of secondary structures may be thought of as using a quasi static mean field equation which approximately is in the form of a Feed-forward Perceptron equation and a time dependent form equivalent to the time evolution of a feed-back Hopfield neural network. On the basis of these results we propose some changes that might improve the neural network schemes used for prediction of tertiary structure of proteins. Our formalism can also be used to analyse the structure of the basin of attractors for the protein neural networks, which might describe the early events in protein folding. Furthermore the formalism can give an insight into the various phases of secondary structure formation in terms of temperature, overlap parameter and randomness.

5a. Contact dynamics of protein folding and neural networks.

Recently neural networks have been demonstrated to be useful in the prediction of secondary structures (ref. 1-3) and tertiary structure (ref. 4, 5, 6 and 7) based on the protein's sequence of amino acids, the latter only when there is a large degree of homology to the training set. When little homology exists the determination of a protein's tertiary structure solely from its sequence, is still a major challenge of biotechnology and represents the practical side of the protein folding problem. Understanding the success and limitations of neural network techniques applied to structure determination may give insight into how to make better predictions and how proteins themselves carry out this daunting computational task (ref. 8). Thus we ask whether there is a way of deriving the neural network equations from the physics of protein dynamics. It turns out, that in the pursuit of a description of the initial down-hill processes of protein folding we end up with equations for neural networks of feed-back type which can be approximated as well in a short time limit by feed forward nets like those used. This can be understood in the sense that feed forward neural networks provide immediate responses to sequences giving structures in return.

Before building up the contact formalism we shall briefly give a crude scenario of what is known of the first stages of protein folding since our dynamical equations first of all apply to that stage.

5a1. A crude scenario of the initial stages of protein folding.

There is some reasonable consensus concerning the experimental data (e.g. ref. 9, 10) about the main features of the various stages of protein folding. Details are still sketchy but most crudely it seems one can divide the folding process into three stages:

1. A fast "down-hill" energy minimization process forming the molten globule state, after a quick formation of most of the secondary structures that are found later in the final native configuration.

2. A slower docking process between preformed structural elements and a partial organization of the hydrophobic core.

3. Slow molecular dynamics where the side-chains rearrange themselves with a delicate change and some H-bonds or torsion angles (ref. 11).

Experiments also indicate that these Molten Globules are compact intermediates and have substantial secondary structure but few if any fixed tertiary structures. The hydrophobic side-chains are not buried as in the native proteins but often exposed to water. It is important to differentiate between different time scales involved in these folding stages. Also at a structural level in a simplified description there are different scales, the first stage consisting primarily of large scale back-bone motions while the last stages have the more fine grained dynamics of side-chain motions. We shall be dealing mostly with the first down-hill energy minimization stages of protein folding, which are dominated by polymer physics, while the later stages are better described by intricate molecular dynamics.

The main motivation behind this presentation is to construct a physical framework in which the various neural network schemes for prediction of protein structures can be derived and explained. At the same time this approach that focuses on the information processing aspect of the problem may help elucidate the role of the different folding processes that participate in the dynamics of the protein folding. In our analysis the fast down-hill physical processes are closely related to the feed forward neural network schemes in the sense that the mean field equations of the former processes are the ones that describe the latter perceptrons.

Another aspect of this approach is that the mapping makes available the use of many of the formal ideas of neural network theory for thinking about folding problems. For example the theories of the size of basins of attraction for nets can give ideas about the probability a sequence can fold really fast. Also Gardners formalism (ref. 12) can be used to understand whether simple codes for folding can exist. We will only hint about these ideas in this preliminary report.

In the next chapters we shall first briefly report on the practical techniques of neural network type for protein structure prediction and then in the light of the formalism for the fast down-hill processes and the dynamics of contacts try to derive and motivate these neural network methods which seem to be intimately related to the early stages of protein folding.

5a2. Contact dynamics for proteins

The basic variables of our description of the protein dynamics are contacts which can be defined as a density variable which we denote by σ_{ij}^{ρ}. This variable measures the correlation between the i'th and j'th residue relative to the distance ρ between the residue (C_α) points on the protein backbone. The correlation measures how close the residues are to each other. The contact variables could be described as binary variables $\{0,1\}$, where 1 stands for contact within a shell of inner radius ρ and outher radius $\rho + d\rho$ and 0 for no contact (see fig. 1). The variables could also be described by a continuous valued function. In this study we shall maintain the binary notion:

$$\sigma_{ij}^{\rho} = \begin{cases} 1 & \text{if j is within a shell of } (\rho, \rho + d\rho) \text{ from i.} \\ 0 & \text{if j is not.} \end{cases}$$

If we integrate up this contact variable over the radial distance ρ we obtain a coarse grained natural variable s_{ij} which describes close contacts. It is defined as:

$$s_{ij}^{\rho} = \int_0^{\rho} \sigma_{ij}^{\rho'} d\rho' \tag{15}$$

that measures if residue j is within the range of ρ from i or not.

From the contact variables σ_{ij} and s_{ij} we can define many useful quantities, one being a quantity for measuring each residue's position relative to the surface defined by, for example, the center of mass for all the side-chains. We have called such quantity the "superficiality" S_i^ρ (see ref. 13) and define it as:

$$S_{ij}^\rho = \sum_{j \neq i} s_{ij}^\rho = \sum_{j \neq i} \int_0^\rho \sigma_{ij}^{\rho'} d\rho' \tag{16}$$

This quantity measures the number of neighbours of residue i within the radial distance of ρ. The quantity tells how much each residue is buried in the interior of the protein (large values) or exposed to the surface (small values).

We can also define the contact variables to be relative to specific atoms of each residue. For example $\sigma_{\alpha ij}^\rho$ could denote a contact measured with respect to the distance between the C_α atoms for the residue i and residue j. Similarly we could define $\sigma_{\beta ij}^\rho$ as a contact measured with respect to the distance between the C_β atoms. The calculation is actually done as: $S_{\alpha i}^\rho = \sum_{j \neq i} \sum_{\alpha,\beta} s_{\alpha i,j}^\rho$. With these two difinitions we could construct a quantity called the vector contact density or the vector superficiality as:

$$S_{\beta \alpha i}^\rho = S_{\beta i}^\rho - S_{\alpha i}^\rho \tag{17}$$

Such quantity $S_{\beta \alpha i}^\rho$ tells about the direction the given side-chain i points. A low value for S will indicate that the side-chain i points out towards the exposed surface of the protein, while a larger value will indicate that the side-chain points towards the interior of the protein. Altogether it is quite clear that the superficiality variable S_i is related, if not proportional, to the solvent accessible surface area. The solvent accessible area is harder to calculate than the superficiality whose calculation involves simply summing up neighbours around each residue. A feed forward neural network with the purpose of predicting surface structures of proteins on the basis of their sequence of amino acids has been constructed (ref. 13) using the superficiality concept. The superficiality was predicted for each residue and then related to the surface area. The neural network was up to 70 % correct in predicting the superficiality of a test set of 10 proteins after trained on 40 other proteins from the Brookhaven PDB data set and with an input window of 7 residues.

It is very important to realize that this binary distance matrix description is a unique way of representing the 3-dimensional structure of a given protein, see ref. 39. Especially the secondary structures are easily recognisable patterns within a binary distance matrix. For example the various helices are elongated structures along the diagonal while the anti-parallel beta-sheets are elongated structures perpendicular to the diagonal, see fig. 4.

It is actually possible (ref. 40) to consider the contact variable matrices $\sigma_{i,j}^\rho$ as operators in a Hilbert space. In fact all types of contact (distance) matrices can be decomposed into elementary 2x2 matrices, such as the Pauli spin SU(2) matrices. Specifically a 4x4 alpha-helix contact matrix can be composed by a tensor product of the Pauli matrices σ_1,σ_2: $\sigma_{\alpha-helix}^{4,4} = \sigma_2 \otimes \sigma_3$ and similarly a 4x4 beta-sheet contact matrix can be composed by the tensor product: $\sigma_{\beta-sheet}^{4,4} = \sigma_1 \otimes \sigma_2$ and it seems resonable to expect that any binary distance matrix can be generated by tensor products of elementary Pauli spin matrices. Therefore the whole dynamics of arbitrary contact matrices can be described by the eigenvalues of the dynamical operators projected on sub-spaces spaned by the Pauli spin matrices. The corresponding wave functions are really the probability distributions of having specific contacts.

5a3. The energy function for protein contacts

In the last chapter we introduced the basic variables σ_{ij}^ρ (and s_{ij}^ρ) measuring the contact between position i and j, or in terms of residues on a protein backbone, measuring whether the i'th residue is mear (within) the distance ρ of the j'th residue. We shall now try construct the free energy in terms of these variables (ref. 20). Basically there are two parts of the free energy to consider. The first part is the free energy (primarily entropic) of a chain and the effect of its various contacts involving hydrogen bonds in secondary istructures and the other part is the contribution to the free energy from the interaction among the side-chains.

5a3a. The general expansion

In this section the free energy is expanded in the basic contact variable σ_{ij}^ρ like a Landau theory. One should note the basic variable σ_{ij} (we shall in the following mostly be omitting the index ρ) is the probability for a contact between i and j rather than an actual contact. This contact variable resembles the density variables used in liquid state theory and much of the analysis of density functional theory can be taken over intact (ref. 21).

We shall first write the free energy in the general form:

$$\mathcal{F} = \mathcal{F}^O + \mathcal{F}^I \tag{18}$$

where $\mathcal{F}^O$ is the free part and $\mathcal{F}^I$ is the interaction term. First we decompose the free energy term into:

$$\mathcal{F}^O(\sigma_{ij}) = \mathcal{F}_{P.G.}(\sigma_{ij}) + \Delta\mathcal{F}^O(\sigma_{ij}) \tag{19}$$

where the first term $\mathcal{F}_{P.G.}$ is an entropy term for a perfect gas, and recalling that the Boltzmann entropy expression for an ideal gas is $\sum_i P_i log P_i$, we write:

$$\mathcal{F}_{P.G.} = k_B T \sum_{ij} \int \sigma_{ij} log \sigma_{ij} d\rho \tag{20}$$

If we use the integrated shortrange contact variable s_{ij} the Boltzman entropy becomes:

$$\mathcal{F}_{P.G} = k_B T (\sum_{ij} s_{ij} log s_{ij} + (1 - s_{ij}) log(1 - s_{ij})) \tag{21}$$

The second term in $\mathcal{F}^O(\sigma_{ij})$ is expanded in orders of σ_{ij}:

$$\Delta\mathcal{F}^O = \sum_{ij} W^o_{ij}\sigma_{ij} + \sum_{ij}\sum_{kl} W^o_{ijkl}\sigma_{ij}\sigma_{kl} + ... \tag{22}$$

being a connectivity expansion where the first term measures the free energy gain for forming the contact (i,j), and the second term measures the free energy gain for forming the contact (i,j) when already having formed the contact (l,k).

A similar Taylor series expansion in terms of s_{ij} can be used:

$$\Delta\mathcal{F}^O(s_{ij}) = \sum_{ij} W^o_{ij}s_{ij} + \sum_{ij}\sum_{kl} W^o_{ijkl}s_{ij}s_{kl} \tag{23}$$

where, of course the Taylor coefficients will be different.

It is important to bear in mind that these terms represent the energies for contacts between points on a free chain. Hence the factors $W_{i..}$ in the expansion can be derived from polymer chain physics. In the first approximation W_{ij} can be derived from free flight statistics (ref. 26), but in order to go further and to include for example excluded volume effects, a whole formalism has been developed (ref. 22, 27) using Feynman path integral techniques. Including excluded volume should give a more realistic picture of the chain with Van der Waal's spheres centered at the atomic coordinates.

Again there are many developments in using free energy functionals for singlet density for polymers (ref. 22-26) and these may be generalized to the pair level used here. Also the back-bone can undergo hydrogen bonding into α-helices and β-sheets. These effects on W^o_{ij} etc. might be treated using the theories of associated fluids (ref. 27).

The interaction term $\mathcal{F}^I$ takes into account the interaction between the residues. The interaction energy is expanded in the same way as the free term above:

$$\mathcal{F}^I = \sum_{ij} W^I_{ij}\sigma_{ij} + \sum_{ij}\sum_{kl} W^I_{ijkl}\sigma_{ij}\sigma_{kl} \tag{24}$$

Now, however, it is more complicated to determine the expansion coefficients $W^I_{ij}(\{q_i\})$. They will depend on a detailed nature of the residues, i.e. sequence information $\{q_i\}$, and this dependence must be **learned**.

As we shall see the minimization of the free energy in this representation leads to a neural network and the learning can be carried out by back propagation (or other algorithms) given some examples. In this representation the associative memory Hamiltonians of Friedrichs and Wolynes (ref. 4) correspond to **instructing** the interaction matrix with examples:

$$\mathcal{F}^I = \sum_{ij}\sum_{\alpha}\int d\rho\gamma^\alpha_{ij}(q^\alpha_i, q^T_i, q^\alpha_j, q^T_j)\sigma^{\alpha\rho}_{ij}\sigma_{ij} \tag{25}$$

where $\sigma^{\alpha\rho}_{ij} =< \theta(r^\alpha_{ij}-\rho) >$. Since the higher terms vanish in this expression the Friedrichs-Wolynes Hamiltonian could be termed a **linear** associative memory Hamiltonian. An associative memory term using only the coarse grained near contact variable s_{ij} can be used too:

$$\mathcal{F}^I = \sum_{ij\alpha}\gamma^\alpha_{ij}s^\alpha_{ij}s_{ij} \tag{26}$$

where s^α_{ij} is the value of the contact variable in the example α. Although this simpler form doubtless diminishes the capacity of the model, it is a useful model interaction for discussing the basin of attraction for early stages of folding.

We are now in the position to write down the total free energy expression with the linear associative memory Hamiltonian (in terms of s_{ij}):

$$\mathcal{F} = \sum_{ij}\int d\rho kT[(1-s_{ij})log(1-s_{ij}) + s_{ij}logs_{ij}] - \sum_{ij} W^o_{ij}s_{ij} - \sum_{ij}\sum_{kl} W^o_{ijkl}s_{ij}s_{kl} - \sum_{ij\alpha}\gamma^\alpha_{ij}s_{ij}s^\alpha_{ij} \tag{27}$$

5a3b. The simplest polymer framework for the calculation of the factors W^o

In this subsection we shall briefly mention a formalism that will enables us to calculate the chain polymer energy terms. It is basically the formalism developed in the references 22, 28, 29.

We first consider smooth chains of residues labelled by an index. The contacts will be denoted by a pair of indices (i,j) meaning a contact between residue i and j. A contact means that the positions for the i'th and the j'th residues are within a certain range ρ of each other. Sequential neighbours are discounted as being able to form contacts. The order of a contact (i,j) is the number of residues between the i'th and the j'th position on the chain. Thus contacts of order 1 are excluded. The contacts of a given order correspond to a loop of a given size.

As explained in reference 29, the effect of a contact can be measured by a reduction R of the chains conformational freedom. If $\Omega(N,i,j)$ stands for the number of conformations of a chain with N bonds and an (i,j) contact and $\Omega_o(N)$ is the total number of accessible conformations, the reduction factor $R(N,i,j)$ is given as:

$$R(N,i,j) = \frac{\Omega(N,i,j)}{\Omega_o(N)} \tag{28}$$

or in terms of our energy contact matrix W_{ij}:

$$W_{ij} = log(R(N,i,j)) = log(\frac{\Omega(N,i,j)}{\Omega_o(N)}) \tag{29}$$

Making use of the results of the simple theory of Jacobson and Stockmayer based on "random flight statistics", ref. 28, the reduction factor can be written as:

$$R(N,i,j) = (\Delta v)[\frac{d}{2\pi \mid i-j \mid}]^{d/2}; (N \geq \mid i-j \mid) \tag{30}$$

where d is the dimensionality of the configurations and Δv is a tolerance volume factor, actuality the volume of a sphere of radius ρ in which the residues qualify as being in contacts.

The effect of two contacts (i,j) and (k,l) in the same approximation of "random flight statistics" can also be measured by a reduction factor $R(N,i,j,k,l)$ and one can define (ref. 29) a topological correlation factor g as:

$$g_{k_1,k_2}(L) = \frac{R(N,i,j,k,l)}{R(N,i,j)R(N,k,l)} \tag{31}$$

where k_1 and k_2 are the two contact orders:($k_1 =\mid i-j \mid$) and ($k_2 =\mid k-l \mid$) respectively, and L is the distance between the two contacts ($L =\mid j-l \mid$). This correlation therefore measures the ratio of the actual number of conformations satisfying the two contacts and their dependences, to the number of conformations in the case of the two contacts being independent. Thus it measures the degree to which one loop effects influence the formation of another. The simple approach of random flight statistics does not take into account any physical geometrical constraints, such as excluded volume effects, and therefore gives a too simple picture of polymer dynamics. However, there exist approaches, ref. 29, to the calculation of these excluded volumen effects via diagramatic methods of polymer theory.

It is interesting to note that there are only 3 topological cases of 2-contact configurations realized by the following index inequalities: 1. case: $0 \leq L \leq k_2 - k_1$, 2. case: $k_2 - k_1 < L < k_2$, 3. case: $L \geq k_2$, when we assume positive separation $L > 0$. In the context of protein backbone contacts and secondary structures the first case correspond to anti-parallel β-sheets, the second case to α-helices, π-helices etc and the third case to random coil configurations.

5a3c. Path integral formalism for the polymer factors

We shall now try to make a path integral construction of these reduction factors on the basis of a more realistic picture where for example excluded volume and chain rigidity effects are included. Usually, see ref. 29, the chain partition function $Q(N)$ is written as:

$$Q(N) = \int [Dc] e^{-\beta H(N,v_o)} \tag{32}$$

where the Boltzmann energy function is:

$$\begin{aligned} H(N,v_o) &= 1/2 \int_0^N d\tau \mid \frac{d\mathbf{c}(\tau)}{d\tau}^2 + v_o \int_0^N d\tau \int_0^{\tau'} \tau' \delta \mid \mathbf{c}(\tau) - \mathbf{c}(\tau') \mid \\ &+ +\frac{1}{2} k_c \int d\tau \cdot (\kappa(\tau))^2 \end{aligned} \tag{33}$$

and where τ is a contour length parameter and $\mathbf{c}(\tau)$ specifies the chain conformation, $\mathbf{c}(\tau) = \sqrt{d}\mathbf{r}(\tau)$ and $\mathbf{r}(\tau)$ is the position vector at τ from the end. Furthermore k_c is the rigidity constant and κ is the mean curvature, which is:

$$\kappa = g^{-3/2} \frac{d\mathbf{r}}{d\tau} \times \frac{d^2\mathbf{r}}{d\tau^2} \tag{34}$$

where g is the metric, $g = \mathbf{r}_\tau \mathbf{r}_\tau$ and the cross-product in 2 dimensions is defined as $\mathbf{a} \times \mathbf{b} = \epsilon_{ij} a_i b_j$ being a scalar. $D[\mathbf{c}]$ is of course a functional integral measure over all possible path (conformations).

The path integral can now be used to calculate the number of conformations $Q(N, l_o, l_o + k)$ restricted by the contact points at l_o and $l_o + k$, having $\mathbf{r}_s = \mathbf{r}(l_o) - \mathbf{r}(l_o + k)$ being within Δv:

$$\begin{aligned} Q(N, l_o, l_o + k) &= \int_{\Delta v} d\mathbf{r}_s \int [Dc \delta \mathbf{r}(l_o) - \mathbf{r}(l_o + k) - \mathbf{r}_s] e^{-\beta H(N,v_o)} \\ &= (\Delta v_o) d^{d/2} \int [D\mathbf{c}] \delta[\mathbf{c}(l_o) - \mathbf{c}(l_o + k)] e^{-\beta H(N,v_o)} \end{aligned} \tag{35}$$

In order to proceed we could now expand the expression in powers of v_o (the excluded volume interaction strength):

$$Q_r = Q_r^o + Q_r^1 v_o + Q_r^2 (v_o)^2 + ... \tag{36}$$

and

$$Q_o = Q_o^0 + Q_o^1 v_o Q_o^2 (v_o)^2 + ... \tag{37}$$

and similarly for the reduction factor:

$$R(N, l_0, l_0 + k) = \frac{Q_r(N, l_o, l_o + k)}{Q_o(N)} = (\Delta v) d^{d/2} \frac{Q_r^o}{Q_o^o} \cdot [1 + v_o (\frac{Q_r^1}{Q_r^o} - \frac{Q_o^1}{Q_o^o}) + O(v_o^2)] \tag{38}$$

We can now for technical conveniences develop a Feynman diagrammatic technique for calculating the free energy contribution to any configuration of the chain in the excluded volume interaction picture. Basically the Feynman rules are obtained by taking v_o as the vertex function and $G_o(\mathbf{c} - \mathbf{c}', l)$ as the free propagator, where:

$$G_o(\mathbf{c} - \mathbf{c}', l) = (\frac{1}{2\pi l})^{d/2} e^{-\frac{1}{2l}(\mathbf{c}-\mathbf{c}')^2} \tag{39}$$

and finally integrate over all coordinate vectors $\mathbf{c}$ (between the vertices) over all space and integrate over all pair of contour lengths (τ, τ') connected by the excluded volume interaction line.

5a4. Dynamics of contacts

The next step is to derive and solve the dynamical equations from the energy expression derived in the last chapter. As we emphasised in the beginning we are interested in the

early stages of the folding process of fast down-hill energy minimization so we will be dealing with a first order differential equation in time. Basically we have a Langevin equation:

$$\frac{ds_{ij}^{\rho}}{dt} = -\frac{\partial \mathcal{F}}{\partial s_{ij}^{\rho}} \tag{40}$$

First we shall study the solutions to the static mean field equation and then the time dependent equation.

This equation is, of course, phenomenological and close in spirit to what is used for magnetic systems. It is more natural, to chemists, to think of s_{ij} as a concentration variable for pairs. A rate law embodying the same free energy function can be written down for s_{ij} but the forward and backward rates will depend nonlinearly on the change in the free energy in making a contact. The pre-factor of such a rate can be estimated in a manner analogous to that used by Karplus and Weaver in their diffusion-collision picture, ref. 30. We do not think the analysis of the nonlinear chemical kinetic equations would be much different from that in the magnetic language.

5a4a. Static Mean Field Equation and the feed forward Neural Networks

The time independent mean field equation for the energy functional introduced in the last chapter becomes:

$$\frac{\partial \mathcal{F}}{\partial s_{ij}^{\rho}} = 0 \tag{41}$$

If we now insert the expression for $\mathcal{F}$ from the last chapter and redefine the variable $s_{ij} \rightarrow (s_{ij}+1)/2$ we get:

$$0 = log(1+s_{ij}) - log(1-s_{ij}) + W_{ij}^{o} + \sum_{kl} \frac{1}{2} W_{ijkl}^{o} s_{kl} + W_{ij}^{I} + ... \tag{42}$$

If we now exclude higher order terms, make use of the relation:$tanh^{-1}x = 1/2log(\frac{1+x}{1-x})$ and insert $W_{ijkl}^{I} = q_i q_j \sum_{\alpha} q_k^{\alpha} q_l^{\alpha}$ we obtain:

$$\begin{aligned} s_{ij} &= tanh[W_{ij}^{o} + q_i q_j \sum_{\alpha} q_k^{\alpha} q_l^{\alpha} s_{kl}^{\alpha}] \\ &= tanh[W_{ij}^{o} + q_i q_j \omega] \end{aligned} \tag{43}$$

where we have absorbed the constant term W_{ij}^{I} in W_{ij}^{o} and defined $\omega = \sum_{kl\alpha} q_k^{\alpha} q_l^{\alpha} s_{kl}^{\alpha}$. The equation above describes a trivial feed forward neural network with two input neurons taking values of q_i and one output neuron with the values of s_{ij} with the rule that:

$$s_{ij} = \begin{cases} 1 & \text{if } q_i q_j \omega > W_{ij}^{o} \\ 0 & \text{if } q_i q_j \omega < W_{ij}^{o} \end{cases}$$

If we now include higher order terms (and absorb the constants in $\bar{W}_{ij}^{o}$, we have:

$$s_{ij} = tanh[\bar{W}_{ij}^{o} + q_i q_j \omega + \frac{1}{2} \sum_{kl} W_{ijkl}^{o} s_{kl}] \tag{44}$$

which can be solved iteratively:

$$\begin{aligned} s_{ij,0} &= tanh[W_{ij}^{o} + q_i q_j \omega] \\ s_{ij,1} &= tanh[\bar{W}_{ij}^{o} + q_i q_j \omega + \frac{1}{2} \sum_{kl,\phi} W_{ijkl}^{o} s_{kl,0}^{\phi}] \end{aligned} \tag{45}$$

With the expansion of the lower order contact variable $s_{ij,0}$, that depends on the side-chain properties: $s_{ij,0} = A^{o} + B^{1} q_i q_j + ...$, it becomes an ordinary feed forward neural network.

This network, which is similar to the Copenhagen network (see ref. 5), predicts contacts or distance matrices from sequence information q_i. In principle the thresholds and weights can be determined partly from polymer chain dynamics. Actually since parts of the thresholds are undetermined (the part with ω) the present analysis suggests that the essential part of the training of the neural network for predicting distance matrices is on the thresholds. It also suggests that the "best" network should be recursive. A similar network, see ref. 13, is one for predicting the integrated contacts, the superficiality $S_i = \sum_j s_{ij}$.

A very similar view of the relationship of mean field theory to neural net architechture has been independently developed by Bryngelson et al. (ref. 31) and used to discuss helix prediction. The connection between energy functions for biomolecules and neural network algorithms has also been discussed by Steeg (ref. 32) for RNA secondary structures.

5a4b. The time-dependent dynamical equation and Hopfield Neural Networks

In this subsection we shall consider the time dependent equation for the fast down-hill folding process. If we insert the expression for $\mathcal{F}$ we have:

$$\begin{aligned} \frac{ds_{ij}}{dt} &= -\frac{\partial \mathcal{F}}{\partial s_{ij}} \Rightarrow \\ \frac{ds_{ij}}{dt} &= -kTlog(\frac{s_{ij}}{1-s_{ij}}) + W_{ij}^{o} + \sum_{kl} W_{ijkl}^{o} s_{kl} + W_{ij}^{I} + ... \end{aligned} \tag{46}$$

or the following difference equation:

$$s_{ij}(t+\Delta t) = [\sum_{kl} W_{ijkl}^{o} s_{kl} + W_{ij}^{I} - kTlog(\frac{s_{ij}}{1-s_{ij}}) + ..]\Delta t + s_{ij}(t) \tag{47}$$

The first order differential equation looks very much like an evolution equation for a Hopfield feed back neural network:

$$s_i(t+1) = sgn[\sum_{j=1}^{N} W_{ij}s_j(t) - \theta_i] \tag{48}$$

if we bring the logarithmic term over to the left hand of the equation and then consider time steps in which the variable $s_{ij}(t+1)$ has relaxed to a local equilibrium. Then we can neglect the differential term and write the equation as:

$$kTlog(\frac{s_{ij}(t+1)}{1-s_{ij}(t+1)}) = h_{ij} = \sum_{kl} W^{o}ijkls_{kl}(t) + W^{o}_{ij} + W^{I}_{ij} + .. \tag{49}$$

and if we consider the low temperature limit with the two cases of s being nearly either 1 or 0 for either positive or negative potential h_{ij} we can transform the logarithm into a sign function and obtain the equation:

$$s_{ij}(t+1) = sgn[\sum_{kl} W^{o}_{ijkl}s_{kl}(t) + W^{o}_{ij} + W^{I}_{ij}] \tag{50}$$

which is a Boolean net.

Such a Boolean net arising from this special low temperature limit is unlikely to be valid for real proteins since it implies a rigid code for the protein folding. Therefore we shall mostly be studying the more general evolution equation:

$$s_{ij}(t+1) = tanh[(k_BT)^{-1}(\sum_{kl} W^{o}_{ijkl}s_{kl}(t) + W^{o}_{ij} + W^{I}_{ij})] \tag{51}$$

that is derived similarly to the evolution equation in the previous chapter. Both the Boolean and the last mentioned equation describe the evolution of feed back neural networks. In the next section we shall discuss the solutions to these equations.

5a4c. Fixed point solutions of the evolution equation and topology of chain connectivity

We shall now try to solve the time dependent equation which we called the evolution equation for contacts in the early stage of protein folding. Let us recall the equation again and write it in a short notation where it is clear what is a dependent variable and what is not:

$$s_{ij}(t+1) = sgn\gamma[\sum_{kl} W^{o}_{ijkl}s_{ij}(t) + \eta_{ij}] \tag{52}$$

where γ is a temperature factor and η is, in the terminology of neural networks, a threshold which has absorbed all the interaction terms W^{I}_{ijkl} and the constants W^{o}_{ij} etc.

The right hand side of this equation can only match the left hand side if the contact variables s basically are the same or, the other way around, if the contacts are the same (e.g. the same over a certain range of residues and otherwise zero) the equation above is full-filled and has a fixpoint at that contact value. The fact that the contacts only have to be equal over a certain range of the sequence and otherwise zero means that the chain or protein back-bone has to have periodic structures along the chain. Such periodic structures are typically secondary structures and give rise to fixed point solutions. We shall now try to show that the typical secondary structure classes of helices and sheets naturally comes out as solutions of the polymer chain topology.

The task is to identify the possible geometrical configurations (see fig. 2) of two contacts interacting and determine the factor W^{o}_{ijkl}. Therefore we consider a chain of length N on which there are two contact configurations (i_1, j_1) and (i_2, j_2) which are separated by $L = j_1 - j_2$ and with contact orders $k_1 = j_1 - i_1$ and $k_2 = j_2 - i_2$. We can assume $k_1 \leq k_2$. One needs to calculate the correlation function (see ref. 11):

$$R_2 = \frac{R(N; i_1, j_1, i_2, j_2)}{R(N; i_1, j_1)R(N; i_2, j_2)} \tag{53}$$

where $R(N; i, j)$ is the reduction factor, introduced in chapter 3., when having the contact (i, j).

First we realize, as mentioned in the previous chapter, that there are (topologically) only 3 types of possibilities (see fig. 2) for the index configurations: 1. case: $0 \leq L \leq k_2 - k_1$, 2. case: $k_2 - k_1 < L < k_2$, 3. case: $L \geq k_2$, when we assume positive separation $L > 0$.

In the Jacobsen-Stockmayer approximation (ref. 10) the reduction factor can be derived to be:

$$R_2(L) = \begin{cases} [k_2/(k_2 - k_1)]^{d/2} & \text{if } 0 \geq L \geq k_2 - k_1 \\ [1 - (k_2 - L)^2/(k_2 - k_1)]^{-d/2} & \text{if } k_2 - k_1 < L < k_2 \\ 1 & \text{if } l \geq k_2 \end{cases}$$

If the contact-orders are equal, and hence $k_1 = k_2$, case 1 is absent. The importance of the above classification is that it is complete. In the case of periodic structures it is clear that antiparallel β-sheet types are of case 1., α-helices, π-helices and parallel β- sheet of case 2, while random coil (periodic or not) are of case 3. For negative L, $L < 0$, the case can be transformed into the former case by the symmetry:

$$R_2(L) = R_2(k_2 - k_1 + \mid L \mid) \tag{54}$$

There is a natural relation between these cases and the secondary structures. The periodicity in the helical structure means that the orders for the non-zero contacts are equal, i.e. $k_1 = k_2$, and that itself exclude case 1. For regular sheet structures the non-zero contact orders are regular series of numbers such that $k_1 = k_2 + k = k_3 + 2k$ etc. occuring in case 1. The question about parallel and anti-parallel structures are determined by a twist on the contact loops.

Including excluded volumen effects basically mounts to multiplying R_2 with a factor D whose coefficient f can be determined from Feynman rules: $D = (1 - \nu_o/2\pi f)$.

We can now under the assumption of some periodic structure go back and solve the complicated summation of the evolution equation and determine the condition for stable fixpoints related to those structures. Therefore we assume that our test protein has some general periodic structure $\{(m_1^o, m_1), ..., (m_K^o, m_K)\}$, where K stands for the number of periodic structures and each structure is labelled by a start position m^o and number of residues m it contains. Non-periodic regions are labelled by m_n. We can now restructure the summation in the evolution equation:

$$\begin{aligned} s_i(t+1) &= sgn\gamma[\sum_{j=1}^{K}\sum_{j=1}^{m_n}\sum_{j=1}^{m_K} W_{ij}s_j(t) + \eta_i] \\ &= sgn\gamma[\sum_{j=1}^{K} m_k \cdot W_{ii+p}s_j(t) + \sum_{j=1}^{m_n} W_{ij}s_j(t) + \eta_i] \end{aligned} \tag{55}$$

where $W_{ij} = logR_2$ is given by one of the 3 cases. We can neglect the non-periodic part since σ_i in that case is zero (early folding stage).

In the real case of contacts s_{ij} with two indices we have similarly:

$$\begin{aligned} s_{ij} &= sgn\gamma[\sum_{kl}^{K} m_{kl}W_{i,j,i+p,j+p}s_{ij} + \eta_{ij}] \\ &= sgn\gamma[\sum_{l}^{K} m_l W_{i,i+p,i,i+p}s_{i,i+p} + \eta] = s_{i,i+p} \end{aligned} \tag{56}$$

where p is the periodicity (i.e. number of residues in each period) and the K periodic structures were assumed not to have contacts between them and where there was assumed no contact outside the periodic structures. Furthermore the threshold for defining a contact ρ was assumed smaller than $2p$ and even smaller than $2h$ where h is the characteristic hydrogen bond length for the given periodic structure.

If the periodic part is denoted p_i the existence of a fixed point is depending on how big the threshold η_i is relative to p_i, and how big the temperature factor γ is. This is seen by looking at a graphical representation of the evolution equation. The fixpoint is given where the line $s_i = s_i$ crosses the function $tanh$. Basically there exists a non-trivial fixed point provided $p_i > \eta$ and $\beta > 1$.

To determine whether the fixed point is attractive we have to solve the stability equation or the following inequality:

$$\delta[tanh(\beta(< s_i > +\eta_i)] = \frac{\partial tanh(\beta(< s_i > +\eta)}{\partial < s_i >}\delta < s_i > < \delta < s_i > \tag{57}$$

which can only be fullfilled for $\beta > 1$.

5a5. Basins of attraction of neural network models of protein dynamics

We have shown that the dynamical equations for contact formation can be formulated as a neural network with an interesting architecture. One of the values of this approach is it allows us to understand the relationship between the approaches based on feed-forward models taken by the Copenhagen group and others, and the associative memory and spin glass models that have been studied so much in both a practical way and by the use of equilibrium statistical mechanics (ref. 4 and 14-19). The other outcome of this analysis is that we are now in a position to take over many of the ideas applied to spin models of feed-forward neural nets (e.g. ref. 33, 34) in order to think about the early stages in protein folding. The strict down-hill dynamics discussed in the last chapter which is equivalent to a neural net would amount, in physical terms, to folding in very short amounts of time of the order of nanoseconds to microseconds. The down-hill dynamics really models the early events of folding while the spin glass analysis focus on the longest timescale processes.

We will point out how some of the formal analysis of neural networks can tell us something about the basin of attraction of this fast downhill motion. This can be examined rather directly if the associative memory instruction model is used for the contacts.

Another question that often arises is whether there could exist a code for rapid downhill folding of proteins. We see that in this representation this question is entirely analogous to the problems raised by Elizabeth Gardner in asking about the statistical mechanics of the space of interactions in the feed-forward neural nets (ref. 12). Considerations of that type have also lead one to ask about optimal learning strategies and we will say a bit about this.

5a5a. The dynamics of overlaps in the associative memory model

Consider the case where the linear associative memory rule is used to develop the interaction Hamiltonian. In the contact representation, we can see that the Hamiltonian is the same as a regular partially frustrated spin glass with an external field which is, in the main, ferromagnetic, i.e., favoring a single configuration and a random part coming from other memories:

$$\mathcal{H} = -\sum_{ijkl} W^o_{ijkl}s_{ij}s_{kl} + \sum_{ij} W^o_{ij}s_{ij} + \sum_{\alpha kl} \gamma^\alpha_{ijkl}s^\alpha_{kl}s_{ij} \tag{58}$$

where the second term $\sum_{ij} W^o_{ij}s_{ij}$ represents an external field and the third term $\sum_{\alpha kl} \gamma^\alpha_{ijkl}s^\alpha_{kl}s_{ij}$ is the random part. The corresponding evolution equation for the spins in the fast downhill motion is:

$$\begin{aligned} s_{ij}(t+1) &= sgn[\sum_{kl} W^o_{ijkl}s_{kl}(t) + W^o_{ij} + \sum_{\alpha kl} \gamma^\alpha_{ijkl}s^\alpha_{kl}] \\ &= sgn[h_{ij} + h^o_{ij} + \tilde{h}_{ij}] \end{aligned} \tag{59}$$

This equation corresponds to a single spin reorienting the field, partly coming from the correct instructions, partly from the neighbouring interacting spins, and a random part from the incorrect instructions. At this level, the evolution equation, therefore, looks

like the evolution equation in a conventional feed-forward neural net that carries out association and we can follow precisely the analysis given by Amit et al., ref. 35, and Krauth et al., ref 36, 37, to write down an equation for the overlap between a contact and the correct pattern:

$$m_{ij}(t+1) = \Phi(\frac{\sum_{kl} W_{ijkl} m_{kl} + h^o_{ij}}{\sqrt{\alpha r}}) \tag{60}$$

similarly to the spin overlap equation:

$$m_i = \Phi(\frac{m_i}{\sqrt{2\alpha r}}) \tag{61}$$

where Φ denotes the error-function and r is the parameter of overlap with non-condensed memories. Written in the integral form our overlap equation looks like:

$$m_{ij} = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} exp(-z^2/2) tanh[\beta(\sum_{kl} W_{ijkl} m_{kl} + h^o_{ij} + \sqrt{\alpha} z)] \tag{62}$$

This is however a complicated matrix equation to solve in general. The size of the basins of attraction are determined by the fixed points of this equation. If there are few incorrect memories, it is clear that there will be a fixed point near $N = 1$, and the size of the basin of attraction can be determined from the slope of the error-function near that point.

A way to get a very crude solution to the matrix equation is to start with a strongly simplyfied picture of the contacts and then iterate the overlap equation. Firstly one could start with all overlaps being zero at time $t = 0$ and then solve for the overlap at the next time instance $t = 1$. One then obtains a value for the overlap at a later stage by inserting the overlap value $m(t)$ at time $t = 1$ in the right hand side of the overlap equation and assuming that the overlap is constant. After some iteration steps one can then try to determine the fixed points of the theory.

Therefore the first equation for $m(1)$ is:

$$m_l(1) = \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-z^2/2} tanh[h^o_l + \sqrt{\alpha} z] \tag{63}$$

which right hand side has to be inserted in the next equation at a later time:

$$m_l(2) = \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-z^2/2} tanh[n \cdot a \cdot m(1) + h^o_l + \sqrt{\alpha} z] \tag{64}$$

where the constant a is determined from the polymer coefficients W_{ijkl}, and n stands for the number of contacts in the overlap. (l is an index for the type of contact.)

It is important that all the parameters in the equation above can be fixed essentially by assuming a specific polymer geometry, i.e. by assuming a specific type of secondary structure contacts. We shall first assume that the only possible contact is alpha-helical or no contact. Therefore the parameters a, n and h^o are determined from polymer dynamics. The parameter α is basically measuring the size of the set of "bad" or non-condensed "random" memories. The critical value of α, i.e. when the memory capacity limit is reached and the network breaks down can in principle be determined from the folding entropy and folding transition temperature. We shall determine α by our numerical fixed point analysis.

The important task is now to look for fixed points. They can be found by plotting $m_1(2)$ as a function of $m_1(1)$ and then look for points (fixed points) where the curve intersects the straight line given by $m_1(1) = m_1(1)$. It turns out that we only find attractive ($\beta > 1.0$) fixed points ($m_1 = f_o$ where $0.02 < f_o < 0.1$) for values of α in the range $0.0 < \alpha < 1.0$ and temperatures in the range $10.0 < T < 0.0$.

Once the non-trivial fixed points are determined the size if the basins of attraction can be calculated corresponding to each fixed point. The size is found as the range of m_1 for which the slope of the curve is less than 45 degrees, see figure 3, where the overlap $m(t+1, m(t)) = m_1(2, m_1(1))$ is plottet for different values of temperatures T and α values denoted by a. For example for the fix-point at $m_1 = 0.7$ for $\alpha = 0.5$ and $T = 0.7$ the size of the basin of attraction is around 0.3, which is quite large. The sizes of the basins of attractions are in some cases of large α and T very small, so the space of interaction looks like containing very steep wells of attraction situated at specific polymer configurations. In Figure 8 we show plots of $m(t+1)$ as a function of $m(t)$. There are curves corresponding to different temperatures. The crossing between the diagonal line ($m(t+1) = m(t)$) and the curves indicate the position af a non-trivial fixed point apart from the one in Zero. The size of the basin of attraction can be read out from the slope of these curves near the fixed point.

5a5b. The evolution of competing structures

We could also try to solve the overlap equations with competing overlap structures and study the time evolution. If we for example start out with two initial overlap structeres $m_1(1)$ and $m_2(1)$:

$$m_1(1) = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} tanh[\beta(h^o_1 + \alpha_1 z)] \tag{65}$$

and

$$m_2(1) = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} tanh[\beta(h^o_2 + \alpha_2 z)] \tag{66}$$

these solutions could then together with a choise of the polymer synaptic weights W_{ijkl}, for example an alpha-helix pattern given by the constant w_{11}, a beta-sheet pattern represented by the constant w_{22} and a mixed term w_{12}, be inserted in an iterative proceedure represented by the coupled integral equations evolving in time:

$$\begin{aligned} m_1(t=2) &= \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} tanh[\beta(n_{11} w_{11} m_1(1) + n_{12} w_{12} m_2(1) + h^o_1 + \alpha_1 z)] \\ m_2(t=2) &= \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} tanh[\beta(n_{21} w_{21} m_1(1) + n_{22} w_{22} m_2(1) + h^o_2 + \alpha_2 z)] \end{aligned} \tag{67}$$

where n_{ij} measure the number of overlaps in each pattern (ij). This coupled system is solved numerically by inserting the solutions at time $t = i$ into the equations for $t = i + 1$ and continue a certain number of time steps. It turns out that for a resonable choice of parameters the patterns will quickly stabilize after, say 10 time steps, while for very unrealistic values of parameters the patterns will oscilate rapidly taking two distinct values for the overlap each occurring at every second time step. This usually happens below the glass transition temperature t_g. A typical evolution of helix and sheet competition is almost seen in any simulation. We could also have tried to iterate the time-dependent dynamical evolution equation and study the evolution of two different compeeting structures.

5a5c. The generic phase diagram

Another important task is to try to get an understanding of the phase diagram for our contact model of the rapid protein folding process of the early stages. We shall be using a strongly simplyfied picture of a protein just being a uniform segment of helix and sheet structure like in the last subsection. This is in order to make the contact dynamics simpler such that the critical quantities, for example the glass transition and folding temperature can be derived. It is in general interesting to see if the phase diagram of the model has any generic structure in common with similar models such as the random energy model or the random field Ising model. We find that the phase diagram will be of the form calculated in figure 9. There are several possibilities for phases depending on the position relative to the folding transition temperature. For example correctly folded and mis-folded phases can occur and there can also be a frozen glassy phase. Beside that there is an intermediate structural phase often called the Molten Globule phase. The phase diagram should be quite similar to that of the random energy model, see ref. 14. An important question is of course what role the frustration plays in our model an if it changes the phase diagram. We remind the reader that our folding model in terms of spin terminology is like a ferro-magnetic, frustrated spin system with a random term. on the other hand the random energy model with not have frustration and that fact is expected for example to change the folding transition temperature relative to the glass transition temperature.

The folding transition temperature is calculated from a comparison of the free energy of the folded and un-folded state. In our simple example of a protein consisting of just an alpha-helical segment and a beta-sheet we can write the free energy as in chapter 4:

$$\begin{aligned}\mathcal{F} = &+ k_B T(\sum_{ij}\sigma_{ij} ln\sigma_{ij} + W^o_{ij}\sigma_{ij} + W^o_{ijkl}\sigma_{ij}\sigma_{kl}) + ... \\ &- \sum_{\alpha(good.mem.)} \epsilon\sigma^\alpha_{ij}\sigma_{ij} - \sum_{\alpha(bad.mem.)} \delta\epsilon\sigma_{ij} \qquad (68)\end{aligned}$$

where we can write $\delta\epsilon = \sum_{\alpha(bad.mem.)} \epsilon\sigma^\alpha_{ij}$ and where $\delta\epsilon = \sqrt{N_{bad}}\epsilon$ (N_{bad} being the number of bad memories and which is proportional to our parameter α from the last subsection).

Now we specialise to our case of one alpha-helix and one beta-sheet and the contacts we should include are therefore: $\sigma^{\rho(\alpha)}_{i,i+4}$ and $\sigma^{rho(\beta)}_{i,j>4}$.

We shall now use a trick for resuming higher orders if the expression for the free energy. We shall be using the stoichiometry constraints on the probability numbers n_α, n_β for alpha-helix or beta-sheet occurrence:

$$0 < n_\alpha + n_\beta < 1 \qquad (69)$$

which then allows us to resum $\mathcal{F}$:

$$\begin{aligned}\mathcal{F}^{resum}_o &= k_B T(\sum_i \sigma^{\rho(\alpha)}_{i,i+4} ln\sigma^{\rho(\alpha)}_{i,i+4} + \sum \sigma^{\rho(\beta)}_{ij} ln\sigma^{\rho(\alpha)}_{ij} + (1 - \sum_j \sigma^{\rho(\beta)}_{ij} - \sigma^{\rho(\alpha)}_{ij}) ln(1 - \sum_j \sigma^{\rho(\beta)}_{ij} - \sigma^{\rho(\\ &+ \sum_{ij} W^{o\alpha}_{ij}\sigma^\alpha_{ij} + W^{o\beta}_{ij}\sigma^\beta_{ij} + \sum_{ijkl} W^{o\alpha\beta}_{ijkl}\sigma_{ij}\sigma_{ij} - \sum_{\alpha(good.mem)} \epsilon\sigma^\alpha_{ij}\sigma_{ij} \qquad (7\end{aligned}$$

Next we shall have to find approximate values for the polymer factors:

$$\begin{aligned}\frac{1}{k_bT}W^o_{ij} &= \gamma - ln \mid i - j \mid^{d/2}; \gamma = ln(\frac{\triangle v}{D^3}); D \sim \sqrt{2}l_p \\ W^\beta_{ij} &= W^\alpha_{ij} - ln(\frac{i-j}{(i-j)_\alpha})^{3/2} \\ \alpha - case &: W_{ijkl} \sim W_{i<j>k<l>} = W^{\alpha\alpha}_{ik} \sim 3/2 k_B T(k = 1,2,3,4) \\ \beta - case &: W^{\beta\beta}_{ik} \sim 2W^{\alpha\alpha} \sim 6/2 k_B T \qquad (71)\end{aligned}$$

where d is the dimension of the configuration space, l_p is the persistence length and γ is a parameter that can include hydrogen bond contributions. It is important to note that the theory in this approximation just has three parameters $\gamma, \epsilon, k_B T$.

The folding transition temperature T_F is now found by setting $\mathcal{F}_o = 0$ and using the contact variables for the folded structure. Let us find the temperature T_F for different cases of simplified situations.

1. In the case of $\gamma = 0$ and just alpha-helical structure, $n_\alpha = 1$, we get:

$$\gamma = 0 : k_B T_F = \frac{N n_{\alpha(good)}\epsilon}{N \cdot 4} = n_{\alpha(good)}\epsilon/(-log4 + 9/2) \qquad (72)$$

2. Similarly in the case of just beta-sheet structures and $\gamma = 0$ we get:

$$\gamma = 0 : k_B T_F = \frac{N n_{\alpha(good)}\epsilon}{N \cdot -log8 + 9/2} = n_{\alpha(good)}\epsilon/(-log8 + 9/2) \qquad (73)$$

3. In the case where half of the content is alpha-helical structure, $n_\alpha = \frac{1}{2}$, and beta=sheets, $n_\beta = \frac{1}{2}$, we have:

$$\gamma = 0 : k_B T_F = \frac{n_{\alpha(good)}\epsilon}{log1/2 - log4 - log8 + 9/2} \qquad (74)$$

By choosing a sufficient number of configurations and then solve the equation above for the corresponding values of T_F one can construct the phase diagram of this two-component model in terms of temperature and the ϵ parameter, see figure 5. The constructed phase diagram has much in common with the random energy field model with a destinct Molten Globule phase of a mixture of good and bad memories and separated from the folded and un-folded phase. Above the critical capacity of $\epsilon \sim \alpha = 1.1$ the folded phase is replaced with a glass state where the protein has fallen into a wrong minimum.

To conclude, we can numerically study the evolution of folding of simple patterns from the feed back neural network equations by an iterative procedure described above. In the very simple case of studying recall of a desired (good) memory pattern, describing a particular secondary structure, and besides having a number of undesirable (bad) memory patterns stored too, we find that the correct structure can be retrieved at a temperature (signifying the folding transition temperature T_G) of around $k_BT = 2$ and with up to around 5 "bad" memories (for a numerical set-up with quite large protein chains, unlike the study above where non-equilibrium properties are analyzed. Since the proteins are only described by simple contact patterns for the backbone we envisage that adding the side chain specificity mounts to approximately a ten-fold increase of the memory capacity. Hence we end up with the result that during the early stages of the folding processes the protein can collapse into 50 specific classes corresponding to the classes or families known in the literature, ref. 41.

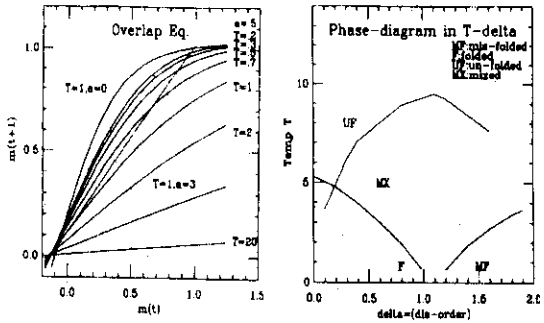


Figure 8,9.: Fig.8: The overlap equation solved numerical at different temperatures. Fig.9: The Phase diagram showing the different protein phases, altogether 4.

5a6. Discussion and future prospects

We shall at this stage summarize the results we have obtained from our analysis and briefly mention some problems of interest for a more general investigation. The down-hill dynamics of associative memory nets has been studied by Amit et al., ref. 35, and by Krauth et al., ref. 36, 37. Their analysis provides a set of dynamic recursion for the overlap of a spin configuration with a memory. The analogous development for the protein dynamic equation that was studied in the previous chapter can answer such questions as: 1. For a certain associative memory Hamiltonian when does a given structure exist as a fixed point? 2. How close to this fixed point must one be for a direct free energy minimization to work? 3. On the average what fraction of contacts can be formed quickly? 4. Does this for example explain the strange coincidence between the fraction of secondary structure found in the molten globule and the prediction rate of the best neural nets for secondary structure determination (around 70 %).

Another approach would be to ask whether any assignment of the W_{ij}^l can give rapid folding for all natural sequences. In the present spin framework this can probably be answered using the ideas pionered by Elizabeth Gardner about the space of connections in nets, ref. 12, 38, although a space of thresholds is apparently the proper concept for protein folding.

Finally it is clear that the optimal code for a feed-forward net would involve maximizing the size of the basin of attraction of the folded structure. The relation of this code to codes based on optimizing the long time features captured by spin glass models will be interesting.

List of references to chapter 5a.

1. Qian, N. and Sejnowski, T. J., J. Mol. Biol. 202, 265 (1988).

2. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, Olsen, O. H. and Petersen, S. B.,FEBS Lett. 241, 223 (1988).

3. Holley L. H. and Karplus, M., Proc. Natl. Acad. Sci. USA 86, 152 (1989).

4. Friedrichs, M. S., and Wolynes, P. G., Science 246, 371 (1989).
Friedrichs, M. S., and Wolynes, P. G., Tetrahedron Comp. Meth., 3 (1990).
and Friedrichs, M. S., Goldstein, R. A., and Wolynes, P. G., J. Mol. Biol., in press (1991).
5. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, Lautrup, B. and Petersen, S. B.,FEBS Lett. 261, 43 (1990).

6. Bryngelson, J. D., Hopfield, J. J., and Southard, S. N. Jr., Tetrahedron Comp. Meth., 3, 129 (1990).

7. Wilcox, G. L., Poliac, M., and Liebman, M. N., Tetrahedron Comp. Meth., 3, 191

(1990).

8. Wolynes, P. G. , Search and recognition: Spin glass engineering as an approach to protein structure prediction. Cargese Summer institute (1990).

9. Bycroft, M. Matouschek, A., Kellis, J. T., Serrano L., and Fersht, A. R.,Nature 346, 344 (1990).

10. Ptitsyn, O. B., Pain, R. H., Semisotnov, G. V., Zerovnik, E. and Razgulyaev, O. I., FEBS Lett 262, 20 (1990).

11. Shakhnovich, E. I., Finkelstein, A. V., Biopolymers 28, 1667 (1989).

12. Gardner, E., J. Phys. A 21, 257 (1988).

13. Bohr, H., Goldstein, R. A. and Wolynes, P. G., AMSE Periodicals, Modelling, measurement and control, C, Vol.31, 55 (1992).

14. Bryngelson, J. D., and Wolynes, P. G., Proc. Nat. Acad. Sci., U.S.A., 84, 7524 (1987).

15. Garel, J. R., Garel, T., and Orland, H. J. Physique, 50, 3067 (1989).

16. Garel, T., and Orland, H., Europhys. Letters, 6, 307 (1988).

17. Garel, T., and Orland, H., Europhys. Letters, 6, 597 (1988).

18. Sasai, M., and Wolynes, P. G., Phys. Rev. Letters, 65, 2740 (1990).

19. Shakhnovich, E. I., and Gutin, A., Studia Biophysica, 132, 47 (1989).

20. Bohr, H. and Wolynes, P. G., "The initial events of protein folding from an information processing viewpoint", University of Illinois preprint (To appear in Phys. Rev. A.) (1992).

21. Oxtoby, D., in Liquids, Freezing and the Glass Transition, ed. Hansen, J. P. , Levesque, D., and Zinn-Justin, J., pt. 1, p. 147-192, North Holland (1990).

22. Freed, K. F.: Renormalization group theory of macromolecules, J. Wiley and Sons, New York (1987).

23. Oono, Y., and Freed, K. F., J. Phys. A 15, 1931 (1982).

24. Mc Coy, J. D., Honnell, K. G., Schweizer, K. S., and Curro, J. G., Chem. Phys. Lett, 179, 374 (1991) and Univ. of Illinois preprint (1991).

25. McMullen, W. E., and Freed, K., J. Chem. Phys., 92, 1413 (1990).

26. Tang, H., and Freed, K. F., J. Chem. Phys., 94, 1572 (1991).

27. Andersen, H. C., Cluster Methods in Equilibrium Statistical Mechanics of Fluids, in Statistical Mechanics, Ed. Berne, B. J., Plenum Press, Vol 5, p.1 (1977).

28. Jacobson, H. and Stockmayer, W. H., J. Chem. Phys. 18, 1600 (1950).

29. Chan, H. S. and Dill, K. A., J. Chem. Phys. 92, 3118 (1990).

30. Karplus, M., and Weaver, D., Nature 160, 404 (1976).

31. Bryngelson, J., Stolorz, P., and Lapedes, A., Los Alamos Preprint 1991.

32. Steeg, E. W., University of Toronto technical report, CRG-TR-90-4 (1990).

33. Müller, B. and Reinhardt, J.: Neural Networks, An Introduction, Springer Verlag (1990).

34. Hertz, J., Krogh, A. and Palmer, R. G.: Introduction to the theory of Neural Computation, Addison Wesley Co (1990).

35. Amit, D. J., Gutfreund, H. and Sompolinsky, H., Ann. of Phys., 173, 30 (1987).

36. Krauth, W., and Mezard, M., J. Phys. A 20, L 745 (1987).

37. Krauth, W., Nadal J. P. and Mezard, M., J. Phys. A 21, 2995 (1988).

38. Gardner, E. and Derrida, B., J. Phys. A 21, 271 (1988).

39. Bohr, J., Bohr, H. G., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Petersen, S. B.:"Protein Structures from Distance inequalities",submitted to Journal of Mol. Biol. (1992).

40. Private communication with Dr. Tugrul, University of Arizona (1992).

41. Richardson, J. S., Adv. Protein Chem. 34, 167 (1981).

42. Wang, J., Bohr, H. G. and Wolynes, P. G., A Field Theoty for growth of Domains in Stages of Protein Folding, (University of Illinois Preprint, 1992).

5b. Domain growth in protein folding.

The main aim of the work in this chapter is to study the formation of contacts and structures in protein folding in the language of statistical mechanics of domain growth. Important for such a discription is the appearant analogy between protein folding and Ising spin systems with randomness.

We shall first construct a general basis for studying domain growth by chosing an appropriat representation for the patterns of 3-dimensional protein structure. Such appropriate representation is given by the distance matrix geometry in which contacts between any pair of residues are the fundamental variables. Earlier we were able[?] to describe development of contacts between residues in biopolymers during the early stages of folding. Here we shall extend that study to domain growth and somewhat beyond the early stages.

In a distance matrix geometry one can study growth of domains of contacts which in turn stands for specific 3-dimensional configurations of protein structures. The energy function used for describing contact formation in protein molecules is constructed from polymer dynamics containing entropy and interactions among smooth strings[?, ?] and with side-chain interactions contained in memory terms of known protein structures.

We can describe early stages of protein folding by a set of first order differential equations standing for down-hill energy processes. These equations can be solved analytically if we assume that they are only weakly coupled but else, in the general case of a fully coupled system of equations, we will have to resort to numerical computer simulations.

The pattern of these contacts can easily be translated into the corresponding 3-dimensional structure of the protein. The contacts are usually represented by a 2-dimensional distance matrix with the collumns and rows designating the sequence of residues, as shown in figure 10 in the case of the native structure of $6pti$. The important thing in this 2-dimensional distance matrix representation of residue contacts is an ideal framework for describing growth of domains of contacts. As in a lattice gas a "positive" definite contact corresponds to a lattice site being occupied and no contact to an empty site. In turns the growth of domains of positive contacts signifies the formation of a given substructure in the protein, and is relatively easy to study in terms of domain growth laws and scaling. Thus by applying such domain growth techniques we can analyze how substructures in the protein get formed and thereby how the protein folds. Furtheremore the analogy to neural networks and chemical activation processes can be used to steer the domain growth since the neural network methodology gives prescriptions for how to include and use protein "memories" as a scafolding and nucleation of protein structure formation.

The main results coming out of the domain growth analysis and through a numerical study is that the domain growth of protein substructures in the early stages of the folding processes is governed by power laws while later stages exhibit a slower logarithmic growth. Furtheremore the nucleation in the start is predominantly of a local nature which means that the earliest contacts are formed between neighbouring residues. These facts resemble the dynamics of first order phase transitions in disordered media where bubles (here contact domains) grow slowly in time. What concerns the formation of secondary structures the helical patterns are starting to form earlier but once the sheet patterns are nucleated they grow faster.

We shall first introduce the basic framework for studying protein contacts and construct the theory for evolution of these contacts.

5b1. Domain growth

In order to make a good semiquantitative illustration of domain growth we consider first the continuous case even though the descrete case may be more relevant to proteins.

Basically we are interested in the dynamical equation for fast down-hill processes and we consider a kind of Langevin equation whithout the noise term. A similar equation arises when we are to study domain growth which is a non-equilibrium thermodynamical process. The corresponding non-equilibrium equation of motion is derived from phenomenological thermodynamics by using the suitable order parameter ϕ and equate its displacement with the present thermodynamical force. We therefore write the equation for the non-equilibrium thermodynamical domain growth as:

$$\frac{d\phi}{dt} = -\Gamma\frac{\delta\mathcal{F}}{\delta\phi} \tag{75}$$

with $\frac{d\mathcal{F}}{dt} < 0$ and where Γ controls the time scale of the system.

Next we need a usefull quantity for studying the growth of domains. We shall here be using the domain size of contacts as a sort of radial parameter:

$$R^2 = \int \phi(x,y)dxdy \tag{76}$$

5b2. Fast growth

In this section we study the very first stage of the process where we have no random term h.

We consider our energy functional which in the continuous case becomes:

$$F = \frac{1}{2}\nabla\phi^2 + \frac{\mu}{2}\phi^2 + \frac{b}{4}\phi^4 + H\phi \tag{77}$$

in the case of an Ising Ferro magnatic system, and the kinetic equation in d dimensions is:

$$\frac{d\phi}{dt} = \Gamma[K\frac{\partial^2\phi}{\partial r^2} + K\frac{(d-1)}{r}\frac{\partial\phi}{\partial r} - \frac{\partial\mathcal{F}}{\partial\phi} + H] \tag{78}$$

which is derived from the equation: $\frac{d\phi(r)}{dt} = -\Gamma\frac{\partial\mathcal{F}}{\partial\phi}$ and where ϕ again is the order parameter (the magnetization) and where $\phi \sim \phi(r - R(t))$. (we can put $\Gamma = 1$).

Our equation in 2 dimensions, with $\phi = \phi(r - vt)$, is (in the sperical symmetric case):

$$K\frac{d^2\phi}{dr^2} + (K\frac{(d-1)}{R} + \frac{v}{\Gamma})\frac{d\phi}{dr} - \mu\phi - b\phi^3 - H = 0 \tag{79}$$

and by integration we get:

$$\Delta(f - \phi H) = 2\phi_* H = \frac{v}{\Gamma}\frac{\sigma}{K}\frac{(d-1}{R} + \frac{K(d-1)}{R^*} \tag{80}$$

In these equations Δ is the change in the argument when crossing the interface, σ is the equilibrium surface tension and ϕ_* is the minimum of f, $\phi_* = \sqrt{-\frac{\mu}{b}}$. We have introduced an effective radius where the energy is maximized:

$$R^* = \frac{1}{2}(d-1)\frac{\sigma}{\phi_* H} = \frac{(d-1)}{c} \tag{81}$$

We can now follow the prescription from the standard theory of domain growth [?] and write the growth equation for the domain boundary as follows:

$$\begin{aligned} v = \frac{dR}{dt} = \Gamma(\frac{1}{R^*} - \frac{1}{R}) \quad &\Rightarrow \\ \frac{R^* R dR}{R - R^*} = \Gamma dt \quad &\Rightarrow \\ R^*(R + R^* log(R - R^*)) = \Gamma t + c_o \end{aligned} \tag{82}$$

In the case of large R: $R > logR$ we have: $R^* R = \Gamma t$, i.e. linear growth.

In the case of small R: $R <\mid logR \mid$ we have:

$$(R^*)^2 log(R - R^*) = \Gamma(t - t_o) \tag{83}$$

so in this limit:

$$R - R^* = e^{\frac{\Gamma}{(R^*)^2}(t - t_o)} \tag{84}$$

In the late time limit the behaviour of ϕ will have exponential growth and the field ϕ will spread out and grow like a spherical bubble.

So far we have been treating the cases of spherical symmetry but if we also want to extend the study to aspherical cases we could replace the term $\frac{1}{R}$ in the equation above with the mean curvature K for the curve that is describibg the profile of our domain of contacts.

5b3. Slow growth

Next we would like to examine the growth when we have included the random term $\eta(x, y, t)$ We shall also like to give some physical scenario of what happens when a random term is introduced, see also [?, ?].

Concerning barrier crossing, if we include the random term h in the energy function we can write the barrier height as a surface potential energy term and a volume term:

$$\Delta = -JR^{d-2}\delta - h(\delta R^{d-1})^{\frac{1}{2}} \tag{85}$$

and minimization with respect to δ (the amount the radius is allowed to shrink) gives:

$$\Delta = \frac{Rh^2}{J} \tag{86}$$

so the time τ it takes to overcome the barrier is:

$$\tau = \tau_o exp((\beta)[\frac{Rh^2}{J}]) \tag{87}$$

and

$$R \simeq \mu\frac{Jk_B T}{h^2} log(\frac{t}{t_o}) \tag{88}$$

Hence we have $R \sim log(\tau)$, i.e. logarithmic growth. The specific protein chemistry is contained in the constant J that basically is a function of μ and hence our contact factor W^o_{ijkl}. It is very important that the power law growth of contact domains in the first stages of the protein folding process, that we are to describe, is slowed down in the later stages and become a logarithmic growth that eventually will stop. This is of course due to the noise term from the spurious memory terms that become more important in the later stages of the process. This conclussion is true for all dimensions $d \geq 2$.

5b4. Domain wall dynamics and the jump condition in the descrete case

We start with the free functional:

$$F = W_{ij}\sigma_{ij} + W_{ijkl}\sigma_{ij}\sigma_{kl} + \sigma_{ij} log\sigma_{ij} + (1 - \sigma_{ij})log(1 - \sigma_{ij}) + W^I_{ij}\sigma_{ij} \tag{89}$$

We use Einstein convention that when the index repeats itself, it means sum over the same index. The dynamical equation for the contacts is:

$$\frac{d\sigma}{dt} = W_{ij} + W_{ijkl}\sigma_{kl} + log\frac{\sigma_{ij}}{1 - \sigma_{ij}} + W^I_{ij} \tag{90}$$

We would like to impose the thin wall ansatz to the problem we study, chosse the proper coordinates our current two dimensional complex problem can be simplified to one dimensional form:

$$\sigma_{ij} = \sigma(i - V_i t) \tag{91}$$

We choose this particular form for the wall ansatz, because α helix and β sheets in our contact coordinates are approximately all one dimenstional perpendicular with each other. i and j are not independent quantities. They are linearly related to each other in

helix and sheets cases. After substituting this ansatz into dynamical equation of contacts, we obtain:

$$-V_i(\sigma_{i+1} - \sigma i) = W_i + W_{ikl}\sigma_{kl} + log(\frac{\sigma_i}{1-\sigma_i}) + W^I_{ij} \tag{92}$$

We multiply each side of equation and then sum over i, and obtain a "jump" condition for the velocity across the boundary of the contact domain:

$$V_i\sigma_0 = \sigma^0_i(W_i + W^I_i) \tag{93}$$

Where σ^0_i is the solution of stationary equation $\frac{\delta F}{\delta \sigma_i} = 0$. σ_0 is the surface energy of the wall:

$$\sigma_0 = \sum_i (\sigma_{i+1} - \sigma)^2 + F(\sigma_{ij}) - F(\sigma^0_{ij}) \tag{94}$$

Since this is an approximate one dimenstional problem, when there is no random source term, we obtain:

$$X_i = A_i t + B_i \tag{95}$$

Here A_i and B_i are constant term depending on W_i. X_i is the position of the wall in sequence space. When we include the random term in, we obtain:

$$X_i = A_i(lnt)^2 + B_i \tag{96}$$

So we see that in contact space, the domain of contacts (helices and sheets) is growing linear in time when there is no random source term, and it is slowly growing logrithmically in time with the random source term. This might be approximately the domain growth law of early stages of folding. In this derivation we have not been using any spherical symmetry assumptions like in the other chapter so in this sense the descrete case in more general and more appropriate for protein applications.

5b5. Numerical studies

In this chapter we present a numerical study of domain growth of protein structures described by the evolution equations. These first order differential equations were derived in the previous chapters and represented the time evolution of contacts between residues in protein during the earlier stages of protein folding. Thus it is in the 2-dimensional plane of contacts, where each point signify a contact σ_{ij} between residues i and j ordered along the axis of the plane, that the domain growth is studied. Thus each domain is consisting of points (i, j) that stand for a close contact between residue i and j, i.e. residues having a distance between them being less than a given threshold. In figure 10 is pictured a typical 2-dimensional distance matrix of contacts σ_{ij} for the folded structure of the test protein $6pti$. We are mainly concerned with contacts related to the most common secondary structures, the helical and the beta strand structures, but the analysis can easily be carried over to other structures. the reason for being primarily concerned with these two types of structures is that they are easily being distinguishable in the contact space where the helical structures are forming close to and along the diagonal $(i = j)$ while the parallel beta-strands are far away and parallel to the diagonal and the anti-parallel beta-strands are structures orthogonal to the diagonal. These circumstances are easily verified on figure 10.

FIGURE 1A

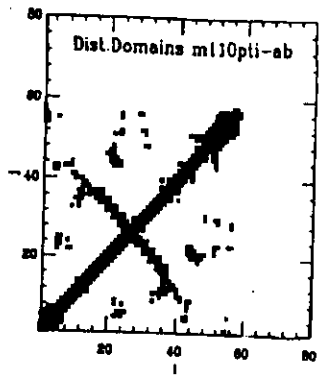


Figure 10.: This figure shows the distance matrix plot of 6PTI.

In the evolution equation equation we have included the polymer factors corresponding to these types of structures. The evolution equations for the contact variables were in the earlier chapters derived from the dynamical equation for energy down-hill processes:

$$\frac{d\sigma_{ij}}{dt} = \frac{\partial F}{\partial \sigma_{ij}} \tag{97}$$

where F is the energy functional

$$F = \sum_{ij} W_{ij}\sigma_{ij} + \sum_{kl} W_{ijkl}\sigma_{ij}\sigma_{kl} + \sigma_{ij}log\sigma ij + (1-\sigma_{ij})log(1-\sigma_{ij}) + W^I_{ij}\sigma_{ij} \tag{98}$$

The 2-loop factors W_{ijkl} can be derived[?, ?] for the specific secondary structures, α, β, under consideration:

$$W^{\alpha}_{ijkl} = log([1 - (\mid n - L \mid /n)^2](\frac{\Delta V)^2}{(n2\pi)^d}); d = 3, L = l - j, n = 3 \tag{99}$$

and

$$W^{\beta}_{ijkl} = log[(\frac{(j-i)}{(k-l-j+i)})^{d/2}(\frac{\Delta V}{(j-i)2\pi})] \tag{100}$$

while the 1-loop factors basically behave like $W_{ij} \sim log[(\frac{1}{2\pi|i-j|})^{d/2}]$ and d being the dimension.

The evolution equation can then, with the stochiometric assumption that there are only α and β secondary structures (i.e. helices and sheets), be written as:

$$\frac{d\sigma_{ij}}{dt} = \sum_{kl} W^{\alpha}_{ijkl}\sigma^{\alpha}_{kl} + W^{\beta}_{ijkl}\sigma^{\beta}_{kl} + W_{ij} + log(\frac{\sigma_{ij}}{1-\sigma^{\alpha}_{ij}-\sigma^{\beta}_{ij}}) + W^{I}_{ij} + .. \qquad (101)$$

and then be approximated by a finite difference equation and integrated on both sides to become:

$$\sigma_{ij}(t+1) = \sigma_{ij}(t) + \triangle(k_B T)[log(\frac{\sigma_{ij}}{1-\sigma^{\alpha}ij-\sigma^{\beta}ij}) + W_{ij} + W^{I}_{ij} + ..] \qquad (102)$$

These equations are solved numerically by calculating the contact variables at a finite time step and integrate. The integration is split up in m integration steps, typically of the order of $m = 100$ to $m = 500$ steps. The integration steps become the time parameter in the evolution of contacts.

We start out with zero contacts at time $t = 0$ and then construct the contacts for all i and j residues at the next time $t+1$. Thus we are able to follow the evolution of contacts up to the time when the desired patterns have converged to a final stage determined by the contact patterns enforced in the equation through the interaction term W^I in the equation. These enforced patterns can be fully folded realistic protein structures or it can be patterns constructed from energy functions calculated in a more general framework of spin glass theories[?]. Basically enforcing patterns from such a framework give the same result with respect to domain growth as adding contacts from real native proteins but the former approach is intellectually more appealing and it is more in the regime of logarithmic growth which makes it possible to study the slower filling of the domains in details. However this approach can only grossly recall the correct protein structure. The growth laws have an exponent that is around half that of the case where patterns are enforced by real protein distance matrices.

Both a "desired" pattern of contacts we wish the trial structure to attain, as well as other structures functioning as noise and a random number generator term, are added to the evolution equation in order to simulate a realistic folding process. The desired pattern is emphasized with a slightly larger factor than the other patterns. We chose Pancreatic Trypsin Inhibitor, $6pti$, as a good test protein for the desired pattern since it had reasonably well-defined and clearly confined secondary structures.

In figure 11 are shown a series of integration steps as evolutionary stages of the domain growth in the distance matrix representation together with snap-shots of the related 3-dimensional protein backbone structure. After a clear stage of nucleation happening along the diagonal of the contact plot predominantly in the middle of the protein stages follow of domain growth of elongated bubles that grow mostly along the diagonal towards the boundaries of each other and then merge together. A later stage is, after the boundaries are marked, filling out or completing each domain.

On figure 12 the growth laws are depictured both on linear and logarithmic paper of the full protein at different temperature and with varying intensity of the random generator. It is clear from fig.2, and also expected from our analytical studies in the proceeding chapter, that the growth laws of the middle stages of domain formation (i.e. not the nucleation or completion stages) is governed by a power law, decreasing in power with decreasing temperature and increasing random factor.

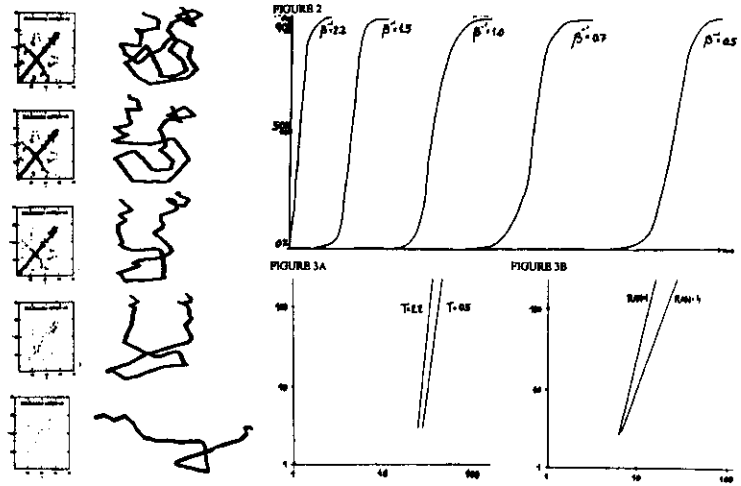


Figure 11,12.: A picture of the growth patterns in 6PTI and the logarithmic growth laws.

References

[1] H. Bohr and P. G. Wolynes:"Protein folding: A physical view of Neural Network Approaches", in Neural Networks: From Biology to High Energy Physics. ETS Editrice, Pisa (1992).

[2] H. Bohr and P. G. Wolynes: The Initial Events of Protein Folding from an Informational Viewpoint", Phys. Rev. A46, 5242 (1992).

[3] H. Jacobson and W. J. Stockmayer, J. Chem. Phys. 18, 1600 (1950).

[4] H. S. Chan and K. A. Dill, J. Chem. Phys. 92, 3118 (1990).

[5] J. J. Hopfield, Proc. Natl. Acad. Sci. USA 79, 2554 (1982).

[6] J. S. Langer, "An Introduction to the Kinetics of First Order Phase Transitions", in Solids far from Equilibrium, (Cambridge Uni. Press) (1991).

[7] R. Bruinsma, "Statics and Dynamics of the Random Field Ising Model (Theory)", in Condensed Matter (Springer Verlag) (1986).

[8] S. Coleman, Phys. Rev. D, 15, 2929 (1977).

[9] T. Nattermann and J. Villain, "Random-Field Ising Systems: A survey of Current Theoretical Views" in Phase Transitions, Vol. 11, (Gordon and Breach Scientific Pub. (1988).

[10] R. Goldstein, Z. L. Schulten and P. G. Wolynes, PNAS, USA, 89, 9029 (1992).

5c. Prediction schemes of protein structure

We shall here discuss some practical implementations of the various neural networks and specific associative memory techniques. The implementations on which we focus try to predict tertiary protein structures on the basis of the protein's sequence of amino acids. These methods will be mentioned first in this presentation in order to motivate and introduce the therminology that is the basis for our later analysis. Most of the detailed results can be found in reference 4, 5, 6, 7 and 13. In some cases we have also done new computer experiments in order to support our analysis. In the first subsection we will briefly mention some work performed with a multi layered feed forward neural network (the Copenhagen network) predicting molecular contacts in the form of distance matrices, similar to a network we shall derive by our mean field equation. In the last subsection we shall deal with some work done with a feed back associative memory system that is used to store and recall memories of protein structures.

5c1. Feed forward neural networks predicting protein distance matrices

Feed forward multi-layered perceptrons have been built and employed, ref. 5, to predict the tertiary structure of proteins. These neural networks were trained on X-ray crystallographic data of protein structures to predict the 3 dimensional structure of folded protein back-bones on the basis of their corresponding primary structure. The input consisted of binary codes representing sequences of amino acids, while the output from the neural networks were binary distance matrices representing inequalities describing the relative 3 dimensional positions of the residues on the back-bone. In a typical distance matrix for a protein the sequence of amino acid residues is numbered along the vertical and the horizontal directions. Every point in the plane corresponds to a correlation (a contact) between two residues, indicating that the two residues are within a certain distance, e.g. 8 Å, to each other. Since the sequence input is ultimately limited by the sequence length, and due to certain computational constraints, the output representing the 3 dimensional protein structure was reduced to a segment of a certain width around the diagonal trace of the C_α distance matrix. After a distance matrix has been predicted the real protein structure is achieved by the use of a computer minimization method that basically consists of a steepest descent algorithm together with some of the most essential chemistry constraints from the back-bone geometry.

The architecture of the network employed in the study of ref. 5 is very much dependent on the form of representation of input and output data that was chosen. Basically the network consisted of an input level with 1220 ($= 20 \times 61$) units, a hidden layer with 300 neuron elements and an output layer with 33 units, such that the first 30 neurons represented a binary number for a contact between the residue in the middle of the (61-unit) input window and 30 residues to the left of this middle residue (see fig. 13). The output data in this configuration then represents a 60-units wide binary band distance matrix. Half of the band is due to reciprocal symmetry. The last 3 units were used for secondary structure output. In one of the test cases the network was trained on 13 very similar (with respect to their function) proteases, and then tested on a similar protease, *2TRM* (73 % homologous to the closest protease in the training set and 223 residue long), that was novel to the network. In a recent test the network was able to predict the double size 2×60 -band distance matrix of *2TRM* up to an accuracy of 97 %, and the minimization could, on the basis of the predicted binary distance matrix generate a structure that had an RMS deviation of about 5 Å compared to the native structure. In an older test (ref. 5) the resulting predicted structure deviated from the correct structure with an RMS deviation of 3 Å when the minimization started from a similar protease *4PTP* instead of starting from a random structure as in the first case (see fig. 14). Today we are able to go down to the homology of 40 % and still maintaining an accuracy of 2 Å RMS. The methodology (neural network and minimization) has no limitation on the size of proteins, and actually the longer the proteins the better the score because larger preoteins have relatively more conserved regions.

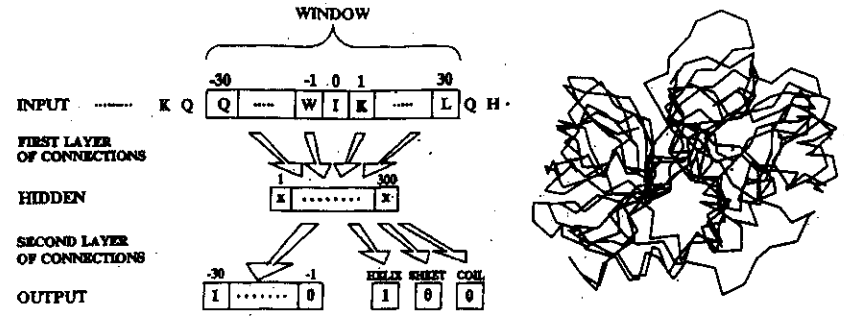


Figure 13,14.: 13: A picture of the Neural Network employed. 14: The predicted structure of 1TRM superimposed on the real.

5c2. Feed back Associative Memory method for predicting protein structure

Next we shall report on another method, motivated by neural network ideas, for determining 3 dimensional protein structures. The basic methodology is more like a feed back Hopfield net. In the ordinary "Hopfield" neural network the Hamiltonian encodes relations of the spins in a set of memory spin states which are then local minima of the resulting Hamiltonian. Thus these states can be recalled by energy minimization methods such that the final chosen memory state is dependent on the initial spin state.

Similarly, in the case of proteins, memories of amino acid correlations are stored in an energy function as local minima that can be recalled through a minimization procedure but the structure corresponding to the free energy minimum should not depend on the initial configuration of the protein chain but rather on the amino acid sequence. Therefore a target protein with a sequence corresponding to one of the memories will have a potential energy minimum at its native folded state. This is achieved by simultaneously encoding amino acid sequence correlations and correlations between pairwise distances among α-carbons. If a pairwise distance matches a memory-distance, the energy is lowered only if the corresponding sequences match. The amino acid sequences are encoded through amino acid charges. The associative memory Hamiltonian that will achieve these goals is written in the form (ref. 4):

$$\mathcal{H}_{am} = \Lambda_{am} \sum_{\mu} \sum_{i<j} \gamma_{ij}^{\mu}(q_i^{\mu}, q_i^T, q_j^{\mu}, q_j^T)\theta(\mathbf{r}_{ij} - \mathbf{r}_{ij}^{\mu}) + \mathcal{H}_o \qquad (103)$$

where Λ_{am} is a scaling constant and $\gamma_{ij}^{\mu}(q_i^{\mu}, q_i^T, q_j^{\mu}, q_j^{\mu})$ is a charge correlation function which originally was chosen to be:$\gamma_{ij}^{\mu} = -(q_i^{\mu} q_i^T q_j^{\mu} q_j^T + q_i^{\mu} q_i^T + q_j^{\mu} q_j^T)$, where q_i^{μ} is a hydrophobic charge for the i'th residue. Furthermore $\theta(\mathbf{r}_{ij} - \mathbf{r}_{ij}^{\mu})$ is a pairwise distance overlap function (e.g. a Gaussian) and finally $\mathcal{H}_o$ is the part that includes the constraints of the back-bone structure (e.g. chain connectivity).

There has been a great deal of work on the equilibrium statistical mechanics of this model and more abstract models related to it (see ref. 14-19).

The Hamiltonian is thus containing a set of representative memory proteins (denoted by the index α) and a target protein charge set (denoted by the index T) whose corresponding 3 dimensional structure is to be found by minimization of the Hamiltonian. Roughly the procedure (ref. 4) is: Once the Hamiltonian is constructed the potential energy is minimized using molecular dynamics with a simulated annealing schedule. First an initial configuration of the protein is chosen using random dihedral angles consistent with Ramachandrian plots. The molecular dynamics (using the Verlet algorithm) are performed using many different temperature values between 1.0 and 0.005 with about 30 timesteps at each temperature. Each time step represents approximately 30 fs of real time. In the Verlet algorithm atom velocities are assigned at every new temperature step consistent with the Maxwell-Boltzmann velocity distribution and constraints, e.g. of the bond lengths, were enforced using a shake algorithm.

The overall system has given resonable results. For example the structure of a 100 residue long protein ($1CCR$) has been predicted with an RMS of around 5 Åprovided the test protein had a sequence homology to the training set of around 60 %. One of the interesting features of this method is that the various folding steps can be studied along with the annealing schedule (ref. 4).

The studies of these practical implementations of neural network techniques to protein structure prediction show that there exist two kinds of variables adequate to describe protein tertiary structure, one set being the α- carbon coordinates and another being the set of molecular contacts encoded in the binary distance matrix that describes which resudues are close to which others. In the following we shall develop a physical theory of the dynamics of the latter variables.

References

1. Qian, N. and Sejnowski, T. J., J. Mol. Biol. 202, 265 (1988).

2. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, Olsen, O. H. and Petersen, S. B.,FEBS Lett. 241, 223 (1988).

3. Holley L. H. and Karplus, M., Proc. Natl. Acad. Sci. USA 86, 152 (1989).

4. Friedrichs, M. S., and Wolynes, P. G., Science 246, 371 (1989).
Friedrichs, M. S., and Wolynes, P. G., Tetrahedron Comp. Meth., 3 (1990).
and Friedrichs, M. S., Goldstein, R. A., and Wolynes, P. G., J. Mol. Biol., in press (1991).
5. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, Lautrup, B. and Pe-

tersen, S. B.,FEBS Lett. 261, 43 (1990).

6. Protein Structure and Chemical Reaction Kinetics

In this chapter we shall use some of all the contact teory we developed in chapter 5 for chemical reaction kinetics in the protein folding processes even though the formalism can be extended to other processes too. It turns out that there almost exists a translation scheme from the contact formalism to kinetic reaction theory. If this formalism is extended to include protein-protein reactions it becomes a theory of molecular recognition which is especially well illustrated by the antibody-antigen reaction complex in immune response processes. Overall one speaks about the chemical reaction network and indeed we shall see that the resulting differential equations describe the evolution of particular Boolean neural networks.

6a. Translation of Contact Dynamics to Chemical Reaction Kinetics

5a4d. An analogy to chemical activation energy

It is interesting that there is actually a deeper reason for the processes of protein folding to resemble the dynamics of neural networks as explained in chapter 5. The reason is a profound connection between chemical reaction kinetics and neural networks. We shall therefore in this chapter give an analogous formulation of our contact dynamics in terms of chemical activation potentials in order to elucidate this neuro-chemistry connection.

We consider a typical allosteric kinetics problem with two states, one (lower) of not being occupied, and one (higher) being occupied and with rate constants for transitions between them being $k_{o\to n.o.}$ and $k_{n.o.\to o.}$ and an activation potential being proportional to the ration of the two rate constants.

If we denote the probability for the occupied state as P_o and for the not-occupied state as $P_{n.o.}$ we can connect these two probabilities to the contact variable σ_{ij}^o as:

$$\sigma_{ij} = P_o + P_{n.o.} \tag{104}$$

or in terms of the activation potential:

$$\frac{\sigma_{ij}}{A-\sigma_{ij}} = e^{E/k_BT} \Rightarrow \sigma_{ij} = \frac{Ae^{E/k_BT}}{1+Ae^{E/k_BT}} \tag{105}$$

which means that the situation of residues i and j being in contact is analogous to being in an occupied state and the situation of no contact is being in the not occupied state. Stoichiometry then tells us that:

$$P_o + P_{n.o.} = 1 \tag{106}$$

We can write the rate of change of the probabilities in terms of the corresponding rate constants $k_{n.o.\to o}$ and $k_{o\to n.o.}$:

$$\frac{dP_o}{dt} = k_{n.o.\to o}P_{n.o.} - k_{o\to n.o.}P_o \tag{107}$$

and introducing the activation energy functional H_A, we have:

$$\frac{k_{n.o.\to o}}{k_{o\to n.o.}} = e^{(\pm)2H_A/k_BT} \tag{108}$$

where the activation energy functional H_A is given in terms of our energy gain expressions W_{ij} for forming contacts (i,j):

$$H_A = W_{ij}^o + \sum_{kl,\rho'} W_{ijkl}^o\sigma_{kl} + W_{ij}^I \tag{109}$$

Hence we can write our equation for fast down-hill kinetics in the form of Glauber kinetics as:

$$\frac{d\sigma_{ij}}{dt} = sgn(H_A^{ij})[e^{-|H_A^{ij}|/k_BT} - 1] - [1 + e^{-|H_A^{ij}|/k_BT}\sigma_{ij}] \tag{110}$$

or with finite time steps:

$$\sigma_{ij}(t+1) - \sigma_{ij}(t) = sgn(H_A^{ij})[-1 + e^{-|H_A^{ij}|/k_BT}] - [1 + e^{-|H_A^{ij}|/k_BT}]\sigma_{ij}(t) \tag{111}$$

where we have changed the differential equation into a difference equation. In the low temperature limit we get:

$$\sigma_{ij}(t+1) = sgn[(k_BT)^{-1}(\sum_{kl,\rho'} W_{ijkl}\sigma_{kl} + W_{ij}^o + ..)]f(H_A) \tag{112}$$

where $f(H_A)$ is given by

$$f(H_A) = (e^{-|H_A|/k_BT} - 1) \to -1 \text{for} T \to 0 \tag{113}$$

This equation describes the evolution of a Boolean neural network similar to the equation of steepest descent for contact dynamics. Thus chemical reaction kinetics can be described as evolution of neural networks. This relation between chemistry and neural dynamics is very much the reason for neural network methodology in protein folding processes that after all are very much governed by reaction kinetics.

6b. Crossing of Activation Barriers and Path Integral Formalism

Let us again start from the dynamical equation for the contact variable with the free functional:

$$\frac{d\sigma}{dt} = -\frac{\partial F}{\partial \sigma_{ij}} \tag{114}$$

We can also write our contact dynamics in terms of a set of chemical kinetic equations. We consider the transition from an occupied state to a nonocupied state, with probability P_{n_o} and define a variable, analogous to the contact variable, as $\sigma_{ij} = P_o - P_{n_o}$ with stoichiometry $P_{n_o} + P_o = 1$. We can write down the rate of change of probabilities in terms of k_{n_o-o} and k_{o-n_o}:

$$\frac{dP_o}{dt} = k_{no-o}P_{no} - k_{o-no}P_o \tag{115}$$

and

$$\frac{d\sigma}{dt} = (k_{no-o} - k_{o-no}) + \sigma(-k_{no-o} - k_{o-no}) \tag{116}$$

The energy cost to change occupation can be obtained from the free energy functional. This gives the activation energy funtional H_a^{ij}:

$$H_A^{ij} = W_{ij}^0 + \sum_{k,l,\rho'} W_{ijkl}^0 \sigma_{kl}^{\rho'} + W_{ij}^I \tag{117}$$

Glauber kinetics then gives:

$$\frac{d\sigma_{ij}}{dt} = sgn(H_A^{ij})(-1 + e^{-|H_A^{ij}|/kT}) - (1 + e^{-|H_A^{ij}|/kT})\sigma_{ij}(t) \tag{118}$$

We make further assumption that $\sigma_k^{\rho'}$ in H_A^{ij} does not change much during the intitial stage of folding. This resembles the spinodol decompostion where the order parameter changes very slowly. We can treat it as constant. W_{ij}^I has a mean value but also a fluctuating part:

$$W_{ij}^I = W_{ij}^G + W_{ij}^F \tag{119}$$

If we assume the fluctuations in W_{ij}^F are Gaussian-white noise, then we can write down the distribution function of W_{ij}^F as $P(W_{ij}^F) = e^{-\frac{W_{ij}^{F2}}{2\Delta}}$ where Δ is the Gaussian squared average of W_{ij}^F. Since H is linearly related to W_{ij}^I, we get the same distribution of H as of W_{ij}^I with different Gaussian squared average Δ'. We can now solve the chemical kinetic equation:

$$\sigma_{ij}' = Ce^{-K_{ij}t} \tag{120}$$

Here K_{ij} is a random quantity defined as $1 + e^{-|H_A^{ij}|/kT}$, and $\sigma_{ij}' = \sigma_{ij} - sgn(H_A^{ij})(-1 + e^{-|H_A^{ij}|/kT})$, C is a constant. Knowing the distribution of H_A^{ij} we can easily obtain the distribution of K_{ij} by substituting H as function of K in the probability distribution:

$$P(K_{ij}) = e^{-\frac{\ln^2(K_{ij}-1)}{\beta^2\Delta'}} \tag{121}$$

We can see that K_{ij} is not Gaussian distributed but follow a log normal distribution law which is typically for a complex energy landscape. Note that K_{ij} here is the decay rate for contact dynamics. The distribution of the rates K_{ij} is basically describing the distribution of the barrier heights in the energy landscape of this complex dynamics.

7. Structure of biological membranes

In this chapter we shall discuss another biological system that is likewise very facinating and interesting to analyze by mathematical modelling tools. We assume a basic knowledge about lipids and shall then start with an introduction about the cases when lipids form aggregates.

7a. Phenomenology of bio-membranes.

Aggregates formed by small amphiphillic molecules in water display a remarkable structural richness, e.g. micellar, hexagonal and bilayer structures [??]. Furthermore monolayer structures can be formed in water-air or water-hydrocarbon interfaces. Amphiphiles constitute a very extensive class of compounds, which counts common substances like soaps, alcohols and lipids. The lipid bilayer structures are of particular interest because they play an essential role in the organization of biological cells (see figure 15). Hydrated lipid bilayer systems displays a wealth of polymorfic transitions, however, but their possible significance in biological membranes are still unrevealed. Although the discussion in this paper is restricted to lipid bilayer structures it may be applicable to a range of other amphiphillic surfactant systems.

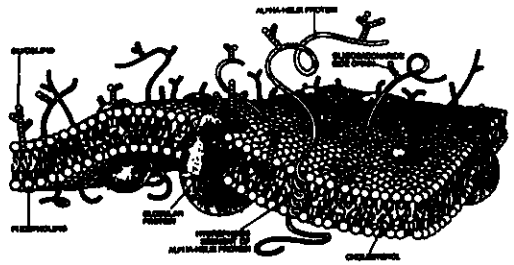


Figure 15. A popular picture of the biological membrane with proteins etc..

The experimental activity in revealing the structure of simple lipid bilayer systems is considerable. This activity is promoted by a range of interests involving medical, physical and biological sciences. The understanding of the equilibrium properties of the simple bilayers are in particular indespensible for progress in the description of structural

stability and dynamical properties of more complicated lipid bilayers like biological membranes. However, an experimental characterization of the lipid-water system in terms of equilibrium thermodynamics is in general quit difficult due to polydispersity, structural complexity and very long relaxation times towards thermal equilibrium.

Theory has been of limited help in the characterization at the large-scale structural transition properties of lipid bilayers in excess water. The stabilisation of the lipids in a bilayer structure is understood in the framework of a thermodynamic theory describing the interplay between molecular interaction free energy, molecular geometry and entropy [??]. This analysis has been suplemented by thermodynamic considerations based on the electrolyte doublelayer theory, which can give a description of the stability of simple bilayer shapes like cylinders and spheres [??].

A popular phenomenological theory for the description of shapes of individual lipid bilayers is the Canham-Helfrich model [??]. This model has even proven to be successfull in the description of shape transformations of biological membranes like erythocytes [??]. Further this model has served as the basis for recent studies on the effects of thermal ondulations on the forces between bilayers [??] and the possibility of order-disorder transitions in macroscopic conformations of membranes [??, ??]. We aim at a full statistical mechanical treatement of geometrical shapes and topologies of membranes.

8. A differential geometrical model of closed membranes

Membrane systems, that are described in differential geometrical terms by a curvature elasticity Hamiltonian (Canham-Helfrich), are analysed especially concerning their topological features and a thermodynamical theory is proposed and solved analytically. The phase behaviour is studied for closed membranes when varying the parametres $\kappa, \bar{\kappa}$ representing respectively the bending rigidity and the coefficient of the Gaussian curvature. The phase diagram displays distinct regions caracterized by many vesicles and closed membranes with many handles. The theory can give estimates of the size distribution for vesicles in terms of the model parametres, including the surface tension μ and the posibilities of an aggregation phase transition is discused.

In this chapter we demonstrate that it is possible, within a mean-field approach, to obtain important information about the phase behaviour of the theory when only closed membranes are considered. The information is concerned with the membrane stability against changes in surface connectivity. Some basic considerations on this topic has already been given in [??].

8a. Topology in membrane phenomenology.

A range of experimental techniques have been applied in the characterization of the phase behaviours and morphologies of lipid-water systems. For low water content, structures with long range order form and diffraction techniques can be applied. Early studies by X-ray diffraction techniques [??] discovered the existence of the lamellar bilayer phases and a number of bilayer phases with bicontinous structures excibiting cubic symmetries. These studies have been complimented and comfirmed with NMR [??] and freeze-fracture electron microscopy [??]. In the more diluted regimes of the phase diagram the characterization of the phases is hampered by the absence of long range order in the phases and coexistence of a large number of bilayer structures. Direct visualization by microscopic tecniques probably gives the best insight in the nature of the phases in this regime [??, ??].

A considerable effort has been directed toward characterization of the phases in terms of surface geometries [??]. The multi-lammellar and cubic phases have been characterized in terms of infinite periodic minimal surfaces (IPMS), e.g. the surfaces having zero mean curvatures and separating the ambient space into periodic subspaces. An IPMS has thus an associated point group symmetry and characteristic dimensions of its unit cell. The topology of an IPMS (see fig.1) can be very complicated, e.g. represented by the number of genus per unit cell. Properties of minimal surfaces can be derived from complex analysis through their representation by Weierstrass-polynomials. However IPMS is still not a fully determined group of surfaces, and the non-periodic minimal surfaces with non-trivial topology has only recently been explored [??]. It is thus evident that surfaces are difficult to treat in a statistical mechanical frame if the surfaces are assumed to be minimal. A second difficulty in dealing with minimal surfaces in membrane physics is that a physical principle, which dictate the crystalline properties of the IPMS, is not known. No packing condition or internal symmetri property of the constituents can guide us, as in the case of molecular crystals. With these difficulties we find that minimal surfaces at present do not provide a good starting point for the description of membranes undergoing phase transitions involving topology. We will restrict ourselves to closed membranes, which actually never can be described as minimal surfaces in $\mathbf{R}^3$.

bf 8b. Topological thermodynamics of closed membranes.

The Model

In this section the Canham-Helfrich model of membrane elasticity will be briefly discribed. This model consider only fluid membranes which at length-scales much larger than the molecular distances and the bilayer thickness can be modelled as a mathematical surface without any internal structure. The lipid bilayers excibit a number of low-temperature solid-like phases with in-plane order of the lipid molecules, but they do no demonstrate the deluge of large-scale structural transitions displayed by the fluid membranes. The model Hamiltionian takes the form

$$\mathcal{H} = \mu \int dA + \frac{\kappa}{2} \int dA (\frac{1}{r_1} + \frac{1}{r_2} - \frac{2}{r_0})^2 + \bar{\kappa} \int dA \frac{1}{r_1 r_2} \qquad (126)$$

where the integrations are performed over the surface area. r_1 and r_2 are the local principal curvatures of the surface and r_0 is the spontanous cuvature, which can arise

in bilayers with an intrinsic assymmetri between the monolayers of the bilayer. In this work we consider full symmetry between the two bilayer halves, which is the case when the bilayer is composed of a single molecular constituent. The notion of spontanous curvature will thus be omitted in the following. The mean curvature $\frac{1}{r_1}+\frac{1}{r_2}$ and the gaussian curvature $\frac{1}{r_1 r_2}$ are surface invariants, i.e. independent of the chosen parametrisation of the surface. The model Hamiltonian can be considered as a Landau theory with an expansion in symmetry invariants (reparametrization invariance in R^3) where the lowest order term is the first term in equation (1).The surface tension μ, that couples to the surface area, which due to the fixed cross-sectional areas of the lipids, must be considered as a chemical potential for the lipids in the membrane. In most thermodynamic problems involving interfaces the chemical potential control the interface.

For free surfactant interfaces μ is generally very small [??]. In a closed system μ must be considered as a Lagrange multiplier insuring a fixed overall amount of lipids in the system. Other terms may be included in Eq.[1] . If the membrane has boundaries a line tension term

$$\mu_L \int_{\text{boundary}} d\ell \tag{127}$$

must be added. However the line tension μ_L are so large that even the presence of small boundaries are suppressed for free membranes [??]. Boundaries can occur if the membrane can be attached to hydrophobic or hydrophillic elements of the experimental setup. We do not consider these cases here and just assume that the membranes are without boundaries. Furthermore anharmonic terms are neglegted in Eq.[1] . The model parameters κ and $\bar{\kappa}$ are difficult to obtain experimentally. However some concensus has been reached regarding the value of κ for artificial membranes. For dimyristoyl phosphatidyl choline bilayers, values of $\kappa \approx 1-2\cdot 10^{-13} erg$. have been obtained by pressure aspiration technicques on individual giant vesicles [??] and fourier analysis of the thermal membrane undulations [??].

8b1.The Wilmore functional

In this section we discuss some results from the mathematical literature concerning the properties of a functional which appears as the second term in Eq.(1), the Wilmore functional. The Wilmore functional is written as

$$W(\Sigma)=\frac{1}{2}\int_\Sigma H^2 dA \tag{128}$$

where $H=\frac{1}{r_1}+\frac{1}{r_2}$ is the mean curvature and dA is the area element of a surface Σ. Here Σ is *any* compact surface in $\mathbf{R}^3$ and we assume that it has no boundaries and no self-intersections. The functional W is invariant under conformal mappings of the ambient 3-space. Thus, if $\tilde{\Sigma}$ is the image of Σ under a Möbius transformation (an isometry, a scaling or an inversion in a sphere with center not in Σ), then $W(\tilde{\Sigma})=W(\Sigma)$ [??]. Recently considerable effort have been directed toward a solution of the Wilmore problem for surfaces of any genus (the infimum of W and the related variational problem). A few results relevant for our purpose will be given.

Following L. Simon [??] we write $\beta_g = inf W(\Sigma)$, where inf is taken over compact genus g sufaces without self-intersections. The following inequality is fundamental:

$$8\pi \le \beta_g < 16\pi \tag{129}$$

Equality holds on the left if and only if $g=0$ and Σ is a round sphere [??]. The right hand side inequality was observed independently by U. Pinkall and R. Kusner, see [??]. Simon then showed [??], that if we put $e_g=\beta_g-8\pi$, then

$$e_g \le \sum_{j=1}^{q} e_{\ell_j} \tag{130}$$

for any integers $q\ge 2$ and $\ell_1,...,\ell_q$ with $\sum_{j=1}^{q}\ell_j = g$. Further he proved the *existence* of W-minimizers in the following sense: For any genus g there exists a genus g surface Σ with $W(\Sigma)=\beta_g$, unless equality holds in Eq.( ) in which case there exists a sequence Σ_k of genus g surfaces and a genus g_o surface Σ_o (with $g_o \le g$) such that $W(\Sigma_k)\to W(\Sigma_o)=\beta_{g_o}$ for $k\to\infty$. Thus for a given surface the minimization of W may cause a drop in genus number. However, the fundamental conjecture [??] is that the equality actually never occurs in Eq.(5), and furthermore that for every genus there is exactly one surface which minimizes W (up to a Möbius transformations in $\mathbf{R}^3$.

8b2. Thermodynamics of surface topology

In this section we will evaluate some thermodynamic properties of lipid membranes governed by Eq. (1). The description suffers from a lack of detailed about $\mathcal{H}$. However, it turns out that the recent mathematical results (mentioned in the last chapter) provide us with sufficient information to give usefull estimates of the phase behaviour. We will consider four different cases corresponding to the introduction of more degrees of freedom. In the first case the available degrees of freedom is g (the number of handles), and in the other cases it is the number and size of vesicles (and of course g).

Going back to the original Hamiltonian Eq.(1) we have in the last chapter given bounds on the second term, the Willmore funtional. The third term is easily evaluated by using the Gauss-Bonnet theorem:

$$\int_\Sigma dA \frac{1}{r_1 r_2} = 2\pi\chi \tag{131}$$

When the surface is without boundaries the Euler characteristic χ is simply related to the genus number by $\chi=(2-2g)$.

The first term in Eq. (1) will be neglected here, since the membrane is considered as an isolated system. From the previous section it is clear that the Willmore functional restricted to compact embedded surfaces Σ without boundaries is realtered to $\mathcal{H}$: $W(\Sigma)=\frac{1}{\kappa}(\mathcal{H}-\bar{\kappa}\chi(\Sigma))$ for $\mu=0$. In particular $inf_g \mathcal{H}(\Sigma_g) = inf_g \kappa W(\Sigma_g) - 4\pi\bar{\kappa}(1-g)$, where $inf_g W$ represents the infimum of $W(\Sigma_g)$ for all boundaryless compact, embedded surfaces Σ_g with genus g.

5 pages cut out.

8b3. Discussion and Conclusion

In the previous section the biophysical analysis on the thermodynamic properties of closed membranes is related to a variational problem on Willmore surfaces. This mathematical problem is still unresolved but the results obtained so far gives sufficient information to provide valuable results on the phase behaviour of membranes. For the three simple cases considered, we can summarize our results in the following way:

In *case 1* a single membrane was considered. When $\beta\kappa$ is large a saddle point evaluation is appropriate. The only remaining degree of freedom of interest is the genus number g of the surface. Detailed information about the saddlepoint is not available but the first approximations to the partition function can be given in terms of limits of the minimum value of the Willmore functional for each g, parametrized by $c = 1$ and $c = 2$. For $c = 1$ the system displays an abrupt, continous change in $< g >$ at $\bar{\kappa} = 0$. Althoug it is accompanied by strong fluctuations in $< g >$, the transition is neighter 1. order or 2. order, but rather ∞ order in the sense that $\frac{\partial^n F}{\partial(\beta\bar{\kappa})} \propto (G_A)^n \to \infty$ for $n \to \infty$ at $\bar{\kappa} = 0$. In an ensemle of weakly interacting membranes this may be changed to a 1. order or a 2. order transition.
For $c = 2$ the transition takes place at $\bar{\kappa} = 2\kappa/G_A$, where G_A represents a cut-off in the number of genus for the surface. Our proseedure thus provide us with detailed information about the thermodynamics of the membrane except in the narrow range of $\bar{\kappa}$-values from 0 to $2\kappa/G_A$, where a transformation from $g = 0$ to $g = G_A$ takes place.

In *case 2* a system of membranes which have not translational degrees of freedom aviable are considered, and the two phases appearing in *case 1* are again present: Phase 1 characterized by membranes with a large number of handles and Phase 2 consisting of simple vesicles ($< g >= 0$). The transition between Phase 1 and Phase 2 is accompanied by an enhancenment of the average size of the aggregates. Phase 1 and Phase 2 are limited by a region in parameter space where this analysis is insufficient in describing the lipid system but a transformation to other lipid structures is expected.

In *case 3* the membranes can move around in space. In this case the description is sensitive in the whole parameter space, i.e. all regions of the parameter space can be structurally determined when the chemical potential μ is kept fixed. There appear phases of vesicles with high genus number for large positive values of $\beta\bar{\kappa}$ or phases of larger or smaller vesicles with no genus number in the region of negative $\beta\bar{\kappa}$. There is a sharp change from the phase consisting of vesicles with no genus and vesicles with a high genus number and a large surface area. In general, if μ is kept fixed, the vesicle size is more or less constant, but if μ is varying and instead the total number of lipids is held fixed, the size of vesicles can vary. The mean number of vesicles will increase when going from one of these phases to the other. The size distribution of the vesicles can be determined for various values of β,$\bar{\kappa}$ when μ is vatying and the number of lipids is fixed.The vesicle sizes are distributed around a peak determined by the chemical potential. An equation of state between the vesicle size and number and the amount of lipid can be derived.

An important result of this study has been the prediction of a topological phase where structures with high genus (many handles) are formed. As we saw, this was due to the fact that the energy would be lowered (at least $2\pi\bar{\kappa}$) by for mig handles. Furthermore the energy distribution (from the Willmore functional) to the formation of a sphere is at least (when $g = 0$) $4\pi\bar{\kappa}$. A possible scenario for the formation of a handle is first an aggregation of a number of vesicles. If the right phase conditions are present (e.g. $\bar{\kappa} > 0$) vesicles fuse together (perhaps through an inverse hexagonal phase [??]) forming a small canal where they face each other. Such an extended structure can attach to a bigger vesicle forming a handle or bending back into a ring structure, a torus, in either case increasing the genus number. Such a picture is just an attempt at visualising the topological process of forming high genus structures.

The principles behind the interplay between three dimensional structures and the structural transitions of proteins and their biological function are to large extend understood. A similar relationship for the biological membranes is still considered at a hypothetical level [??]. Whether the extended lipid polymorfism has any significance in biological system is still unclear.

To conclude, it is well-known that biological membranes provide a large variety of mathematical forms that are realized in aqueous surroundings inside or around the living cell. Some of these forms represent highly non-trivial topology seen e.g. in the intracellular Golgi apparatus, where a large number of handles and tubuli are connecting different compartments between lipid layers. The purpose of this topological structure is a need for filtration of proteins in the cellular liquid. Such topological structures have been verified in various observations [??], [??] (figure 16a) . It has therefore been tempting to see if they among others can be explained from the more general geometrical description presented in this paper. This topological complexity can largely be explained by the phase structures described in the last chapter in the case of $\bar{\kappa} > 0$ and pictured in the phase diagrams of Fig. 2. Here is, under certain conditions, seen a large production of handles and tubuli fixed to the membrane structure and only limitted by material constraints such as a finite lipid size.

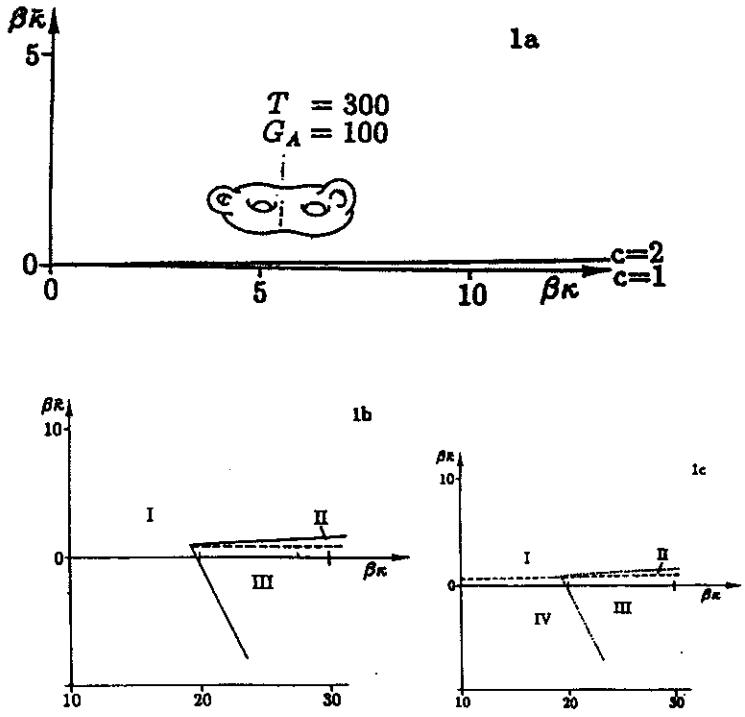


Figure 16a,b,c. Figure 16 shows the phase diagrams in the three cases mentioned in the text.
The phase diagram contains many topological phases as functions of κ and $\bar{\kappa}$.

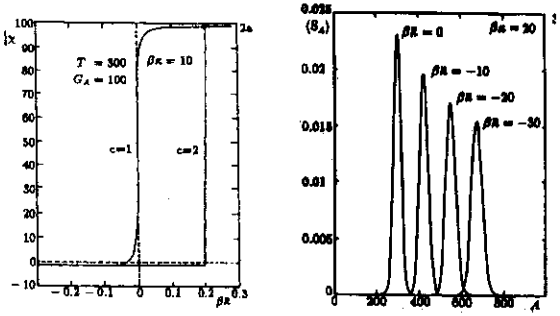


Figure 17a,b.: Figure 17a shows what happens when transversing through genus=0 in the phase diagram.
17b: shows the vesicle distribution at different values of κ.

In general the lipid system displays a number of different statistical mechanical in-plane phases, such as a high-temperature fluid phase, a low-temperature solid phase and perhaps hexatic phases. The last mentioned are characterized by long-range orientational order and short-range positional order. It is also known from experiments that membranes can form a variety of different large-scale structures, a property which is extensively exploited in biological membrane systems. For low water content structures such as lamellar and inverse hexagonal phases are very common and the transition between these phases has, e.g., been described in reference [??] .

REFERENCES

1. Luzzati V., Tardieu, A. (1974) Annu. Rev. Phys. Chem. 25. 79-92
2. Israelachvili, D., Mitchell, J.,Ninham, (1977) Biochim. Biophys. Acta 470, 185-201.
3. Jönsson B.,Wennerström H. (1981) J. of Colloid and Interface Science, 80 482-496.
4. Canham, P.B. (1970) J. Theoret. Biol. 26, 61-81; Helfrich, W. (1973) Z. Naturforsch. 28c, 693-703.
5. Deurling, H. J., Helfrich W. (1980) Biophys. J. 13, 941-?
6. Helfrich W. (1978) Z. Naturforsch. 33a, 305-? and Lipowsky R., Leibler S. (1986) Phys. Rev. Lett. 56, 2541
7. Helfrich W. (1985) J. Physique 46, 1263
8. Peliti L., Leibler S. (1985) Phys. Rev. Lett. 56, 1690.
9. Kantor Y., Nelson D. R. (1987) Phys. Rev. A 36, 4020.
10. Helfrich W., Harbich W. in *Physics of Amphiphilic layers* (Meunier J., Langevin D., Boccara N., Eds) pp 58, Springer, Berlin, 1987.
11. Lindblom G., Wennerström H. (1977) Biophys. Chem. 6, 167-171.
12. Gulik-Krzywicki T., Aggerbeck L. P., Larsson K. in *Surfactants in Solutions* (Mittel K. L., Lindman B., Eds.) Vol. 1 pp 237-257, Plenum, New York 1984.
13. Miller D. D., Bellore J. R., Evans D. F., Talmon Y., Ninham B. W. (1987) J. Phys. Chem. 91, 674-685.

14. Servuss R. M. (1989) Chemistry and Physics of Lipids 50, 87-97.

15. Costa C. J. (1982), see Ossemann R. *A survey of Minimal Surfaces* Dover Publications Inc., New York, 1985.

16. Brochard F., De Gennes P. G., Pfeuty P. (1976) J. Physique, 37, 1099.

17. Lorenzen S., Servuss R. M., Helfrich W. (1086) Biophysics J., 50, 565-572.

18. Bo L., Waugh R. E. (1989) Biophysics J., 55, 509-517.

19. Engelhardt E., Duwe H. P., Sackmann E. (1985) J. Physique Lett. 46, 395-400.

20. Anderson S., Hyde S.T., Larsson K., Lidin S. (1988) Chem. Rev.,88, 221-242.

21. Weiner J. (1978) Indiana Univ. Math. J. 27, 19-35

22. Simon L. (1986) Proc. Cont. Math. Anal. Natl. Univ. 10, 187-216.

23. Kusner R. (1989) Pacific J. Math. 138, 317-345.

24. Li, Yau (1987) Acta Mathematica 156, 192.

25. Willmore T.J., Total Curvature in Riemanian Geometry, Chichester, Wiley (1982).

26. Cullis P.R., Hope M.J., de Kruijff B.,Verkleij A.J., Tilcock C.P.S., "Phospholipids and Cellular Regulation" (ed.: J.F.Kou) CRC Press, Boca Raton, Florida (1985) Vol. 1.

27. Harbich W., Servuss R. M., Helfrich W. (1978) Z. Naturforsch. 33a, 1013-1017.

28. Kirk G. L., Gruner S. M., Stein D. L. (1984) Biochemistry, 23, 1093-1102.

9. Future outlook

We started introducing proteins and then modelled and analyzed them. Then we went to membranes and modelled those too. A nice end on these lecture notes would be to combine the proteins with membranes. Unfortunately there is not so much known in details about membrane bound proteins. The scope of these lecture is the study of protein structures in details in a distance geometry approach so perhaps we should leave the membrane proteins for another time. However, if we were to find a relevant subject as the basis for an outlook into the next century I cannot think about a better subject than protein-protein interactions. In the last part of this century we have concentrated immensely on single proteins and their folding (without super success) but the most important biological processes have to do with docking etc. of several proteins. Good luck!