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

H4.SMR/916 - 32

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

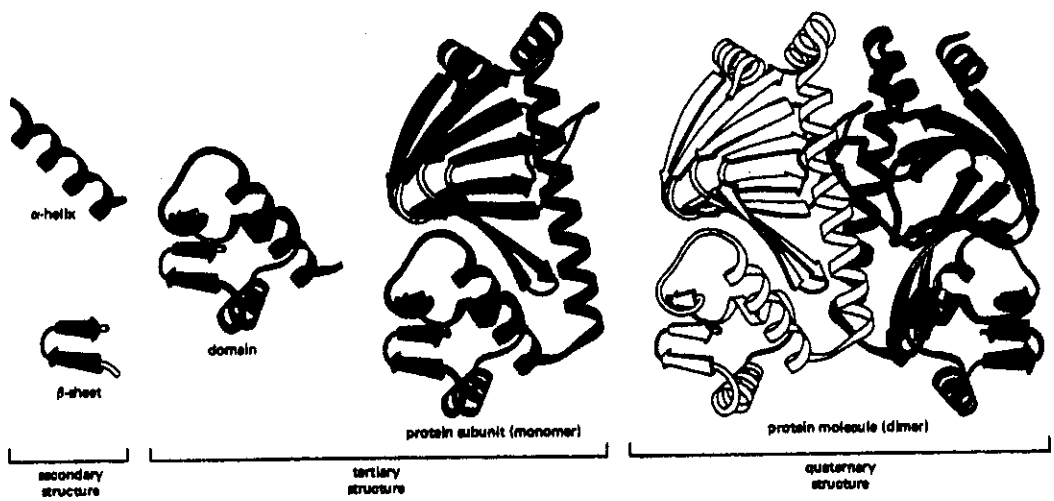
Geometry and Topology of Biopolymers

Henrik. G. BOHR
CBS, Dept. of Physical Chemistry
DTU The Technical University of Denmark
Lyngby, Denmark

GEOMETRY AND TOPOLOGY OF BIOPOLYMERS

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 are for the college in biophysics in March 1996 at the International Center for Theoretical Physics, ICTP, in Trieste, Italy. The lectures are made for an audience interested in working on mathematical models in molecular biology from a theoretical physics perspective.



CONTENT

4b2. Graphical representation of the protein folds.

4b3. Statistics of secondary structure abundance in nature.

4d. Magic numbers.

4d1 Magic numbers and the Euler characteristics.

4e. The Molten Globule and the Parent states.

4f. Conclusion.

5. Implication of topology for protein structure and folding

5a. Introduction.

5b. Winding.

5c. Wringons.

5d. Resonator driven transition.

5e. Linking and writhing.

5f. Wring modes in open chain molecules.

5g. Hydrolysis of proteins and DNA breaking.

5h. Length distributions of prokaryotic and eukaryotic proteins.

6. Aggregation of bio-polymers

6a. Protein aggregates studied by Atomic force and electron microscopy.

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: Magic numbers in protein structures, by Per-Anker Lindgaard and H. Bohr (submitted to Phys. Rev. Letters (1996). Chapter 5 is partly from the article: Resonator driven Protein Folding, by Jakob Bohr, H. Bohr and S. Brunak, in "Protein Folds", CRC Press 1995, Florida, USA. Chapter 6 is a study of microscopy done in collaboration with Anders Kühle and Jakob Bohr. 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). Some of 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.

Connect d 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.

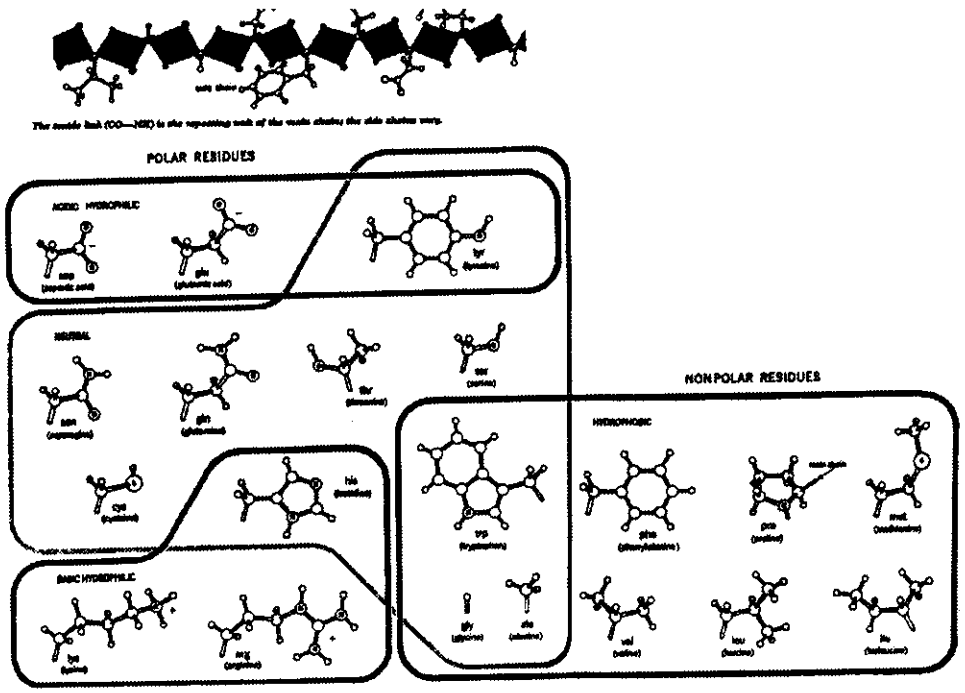


Figure 1 a,b. a: A picture of the unfolded peptide chain. b: All the 20 amino acids grouped.

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.

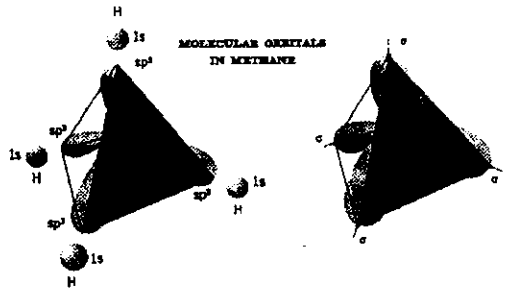


Figure 2a.: The molecular orbitals of Methane forming a Tetrahedron when hybridized.

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.

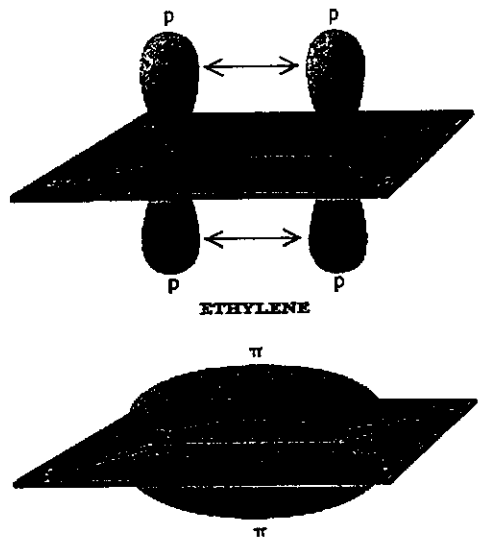


Figure 2b.: The σ and π orbitals of Ethylene.

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 $32kcal/mol$.

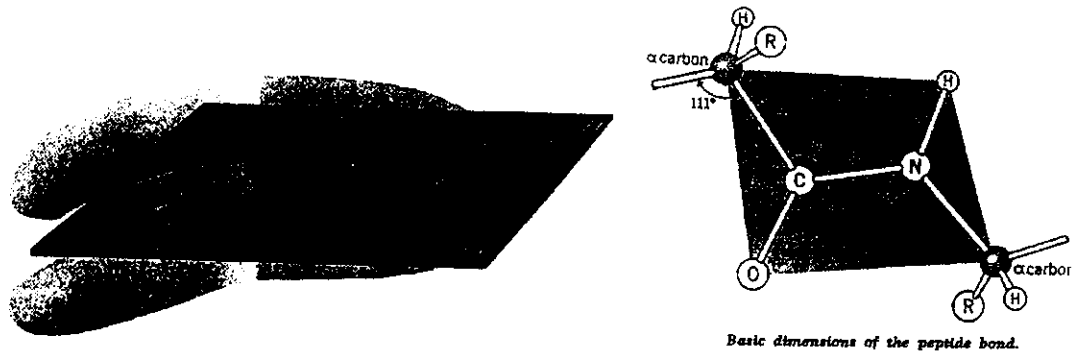


Figure 2c,d c: π bond across the peptide plane. d: The atoms in the peptide plane.

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.

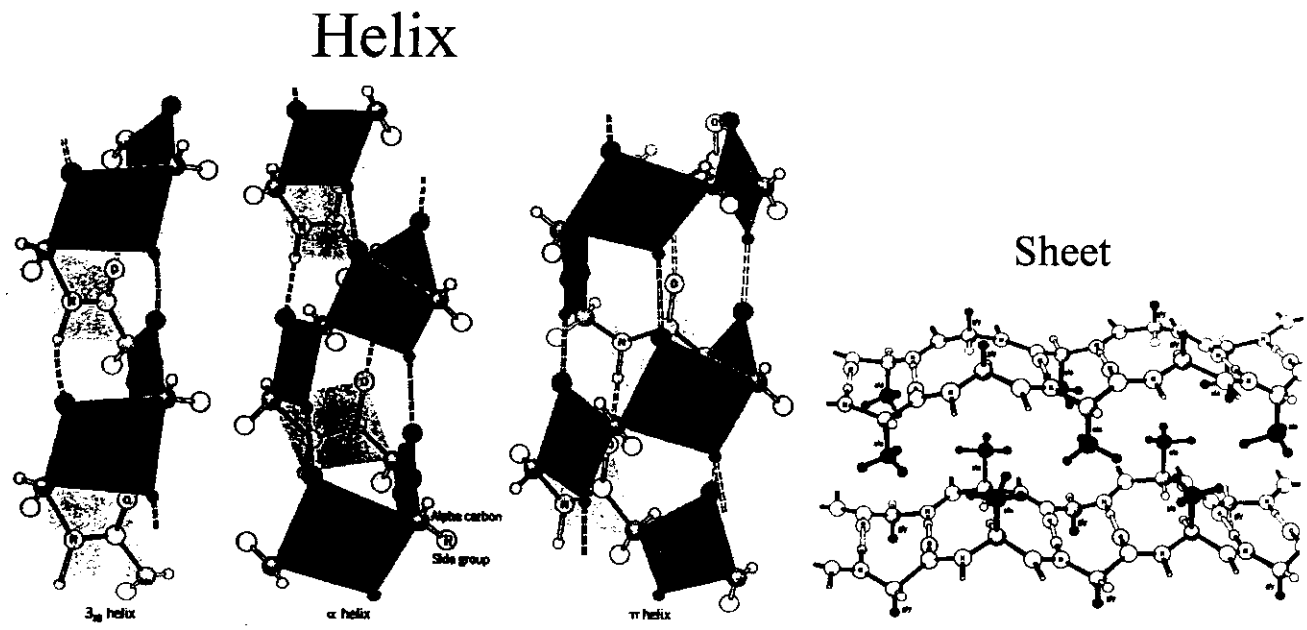


Figure 3a,b. a: 3 different helices, $3_{10}, \alpha$ and π helix.
b: Beta-sheet in Silk.

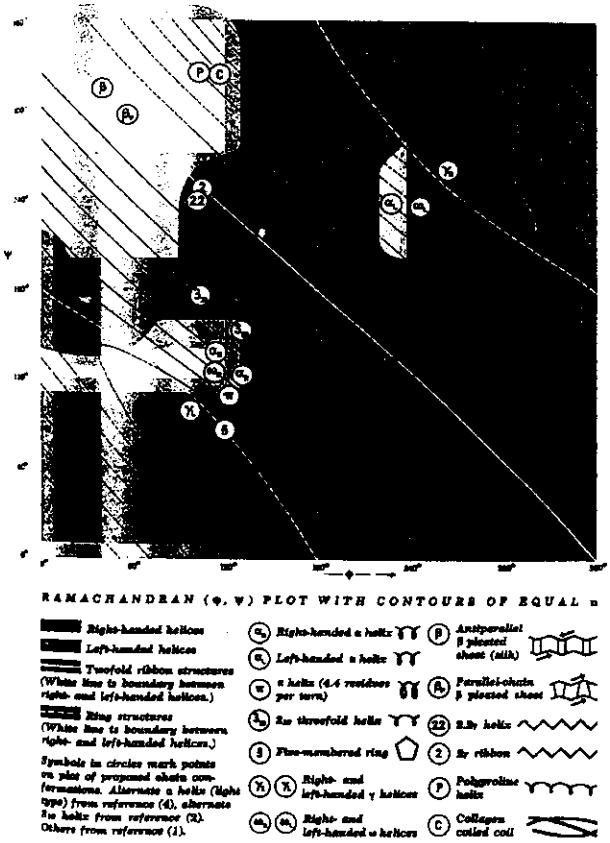


Figure 3c. c: The Ramachandran plot of substructures in proteins.

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.

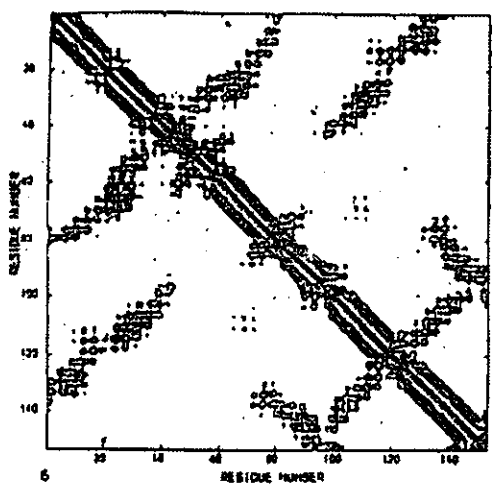


Figure 4: 2-dimensional plot of binary distance matrix of Rubredoxin, Threshold=8 Å.

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

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.

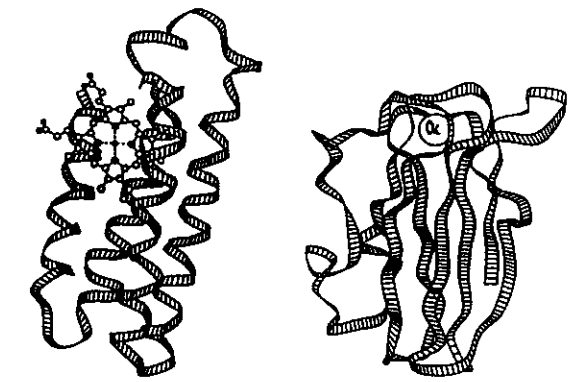


Figure 4a,b: The ribon representation of typical members from the super fold class α (left) and β (right).

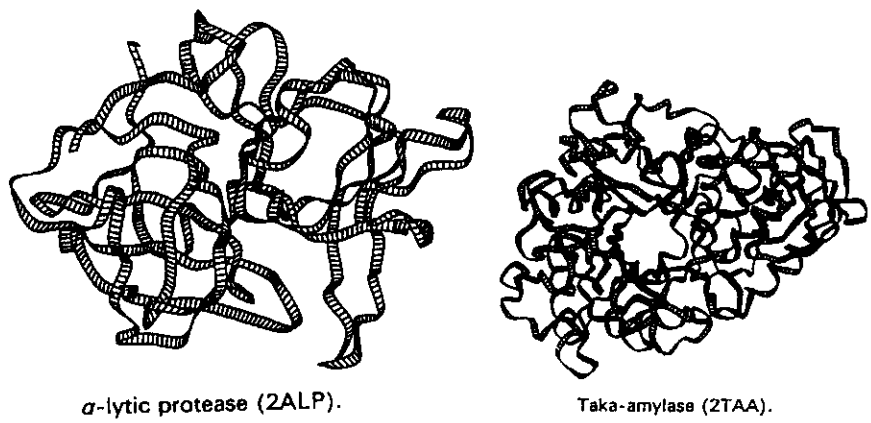


Figure 4c,d: The ribon representation of typical members from the super fold class $\alpha + \beta$ (left) and $\alpha \cdot \beta$ (right).

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.

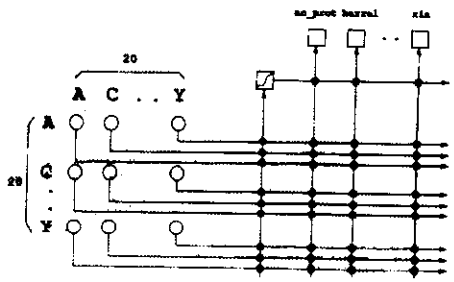


Figure 5.: A picture of the Cascade Correlation Network.

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 $\mathring{A}$ 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.

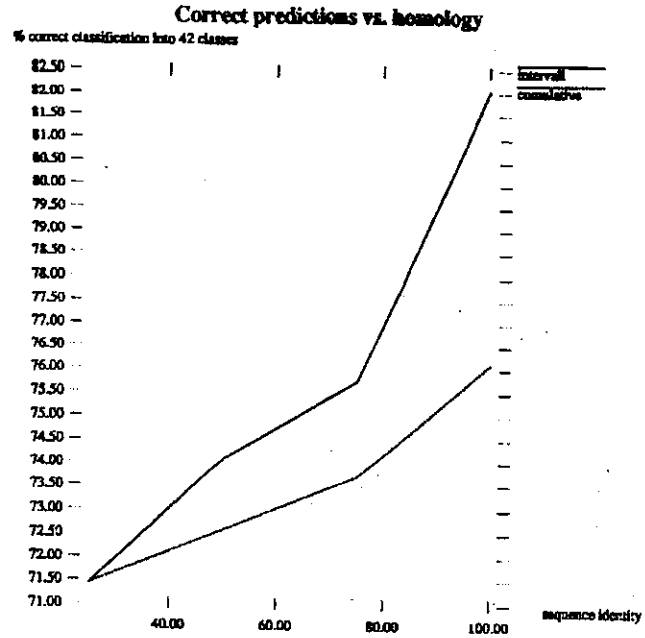


Figure 6.: This figure shows the correctness in the prediction versus homology measured in intervals (upper curve) or accumulatively (lower curve).

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. Statistical mechanics of protein fold classification

In this section we shall present a model or an energy prescription for classifying proteins into structural groups - the fold classes. The structural fold classes for proteins are being defined on a three dimensional lattice and a model Hamiltonian is suggested that can explain the division into such fold classes during the later stages of the folding processes. Proteins are described as chains of secondary structure elements with hinges in between being the important degree of freedom. In such a chain representation protein structures can be represented uniquely by a 1-dimensional string of physical coupling constants describing scalar and vectorial spin interactions. An automated procedure is constructed in which any 3-dimensional protein structure in the usual *PDB* coordinate format can be transformed into this papers chain representation. ¿From more general statistical mechanics arguments one can estimate the upper limit of the total number of possible fold classes to be around 4000. This number is confirmed through an explicit calculation of the possible chain configurations that are tightly packed on a small 3-dim. regular lattice. Taking into account hydrophobic forces we have found a mechanism for formation of domains containing magic numbers of secondary structures and multiple of these domains. We have performed a statistical analysis of available protein structures and found agreement with the predicted preferred abundances of proteins with a magic number of secondary structures.

1 4a. Introduction

In almost half a century databases have been build up of, up to now, over a thousand protein structures and, like the case of for example botany or isotope tables, it is natural to ask for some classification that can group the entities into structurally related families other than just what analysis of the corresponding sequence of amino acids can tell. What we here have in mind is a kind of atomistic grouping where the elements are grouped according to the number and type of their entities.

In the case of the nuclear isotopes the grouping in closed shells of nucleons came historically rather late since it was not obvious that an independent particle description would make sense in the nuclear interaction picture; - and yet magic numbers came out of a fairly simple single particle force potential in agreement with empirical data. Likewise for our case we shall show that magic numbers for the stability in the packing of protein structure elements are revealed in a calculation based on a simple hydrophobic force field. Literally proteins appear to be packed like closed shells of secondary structure elements all connected and making out the polypeptide chain that is wound up to a compact body with subdomain structures corresponding to that of a vegetable. Before showing that let us first review the usual story of protein structure and dynamics.

A characteristic feature of proteins is that their observed structures are densely folded in a complex manner of secondary structures and intervening irregular structures [?]. Pauling [?] was the first to describe the two dominant kinds of building blocks, the α-helices and the β-sheets. Later others have been proposed (i.e. inverse turns and Ω-loops

[?]).

In aqueous solutions proteins form dense globule, which neither dissolve nor phase separate, as emphasized by Dill [?], who derived a thermodynamic theory for these. A main reason for this is the action of the *hydro-phobic/phillic* force, which is an unspecific interface-tension-like force [?]. Yet, a protein with a specific amino acid chain folds, paradoxically [?] in a matter of seconds, to a particular fold, according to an information which must be provided via the underlying linear sequence information. A concise review is given by Wolynes [?], in which the folding problem has been related to the spin glass problem, marginal stability and minimal frustration. Another problem to understand is why proteins seem to have predominant lengths of the chains [?] and separate into subunits, secondary structures [?], domains and finally the functional tertiary or quaternary structures. Berman *et al* [?] have made a statistical study of known proteins and found that the distribution has characteristic peaks near multipla of chain lengths of 125 amino acids. The total length may go up to a few thousands. Only a few hundred structures have so far been determined (in crystalline form). Yet, many thousands of proteins have had their sequence determined. It is of great interest to attempt to classify the possible structures which can exist.

The purpose of this work is to propose a schematic framework for the description of the folding of secondary structures into domains of proteins. A domain is defined as a typical folding motif, which is re-found in many more completely folded proteins in their final, tertiary state. A domain consist of typically up to 125 nuclear acids, which is not surprising in the light of the statistics[?] mentioned above. The domains form usually about ten secondary structures and interconnecting loops. The aim is not that of predicting the detailed structures but to describe general classes of typical possible folds.

Let us start by considering the simple crystalline classes of structures. Group theory tells us that there are only 230 different classes in three dimensions. Many materials assume before they melt, in spite of the possible diversity, a simple open structure, the body centered cubic structure *bcc*, which is stabilized by entropy [?]. At lower temperature the structure transforms by a so called Martensitic transformation to more closed packed structures with generally 'triangular' coordination between the constituents. There can be several such possibilities *hcp*, *fcc*, *dhcp* 9-R, 18-R, $\cdots$, however all resulting from a single 'parent' *bcc* phase [?]. The observed, irregular protein structures may correspond to the such complicated ground state configurations which are the results of the competition between all relevant forces.

A major problem in the protein fold problem is how to understand that the proteins can find their fold without trying all the statistically possible options. We shall here assume that such information is coded linearly in terms of the amino acid sequence, giving rise to a natural tendency for the backbone to fold correctly just from the information about the short range forces along the backbone. The difficult forces between far apart sections of the proteins as well as the hydrophobic and hydrophilic requirements come in at a later stage, providing the final optimization - and the observed complex irregular and twisted patterns. However, it is not possible for such unspecific hydrophobic forces to define a specific fold when *the system is in an unfolded state.* These forces only act as a general condensing force. At the final level however, these forces finally minimize the energy for the compact folds. It is possible that proteins, in the course of evolution, has selected amino-acid sequences for which short range bending forces are in agreement with the compactness requirement. This would help understanding why natural proteins

can find their optimum (native) fold exceedingly fast (in a matter of seconds), whereas if all possibilities had to be tested an astronomic time would elapse, *i.e.* the Levinthal paradox [?].

Recently Chothia[?] 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 as the only effect of the long range forces. A fold means[?, ?, ?, ?, ?, ?, ?], as mentioned above a folded motive with a particular structural topology that a protein-domain, can assume in its native state. The new paradigm is to classify proteins or, more precisely, protein domains 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[?] in the 18th century succeeded in the field of botany to identify the important classification parameters. He thereby solved the difficult *homology* problem defining when plants are *the same* without being *identical* - and when they belong to the same class or not. It gives a systematic, although not natural classification. Here we suggest, that 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. By devising a local projection scheme for systematizing the protein fold on a lattice we have proposed an effective cut through the homology problem. It is well known that global measures for 'similar' folds using the root-mean-square measure (RMS) for the coordinates of the backbones is vastly misleading[?]. If just one secondary structure is slightly rotated the RMS can become very large; this is not expedient. In traditional classification in physics, as in the periodic table or in the crystal groups, a certain capaciousness in the homology concept is neither needed nor warranted. In the protein folding case, as in botany, it is. Yet the final classification criteria must be unique.

Similar simplifications with idealized elements have previously been proposed by Murzin and Finkelstein [?] for describing the domains of α-helices. They considered the α-helices as cylinders and considered a close packing of these as edges of polyhedra with triangular faces. They demonstrated a high degree of coordination of the possible and the observed structures, except for bundles of larger numbers of long helices, which seems to align more parallel. It is interesting to note that their structures in all cases can be regarded as twisted structures of a simple parallel bundle. Their work describes the number of distinct twists. In the above crystal analogy, they classify the various possible closed packed structures belonging to a simple 'cubic' parent phase. The polyhedron method has the draw back that is does not work for β-sheets. However the 'cubic' representation describes equally well the β-sheets and the β-sandwiches, as for example schematized in a different representation by Finkelstein and Reva [?]. Recently, compact lattice models for late stages of protein folding have been intensively studied [?, ?, ?]. There secondary structures is modelled schematically as sequences of monomers with a persistence length

of more than two. These are supposed to be formed at the compact folding stage in a search for the minimum strong inter-chain interactions, or for a state of 'minimum frustration' as discussed by Wolynes [?]. This approach is very different from the present.

2 4b. A model Hamiltonian system for protein folding

Let us introduce a highly simplified model in order to establish a necessary frame work in which we can systematically name the structural classes of the native folds, and in addition calculate the energy cost for structures not belonging to these classes.

We shall first discuss some well-known examples from theoretical physics. The important issue is if we can understand how weak, global forces can be important compared to stronger forces. There are several examples in physics, where it is the weak forces which determines the gross structure and the strong forces which determine the details. A well know example gravity. Another, which in fact will be close to the present approach, is the Heisenberg ferromagnet. It represents a system described by three dimensional spin vectors $\mathbf{S} = (S_x, S_y, S_z)$. The spins interact with an isotropic interaction,*i.e.* it is invariant under 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 ground state, *i.e.* the lowest energy state has all spins parallel. However, it cannot determine the direction in which they point; and the ground state is infinitely rotationally degenerate. The rotational symmetry can be broken, however, 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.

2.1 4b1. Defining the structural elements

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 or ribbon, *i.e.* almost taking literally the familiar ribbon representation of proteins. Of course it is well known that the (not depicted) side chains fill out a considerable amount of space and contribute to both the rigidity of the backbone and the inter chain forces. Thus we assume that the tape, although being flexible, can transform information about angular twists along itself to some distance. The protein is characterized by a linear information in terms of the twenty different the amino acids. This given 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; 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. However, initially we switch off these forces (the detailed mechanism by which this can be done physically is not important; a possible way could be that water is playing an important shielding role before the very dense, water-free native state is reached). Imagine we start the protein in a fully extended state in aqueous solution. 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 native close folding which we know is the 'ground state'. The extended state therefore corresponds to a high temperature configuration of a statistical system. We shall use the notion temperature about a certain dynamic state of the system. However we are strictly speaking dealing with a non-equilibrium system and the use of concepts from equilibrium statistical mechanics might be misleading. Thus an actual final state depends on the intermediate reaction rates rather than the energy of the final product, which may well be higher than the optimum state. 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, which may be of importance for the later folding. Experimentally the α-helix is usually seen to occur in the early stages of the folding process, although not perfectly (circular dichronism). By the helix formation hydrogen bonds between every third amino acid along the helix are satisfied. This is clearly a strong driving force, making it plausible that this process is one of the first to occur. The helices are typically between 6-10 amino acids long. It is not known why they have this length In respect to the analogy to the Heisenberg model eq.(4) it corresponds to the formation of small ferromagnetic domains. At the considered temperature (*i.e.* at the molten Globule state) let us assume that the extended tape is sub-structured according to the underlying amino-acid letter code into two groups of secondary structures, as used in the well known tape representations of proteins (ref to book with the pictures). One set, which we denote by large letters $A, B, C, ...$, is representing the described α-helix and also potential pieces for the formation of 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. Yet, the β-sheets need to be folded into the correct relative position in space. All these elements are assumed to be approximately linear with a well defined start and end point (amino acid). The second group consists of the remaining connecting pieces of the tape, the irregular strands and the turns. These can be replaced by the straight connection line, $a, b, c, \cdots$ between two consecutive secondary elements of the first group. Then all elements can be considered straight. Two elements are connected by a 'hinge' which is characterized by a direction in space, perpendicular to the plane in which the two joining elements can rotate. The position and action of the hinge is determined by the underlying amino acid sequence (or their side chains) - and it could also be determined by the action of chaperones. 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, simplifying 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 could include such interactions between the elements of the first group $A, B, C, ...$, in particular the potential β-sheet elements,which come into operation when the water is squeezed out of the final dense fold.

2.2 4b2. A realistic fold Hamiltonian

To introduce a realistic Hamiltonian function for protein structure formation we will start by looking at the final, known structures. These consist of of secondary structures [?] of principally α-helices (spiral, stiff subunits of on the average ten amino acids) and β-sheets which are semi planar collections of a number of almost straight chain elements (β-strands) of around ten amino acids. Between these elements the protein is connected with more irregular loops. The structures are twisted and deformed in a characteristic biological way. The first problem which arises is the homology problem: how to define when two protein structures are similar, *i.e.* belonging to the same structural class or not. A strict identity measure as for example that of a minimal root mean square sum for the backbone coordinates, is clearly too strict - and even misleading since similar - but differently twisted - structures might be judged as unrelated. In crystal structures it is known that most materials - and in particular shape-memory-alloys - assume a high symmetry, simple and open cubic (b.c.c.) phase at high temperatures, just below the melting temperature. This is called the *Parent* phase. At lower temperature, at the *Martensitic* transition, they condense into more complicated structures. We wish to describe the protein folds on a similar high symmetry level. To solve the homology problem, we consider the secondary structures as straight sticks and replace the loops by the interconnection lines between the endpoints of the secondary structures, which are defined by the sequence information.

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 $\hat{\mathbf{e}}^P$ along one of the axis in a Cartesian coordinate system. Similarly each element $p = a, b, c, ...$ is characterized by $\hat{\mathbf{e}}^p$. 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

a and A with a hinge spin $\mathcal{S}_1$ is given by

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

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° (we neglect 180°) 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 $\hat{\mathbf{e}}^P$ and $\hat{\mathbf{e}}^p$. A given sequence of spins $\mathcal{S}_i$ and start direction $\hat{\mathbf{e}}^a$ is a rigid building prescription, by which any later element direction $\hat{\mathbf{e}}^i$ 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, e.g.:

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

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 eq. (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 i 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 stearic hindrance in doing so (this could in fact also be introduced in the model). Expressed in an other way we do not care how the system has arrived at any state which we can measure the energy for, which is reasonable when discussing the ground state. In order to be able to describe the energy cost for violating the optimum fold we write the argument as a 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- dibic direction. 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 z^{2N-1} possibilities for a chair of $2N + 1$ elements, with $z = 4$ in the described case. 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. Thus one could characterized 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. 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): Suppose the above mentioned elements are viewed in the untwisted 'cubic' state, and that they are essentially linear of various lengths, the direction is given by a unit vector $\hat{e}$. By projecting the system onto a simple cubic lattice a dense packing can be defined as one in which all vertices have the maximum number of nearest neighbours. This measure has been used earlier by Thriumalai[?]. One could also define another measure, which is more appropriate in our case, were we by the elements model the secondary structures. Thus a packing which is optimal for the hydrophobic forces is one in which the secondary structures, (which usually have a hydrophobic side), can be packed as closely as possible. Accordingly we need a subset of the above in which the secondary elements (characterized by the interaction constants $\pm J$ and $\pm K$) have as many parallel neighbours elements as possible. By Monte Carlo computer simulations the model is able 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.

For the classification to be useful for actual proteins it is important to have a unique and easy identification of the class to which any given protein belongs. Since the observed low-temperature are usually strongly twisted, we wish to devise a local identification method as follows. Find the unit vectors along the elements, for the turns this represents the interaction line between two connected type 1 elements. For any three consecutive unit vectors $\hat{e}_o, \hat{e}_1, \hat{e}_2$ it is the condition $\hat{e}_o \cdot \hat{e}_2 > 1/\sqrt{2}$ defines the interaction constant $\bar{j}$. Similar for others..... This provides the basis for the classification of sequences into fold-classes. As a simple example we show on Fig. 1. the projection of the 4-α-helix bundle, which is given by the descriptor $jKj\bar{K}j$, where $\bar{K} = -K$. The descriptor depends on the direction in which the chain is traversed, but it is invariant under rotation. We find that the cytochrome family belongs to the mentioned $jKj\bar{K}j$-class, as well as for example Haemarythrin Tulysozym and Rabbit uteroglobin. Some proteins are 'embellished' by the

By this definintion of a distance between fold classes the *4-helix bundle* class and the *β-sandwich* Plasto-cyanine class will have a certain overlap (due to the fact that helices and strands are counted the same) and therefore a small distance between them while the *4-helix bundle* and the *TIM-barrel* will have a large distance between them due to the very great differences in size and geometry.

One could invent another measure that would distinguish the fold classes by their content of specific types of secondary structures. In such a measure the most distant classes are the ones containing proteins purely build up of either α-helices or β-strands. However, we are here more interested in quantifying geometrical and topological (or morphogenetical) aspects of the structures of proteins more than their content and for that purpose the measure based on similarities of the mentioned "names" is more suitable.

3 4c. Numerical calculation of the fold classes

The purpose of the numerical calculation is to find precisely how many tightly packed configurations of a given chain, that can exist on the 3-dimensional regular lattice. From this number we estimate the number of specific folds and the total number of possible fold classes beside gaining statistical knowledge of configurations for a particular number of links and lattice sizes. The latter turns out not to be crucial since the statistics of the configurations quickly converge to definite value for larger lattices.

In the description above we constructed a chain model that could describe folding patterns of proteins that were configured on a 3 dimensional regular lattice. this chain model is constraint to self avoidance so that the lattice links are only occupied by one structural elements or an interconnecting strand between two elements. For mapping out the ground state, it is most straightforward to operate directly on the element direction vectors $\hat{e}_P$ and $\hat{e}_p$. We start by placing two perpendicular elements (a, A) and their spin: $\hat{e}_1^a \mathcal{S}_1 \hat{e}_2^A$. The next element direction vector $\hat{e}_3^b$ is then placed in any of the four possible direction according to the values of the interaction constant for element A: $\pm J$, corresponding to a link parallel and anti-parallel to $\hat{e}_1^a$; and $\pm K$ corresponding to a link perpendicular or anti-perpendicular to $\hat{e}_1^a$. This determines the direction of $\mathcal{S}_2$, which is not essential for the ground state calculation, since all spins follow the direction dictated by the interaction constants. However, they are important for r the excited states since they, describe the excursions from the optimum folds. The process is then continued, under the constrain that the path is self avoiding. To find the dense folds we consider all configurations in simple confinements, such as those in a $2 \times m \times n$-box. This gives still a large number of folds, as can be seen from table 1 in the case of a $2 \times 1 \times 1$.

In order to find the number of unique folds, which are not symmetry related - f.ex. by a simple rotation, it is imperative to be able to select only those with different names. Next we find among those all which are closest packed in the sense of having the largest number of neighbours. This is plotted as the heavy full line. We notice it is very irregular with dips at a number of 'magic' numbers. A simple analysis shows that these corresponds to (in sequence) filling a $1 \times 1 \times 1$-box at the number of element $N = 7$, a

$2 \times 1 \times 1$-box at $N = 11$, a $2 \times 2 \times 1$-box at $N = 17$ and a $3\times$ packed $2 \times 1 \times 1$-boxes at $N = 23$. It is not possible with $z = 4$ to completely fill a $2 \times 2 \times 2$-box. This can be done if we allow also straight continuation of the elements, *i.e.* using $z = 5$. The numbers are in this case much larger. The dashed line on both plots indicates the mean field estimate $(Z/e)^N$ (ref). Although it represents the data reasonably well there seems to be systematic deviations for large N. The number of closed packed structures seems to be systematically over estimated by the mean field theory. A mush better agreement was found by Thriummalay for the 2-d case. We have given the complete statistics for all lattice sizes up to $3 \times 2 \times 2$.

In the context of protein folding we find (arguments) that the simple case with $z = 4$ is most relevant. It is then interesting to note that for $N \leq 18$ the number of possible fold classes is relatively small. The accumulated number is 3244 if we include only those which can pack in a $2 \times 2 \times 2$-box. If we include also those which can pack in the more elongated structures, *i.e.* including a $3 \times 1 \times 1$-box we get 4066. However by increasing the number of elements in a domain by just a few one increases the number of possibilities dramatically by about 4000 for each added element. We believe that there is a connection between the simple geometrical magic numbers found in the close packings and the magic numbers in protein domains, where there is a pronounced maximum at about 125 amino acid, corresponding to typically 10 secondary structures, or 18 of our elements.

There is experimental support for this observation which can be seen by studying the statistics of the length distribution of protein chains[?] in the databases. Those distributions clearly show optima in protein length around 125 amino acids, 250 a.a. etc. for eukaryote and similarly 150 a.a. and 300 a.a. for prokaryote. The origin of this remarkable periodicity has yet to be explained in details but it can have something to do with the topology of the polypeptide chain in early stages of protein folding[?] or the phenomena behind could be a remnant of the DNA/RNA structures). Here we shall argue that these periodic optima have something to do with the packing of the polypeptide chain at the later stages of protein folding. The position where the curve in fig.2 has a minimum is a special economical configuration for domain sizes such that the most common protein domains are the ones with an amount of secondary structure elements (around 6-8) corresponding to the optimum point of chain links. Once the domain size is understood and correspond to 125 amino acids the periodicity is then easily explained be being multipla of elementary domains.

One might argue that the restriction of the chain to have links being only orthogonal to the proceeding ones is too limited in the sense that two consecutive parallel links should also be considered and counted for in the total energy. Therefore we also carried out a study where we included the case (with coupling L) of links going straight. This means that when it is being decided whether a link is going orthogonal to the previous link in the plane (J) or out of the plane (K) we shall also include the possibility of the link going straight ahead (L) from the previous one. This extra move possibility gave rise to a new list of configurations shown in the last column in Table I. The possibility of including the straight moves gave a much larger set of unique configurations but the behaviour exposed in fig.2 of minima at $7, 11,...$ number of links were still maintained in these extended numerical calculations.

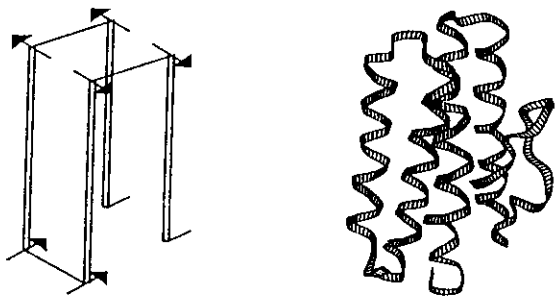


Fig. 1.

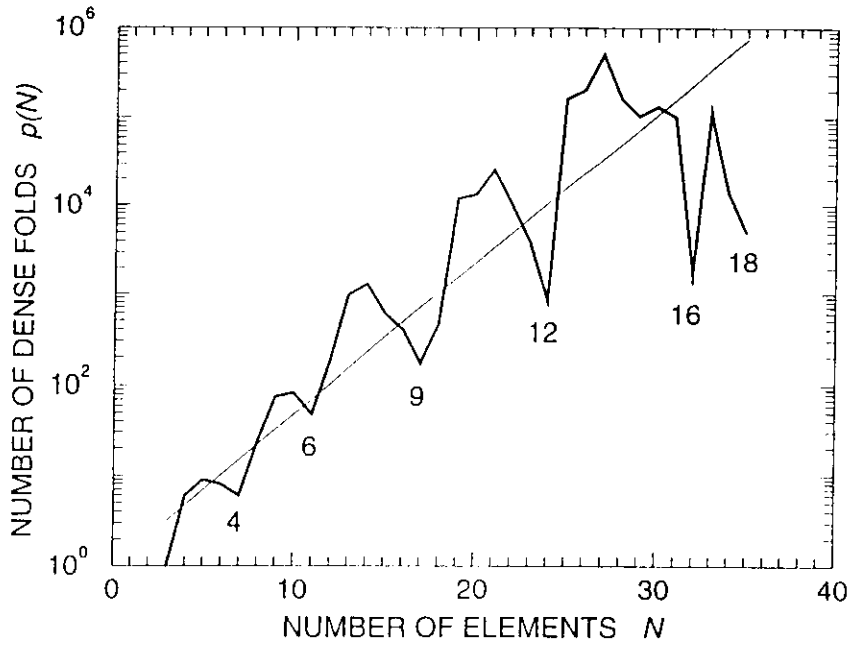


Fig. 2.

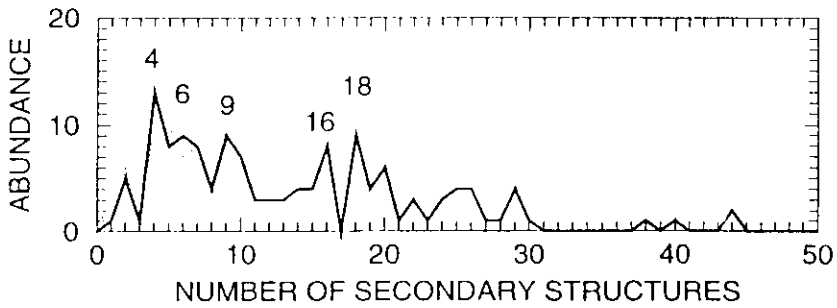


Fig. 3.

$2 \times 1 \times 1$ **Lattice**

Number of chain links	Number of (JK) dense config.	Number of (nn) nearest neighb.	Number of (JK) all config.	Number of (JKL) dense config.(inc.L)
3	1	8	4	1
4	6	10	14	8
5	9	14	39	12
6	8	18	80	8
7	6	24	130	6
8	24	26	160	36
9	76	30	180	164
10	84	34	100	192
11	48	40	48	146
12	0	0	0	0

Table 1: Calculation of unique chain configurations (most dense and all and straights) in a $2 \times 1 \times 1$ lattice for increasing number of links.

Furthermore one can make a study, of the statistics of optimal packed configurations for specific chain length as a function of different lattice sizes. As expected the number of configurations with the magic number of links for the $1 \times 1 \times 1$-lattice will remain the same for all greater lattices.

3.1 4c1. Graphical representations of the protein folds

Basically the philosophy behind our representation of folds is that the 3-dimensional protein structures can be represented in a uniquely way by a 1-dimensional string of coupling constants (J, K). We have given the prescription for how that can be done once the protein is partitioned in secondary structure elements.

The theoretical idea behind this is that proteins (or protein domains) from the same fold class will have the same string representation and belong to the same minimum of the energy function given in equation (5). However two proteins with different string representation could belong to the same energy minimum, e.g. proteins with different order of the (J, K) couplings can have the same energy.

One thing is theory, another is practice. One might have reservations as to whether it is possible to generate these string representations for any protein based on its raw data, e.g. X-ray crystallographic coordinate sets. In order to convince the sceptical reader we have constructed a big computer program including graphics software that can convert any set of protein coordinates in the PDB format into our representation of ordered chain links on a regular lattice. The representation can be given in a nice graphical form.

We shall here present a test of the prescription in the case of the *four helix bundle* fold

in a given protein indicating that the residue participates in that type of structure.

Secondary structures of a given type are identified as such if they contain at least 4 consecutive residues. The decision of how many residues that constitute a secondary structure is crucial for the statistical analysis of the abundance of secondary structure. In fig. 3 we have displayed the size distribution of secondary structures for all known proteins in the database. This clearly gives support to defining secondary structures as containing at least 4 consecutive residues. The β-strand distribution (see fig. 3) clearly has a maximum at 4 residues while the helix distribution has its maximum spread out over a plateau stretching from 4 to 12 residues. We actually performed statistics with a definition of helices containing more than 4 residues as a minimum requirement but that did not alter significantly the statistics of secondary structure abundance.

In making the secondary structure statistics of fig. 2 we have counted α-helix and 3_{10}-helix assignments as one type and all β-strands as another and then counted them all together. In fig. 2 we have displayed the abundance of the secondary structures as a function of their number. The curve clearly shows local maxima in the abundance, which correspond to the optimal packing we find theoretically.

We find optimal abundance at the following number of structure elements: $4, 6, 9, 16, 18, 24, \ldots$ etc and these optima are stable as to what size of the database we use, e.g. the first half of the dataset has roughly the same distribution as the second half of the set.

4 Magic Numbers

We shall now turn to the question of an atomistic grouping of packed structures of protein chains as considered in the previous chapter. From an analysis of packing and the effect of hydrophobic forces [26] we shall understand the appearance of 'magic numbers' of occurrences of proteins with specific size - and test the paradigm by a statistical analysis of available structural data.

Using the fact that the hydrophobic forces tend to confine the proteins and make them contain as little as 3% water [27] in the *native* state, we want to find all folds which are self-avoiding and densely packed. A scaling and mean field theory [32] of this problem gives the estimate that the number of folds for N elements increases as $(z/e)^N$, where z is the coordination number, in our case $z = 4$, and $e = \ln(1)$. For a protein with nine secondary structures and consequently eight interconnecting loop-elements we have $N = 17$, and the above theoretical relation gives the number of folds as $(4/e)^{17} \sim 711$. This is already a quite small number. However, the discreteness gives rise to *magic* numbers at which there are particularly few, different folds. Fig. 2. shows the exact enumeration of all possible folds for elements up to $N = 35$. For $N = 17$ there is a pronounced minimum with only $p(17) = 172$ distinct and predictable folds. The mean field theory overestimates this grossly. Between the magic numbers the abundance is on the other hand much larger. The magic number at $N = 7$, corresponding to the 4-α-helix bundle, is a close packing of a $1 \times 1 \times 1$-box. The next closed confinement is the $2 \times 1 \times 1$-box, which we call B. The magic numbers at $N = 11$, 17, 23, 32 and 35, can be understood as the optimal packing in closed polyhedra (analogous to shells) consisting of 1,2,3,5 and 6 B-boxes. On Fig. 3. is shown the statistical distribution of proteins with a specific number of secondary structure elements (as described in the last subsection).

The dense packing criteria we have used is a simple count of the neighbours of endpoints of the elements. This does in fact represent the hydrophobic force very faithfully. Firstly, it is unspecific *i.e.* independent of which elements are close to each other. Secondly, it depends on the 'curvature' of

the confinement approximately as a surface tension force, *i.e.* the different sites are rated 3,4,5 and 6 for a corner-, edge-, face- and a buried-site, respectively. Only the sum counts, in agreement with the nature of the hydrophobic force. One could, in order to introduce a temperature in the problem, assign energy values for the mentioned sites. This need not be a linear weighting. The found magic numbers are not very sensitive to deviations from a linear weighting which is still consistent with the globular structures. The magic numbers in our model are *universal* in the sense that they do not depend on the specific, chemical interactions between the amino acids: neither between distant parts of the chain nor the interaction along the backbone - they are dictated by the hydrophobic, confining forces. If the weighting is far from linear one can form other families of proteins. For example such that are dissolved in cell membranes. Clearly, for those the hydrophobic/phillic forces act differently. Families could be imagined with higher coordination number z or other projected lattices. We have investigated the closed packed folds for the simple cubic lattice case with $z = 5$, and find again a number of pronounced minima with the same magic numbers as before for the smaller domains.

We suggest that the folds at the magic numbers are particularly favorable for the following reasons. They represent closed confinements having minimal surfaces and thus being energetically favorable from the point of view of the hydrophobic forces. The magic number configurations have a clear energy separation from other folds. This is, according to the theory by Shakhnovich [34], a necessary condition for them to be able to fold rapidly. The configurational entropy for a fold at the magic number is low, and allows the large entropy of the extended chain to be exchanged by energy gain, without significant change in free energy. We are accordingly led to conclude that proteins with the magic numbers elements are more stable and fast folding than others. The minimum at $N = 17$, corresponding to nine secondary structures is relatively well pronounced. There is also a well pronounced minimum at the magic number $N = 35$. The $N = 35$ structure is confined in a $3 \times 2 \times 2$-box. An analysis of the folds shows that a large part are formed of two folds of the $N = 17$ domain interconnected by just a single element, *i.e.* $2 \times 17 + 1 = 35$. This explains why the domain formation is a natural consequence of the discrete packing problem. Given the average size of the elements, the magic numbers also rationalize why the size of the domains [30] is as preferred by nature, being in concord with the overall thermodynamic theory [27]. Next, we can estimate how many distinct fold classes there are to be found. If we restrict ourselves to domain structures with $N \leq 17$ we find in total 3906 possible, distinct globular fold-classes. This is close to Chothia's estimate of one thousand, based on an heuristic argument [35].

4.1 Magic numbers and the Euler characteristics

How to understand and construct the series of magic numbers for packing of the protein chain? As we have seen the magic numbers of secondary structure elements occur when the density of packing has a local minimum. At the position of a magic number there is a jump in the number of closest neighbours around each lattice site occupied by the chain.

We shall argue that the magic number occurs when the chain forms a closed surface (box) within the lattice. A good example is the 4-helix bundle at the magic number, $n = 4$, corresponding to $2 \cdot n - 1 = 7$ chain links which form a closed cube ($1 \times 1 \times 1$) that can be embedded in any other larger lattice.

The reason for the tendency to form closed surfaces in box configurations are the hydrophobic forces that tend to minimize surface area in which hydrophobic side-chains participate.

For closed surfaces we have the Euler equation that connects the number of vertices (corners), c, with that of edges, e, and surfaces, f. The formula is:

$$c - e + f = 2 - 2g \tag{9}$$

where g is the genus number. We shall in the following only be considering surfaces with no genus ($g = 0$) but our considerations are easily extended to surfaces with an arbitrary number of genus.

In case the total surface of the chain configuration is not closed the equation is not fulfilled but becomes instead:

$$c - e + f = n \neq 2 \tag{10}$$

This means that for a given chain with E links, the number E will be:

$$e = +f - n \neq 2 \tag{11}$$

where n is any natural number.

At the magic number n is close to zero and the jump in the number of neighbours is optimal (i.e. $\triangle n^V = 6$. The next magic number is obtained by adding a new closed box to the other in the lattice and see when it is filled out by the chain. For the case of the $(1 \times 1 \times 1)$ lattice alone the magic number can only be $n_{mg} = 4$ but in the case of the $(2 \times 1 \times 1)$ lattice that contains the $(1 \times 1 \times 1)$ box two times we can obtain the next magic number: $n_{mg} = 6$. In table I we can see that the calculation of chain configurations in the $(3 \times 2 \times 2)$ lattice renders all the magic numbers $\{4, 6, 9, 12, ..\}$ up to $n_{mg} = 18$ corresponding to 37 chain links.

Let us try to examine in details the cases where the chain is configured around a $(1 \times 1 \times 1)$ box and then having a few extra links. As we saw the elementary box was filled out well by the 4-helix protein chain and satisfying the Euler condition with 8 corners, 12 edges and 6 faces. This "magic" configuration of 7 links l has correspondingly 4 secondary structure elements $ss = (7 + 1)/2 = 4$. With an extra link added to these chain configurations we obtain one more corner and one more edge but no extra faces. We can count the extra nearest neighbours as being simply the sum of all the attributes, $+nn = +2$. With two more links (see fig.5) we have 2 extra corners, 3 extra edges and 1 extra face. By adding these extra quantities we get 6 minus the 2 from the last case making the extra nearest neighbours $+nn = +4$. By adding one more link we end having again $+nn = +4$. If we sum up all the corners, edges and faces for these cases with extra links including the extra corners etc. we cannot satisfy the Euler relation for this extended surface structure that is not closed in these cases as anticipated above. If we however add one more link (i.e. all together 4 links on the "magic" $(1 \times 1 \times 1)$ box we end up getting 6 more corners, 8 more edges and 4 more faces which altogether is $c = 12, e = 20, f = 10$ which we see satisfy the Euler relation again. We have arrived at the next magic number configuration of 11 chain links corresponding to 6 secondary structures. Furthermore we can count the extra nearest neighbours obtained by this configuration as being $+nn = +6$ which is precisely what is observed in the Table I of the numerical calculation of chain configurations. Going to the magic number configurations the average number of nearest neighbours jumps +6 from the previous configurations with one link less. We have thereby found a procedure for determining a magic number occurrence by using the Euler relation and counting the extra content of corners, edges and faces which thus gives us

the number of nearest neighbours and hence the density of the chain configuration. We can extend this prescription to more complicated lattice boxes, e.g. for the $(2 \times 2 \times 2)$ lattice.

The fact that the magic number configurations are more abundant and where the Euler relation is satisfied is due to the minimalization of surface area to that of volume. This is again due to the hydrophobic forces that tend to minimize the number of hydrophobic side-chains on the surface of the chain configuration.

There is, however, no simple and precise way of deriving the numbers which is why we call them magic like in nuclear physics where a quantum mechanical prescription (the Schrödinger eq. with a spin-orbit coupling potential) is needed for deriving the magic numbers of nucleons in the closed shells.

5 4e. The Molten Globule and the Parent states

So far we have only used the Hamiltonian (??) for enumerating the distinct folds found according to the hydrophobicity criteria. It is likely that the protein folding problem is an essential non-equilibrium phenomenon in the thermodynamic sense, and an energy function is only describing part of the process. Since the experimentally unknown time interval is large, ranging from $10^{-10} \rightarrow 10^{-3}$ sec. a proper theory for the dynamic folding processes in that interval is still far fetched. At most one can make a scenario, the details of which are to be resolved experimentally. First, the temperature is not a well defined concept - and can be replaced by the properties of the solvent; even at room temperature one can fold and unfold proteins by varying the amount of denaturants. However, with this in mind, let us follow common practice [29] and use the word temperature as indicating a measure for the degree of folding. We envisage the following scenario in line with recent observations [36].

At high temperatures the protein will be in an *extended* state because of the large phase space for this. When cooling down, the protein will start to form the α-helices because there is a clear chemical energy gain by forming hydrogen bonds between every third amino acid, and also a certain hydrophobic gain because of the contraction. This curling up does not require any global twist of the chain. The α-helices will have a limited length, because else it would require a global contraction of the chain in a viscous solvent. At this stage we then imagine the chain with a number of α-helices which straighten the intermediate chain pieces to potential β-strands and loops. The limited size α-helices can, however, almost freely run along the chain and find the optimum place according to the underlying amino acid code. The gross partitioning of the chain in secondary and intermediate structural elements is completed at this stage, which could also be partly arrived at *en naissance*. Any interaction between the elements is supposed to be switched off by screening effects of the solvent. It is at this *Molten Globule* stage we introduce our Hamiltonian (4). The hinge-forces (the values for the parameters in the Hamiltonian) are supposed to be given by the underlying sequence. They are of course very weak relative to other forces. How can they matter in the folding process? To see this, it is instructive

to look again at the Heisenberg magnet with the Hamiltonian

$$\mathcal{H} = -J \sum_{ij} \mathcal{S}_i \cdot \mathcal{S}_j - h \sum_i \mathcal{S}_i^z \,. \tag{12}$$

The strong interactions J cannot determine the spin direction in the fully rotationally invariant ordered state given by the dominant first term - but by introducing an infinitesimal field h in the z-direction the rotational symmetry is broken. It is the weak global force which determines the overall structure - the strong force determines the details.

This analogy is in fact deeply related to the present problem. The Hamiltonian (7) is a discrete one of the 'Ising' type. For simplicity in discussion we map the potential native fold (one of the closed packed states) onto a ferromagnetic Ising chain. As we have seen, with respect to the dominating hydrophobic forces this state is degenerate with a large number $p = p(N)$ of other states, which may be thought of as p different staggered Ising states for a protein with N elements. The lowest energy excitation for the chain, with respect to (1), is a soliton mode in which all spins to the left of one are flipped. This violates the valeur of only one letter in the descriptor (change in sign or type), whereas a single spin flip requires change of two bonds. The degenerate models all share the high energy excitation phase space. However, the low lying excited states are very different - in particular because a large number of excitations are prohibited by the non-overlap-constraint for the folds, and the energies of extended folds are augmented by hydrophobic energy. At moderate temperatures the states are essentially independent and separated by large energy barriers. We introduce this regime as a new intermediate stage. It is a volatile, high symmetry *Parent* stage corresponding to the b.c.c. phase. The hinge forces (which may include the effect of chaperones) sum up to give maximum energy gain for the potential native fold. The $p-1$ other states will have a higher energy according to how many letters in the descriptor have been violated. The effect is like that of the uniform field h in (2), and it is not sensitive to whether the hinge forces fit exactly to the final fold. So, without frustration the p-fold hydrophobic symmetry is broken. This demonstrates a natural relation between the sequence information and a preferred folding into the high symmetry fold corresponding to the native one.

6 4f. Conclusion

Summarizing, the hydrophobic forces cannot define a particular fold whereas the weak hinge forces set up a global force which will make a given protein fold predominantly in the right direction; even chaperones may here play a role. We believe that the proposed Hamiltonian in fact also makes sense in modeling the actual folding process from a certain stage. In our model we have at first neglected any forces between the secondary elements. This is an important conceptual aspect in our model for the not too late stages of the folding process. If specific amino acids on different elements could bind strongly it would fix the fold in any arbitrary configuration (imagine trying to fold double-glue-sided tape). The physical justification for switching off these forces is that they could be screened by the water, which accordingly must have an important 'lubricating' role to play during the folding. Only in the final approach to the dense fold the water is supposed to diffuse out and leave a problem for the final optimization of the short range chemical forces

between neighbouring elements. The result of that is undoubtedly the observed twistings and deformations of the actually observed structures. At that stage we have argued that the protein can not make any significant refoldings, so most of these forces would be frustrated if they do not happen to match according to the underlying sequence. We thus have argued that a match is not instrumental in the folding process, whereby our model is very different from previous theories, which precisely focus on this problem of frustrating forces, and led to a comparison between the folding problem and the spin glass problem [29]. In our model there is no frustration in setting up the main part of the folding. The end result will necessarily be frustrated and therefore the native state is not the ground state for the chemical forces from an equilibrium thermodynamic point of view. Results of using our Hamiltonian in an analysis of the dynamical folding process will be published elsewhere. We emphasize that the dynamical interpretation of the model, which is susceptible to future experimental tests, is independent of the already experimentally supported structural classification of the native states, discussed in the main part of this letter.

We have demonstrated that our model of the protein folding can rationalize the experimentally observed increased probability of finding structural domains with the magic number of elements. It may not be possible to asses if the magic structures are more stable at higher temperatures, because the formation of the secondary structures breaks up. In fact we could incorporate such a mechanism and extend the spin model by including internal excitations. At the molten globule stage there may be proteins with an unstable number of secondary structures 'decaying' into the stable ones, in quite close analogy to the shell model for nuclear matter.

References

[1] S. J. Prestrelski, A. L. EWilliams, Jr., and M. N. Liebman, Proteins **14**, 430 (1992).

[2] L. Pauling and R. B. Corvey, Proc. Natl. Acad. Sci. USA, **37** 729 (1951) *ibid* with H. R. Branson **37**, 205 (1951).

[3] P. Wolynes,

[4] P. -A. Lindgård, J. de Phys. IV, Colloque **C4**, 3 (1991).

[5] J. Krumhansl and R. J. Gooding, Phys. Rev. B **37**, 3047 (1989), R. J. Goodings and J. Krumhansl *ibid* **38**, 1695 (1988)

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

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

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

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

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

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

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

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

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

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

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

[17] A. G. Murzin and A. V. Finkelstein, J. Mol. Biol., **204**, 749 (1988).88

[18] A. V. Finkelstein and B. A. Reva, Nature, **351**, 497 (1991).

[19] E. Shakhnovich, Phys. Rev. Lett. 1994.

[20]

[21] D. Thirumalai, Phys. Rev. Lett. 1993.

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

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

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

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

[26] C. Tandford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, J. Wiley & Sons, New York, (1980).

[27] K. A. Dill, Biochemistry, **24**, 1501 (1985).

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

[29] P. G. Wolynes, *Protein Folds* Eds. H. Bohr and S. Brunak, CRC Press, New York p.3-17 (1995), M. Sasai and P.G. Wolynes, Phys. Rev Lett., **65** 2740 (1990).

[30] A. L. Berman, E. Kolker, and E. N. Trifonov, Proc. Natl. Acad. Sci. USA, **91**, 4044 (1994).

[31] L. Pauling and R. B. Corvey, Proc. Natl. Acad. Sci. USA, **37** 729 (1951) *ibid* with H. R. Branson **37**, 205 (1951).

[32] H. Orland, C. Itzykson, and C. de Dominicis, J. Phys. (Paris), Lett. **46**, L353 (1985).

[33] B. Rost and C. Sanders, J. Mol. Biol,, **232**, 584 (1993).

[34] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

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

[36] C. Redfield, R. A. G. Smith and C. M. Dobson, Nature Struc. Biol., **1**, 23 (1994).

[37] The proposed hinge-force assisted folding is in fact a much more direct and reliable process than the corresponding defect assisted selection of variants, which occurs in 'trained' shape memory alloys. Given the code for the hinge-forces the descriptor *i.e.* fold classes can be directly predicted from the sequence information.

[38] A. G. Murzin and A. V. Finkelstein, J. Mol. Biol., **749** (1988).

[39] P. -A. Lindgård and O. G. Mouritsen, Phys. Rev. Lett., **57**, 2458 (1986), Phys. Rev. B **41**, 688 (1990).

[40] T. Castán and P. -A. Lindgård, Phys. Rev. B **40**, 5069 (1989).

[41] E. Vives, T. Castán and P. -A. Lindgård, Phys. Rev. B. (1996). (accepted for publication).

[42] For the finite chain there is of course no phase transition in the strict sense, but it is replaced by a smoothed transition region.

5. The Implication of Topology for Protein Structure and Folding

In this section an analysis of the geometrical nature of the protein backbone is presented resulting in a winding topology for the peptide chain. The relationship to the writhing number used in knot diagrams of DNA is discussed. The winding state defines a long range order along the backbone of a protein and long-range excitations will exist (twist'ons). Energy can be pumped into these excitations, either thermally, or by an external force. A mechanism for the folding of proteins occur when the amplitude of a twist excitation becomes so large that it is more energetically favorable to curve the backbone. It is proposed that protein folding is a resonance phenomena.

5a. Introduction

The length of the polypeptide chain of globular proteins is much longer than typical diameters of the molecules. Yet, with the possible exception of some very short proteins, the polypeptide chain never displays a knotted topology [1, 2]. Thus, the disparity between different protein folds [3, 4, 5, 6] cannot be due to differences in their knot topology. Rather, the number of possible different protein structures is indeed limited by a "knot preventing mechanism". In this paper we address what further reduction of the phase space one may infer from topological arguments, and we present a model for the phenomenon of protein folding based on winding.

constraints can lead to a deterministic model of folding. In contrast to earlier views, topological folding may not necessarily lead to a folding path that minimize the sum of self-interactions. Rather, the physical reason for folding to follow topological constraints is the interaction of a protein with its environment.

2 Winding

Next we investigate the winding of the protein backbone. By this, we refer to the ubiquitous winding phenomena which may be observed in everyday life when dealing with items such as telephone cords, water hoses in gardens, pump hoses at self service gas stations *etc.* Basically, unless these tubes are handled with great care, their unwinding requires large motions in the space of the tube. The point of view put forward in this paper is that, as the protein folds in interaction with its environment (e.g. *in vivo*), such large motions of the protein backbone are unlikely. This is in contrast to spin-glass and lattice models which are based on self-interactions [11, 28]. Notice, that it would be a lot easier to wind the water-hose on the reel in the garden if there was no gravity. This is of course not physical.

Examples of winding considerations, a) a traight tube with no winding, b) an almost flat double loop structure with zero winding; the two loops have opposite chirality, c) an almost flat double loop structure with 4π winding; the two loops have identical chirality. For a shortened structure such as the one depicted in d), it is not possible to say whether it is a part of b) or c), and hence it is not possible to assign an unambiguous winding to it).

An illustration of what we mean by winding is shown in figure 1. The first part, a), shows a straight tube which has zero winding. The winding of the tube is defined as the number of rotations the end of the tube has made relative to the other end. This number is determined by the path of the tube. But, winding cannot be calculated as a continuous measure depending uniquely on the local geometrical progression of the curve. This is illustrated in figures 1b and 1c, where two nearly identical approximately flat structures are depicted. While, one is not winded (fig. 1b), the other is winded by two turns (4π), (fig. 1c). At first, it is therefore not possible to assign an unambiguous winding to a shortened path such as the one displayed in figure 1d.

A consequence of the above consideration is that in order to assign winding to a polypeptide backbone it is necessary to know what path the backbone has taken during folding. But, this folding pathway is most often unknown. Instead one must work with protein backbones that are extended to form a closed curve (or are extended to infinity). The winding of the backbone can thereafter be found as the *linking* of the curve. The linking of a closed curve is topological conserved. Below we discuss the relation between linking, twisting, and writhing [29].

Three examples of twist excitations displayed on a tube. a) Solitons, local increment (or decrement) in the twist amplitude. Two solitons of opposite twist have no long range implication for the twist amplitude. b) and c) show collective twist excitations over the entire tube. We call such excitations twistons. b) and c) have different boundary conditions.

3 Twistons

Associated with a particular path of the backbone is a geometrical orientation. It is normal practice to define a ribbon, or a frame, by assigning a vector-field for this purpose. Nevertheless, even in the absence of a ribbon, a vector-field can be defined. Observing

a line drawn by a pencil on an elastic tube one may easily be convinced about this. For a tube with rotational symmetry the incremental twist equals the torsion of the curve, equation (1), as such a geometrical-frame minimizes the twist energy. In equation (1), $\vec{r}$ is a vector representation of the curve and the primes denote derivatives. Notice, that the above geometrical frame may be different from the physical frame imposed by the backbone itself due to additional twist of the physical backbone.

$$\frac{\vec{r}' \times \vec{r}'' \cdot \vec{r}'''}{|\vec{r}' \times \vec{r}''|} \tag{1}$$

As there is a geometrical frame with long range order, twist excitations of the backbone will exist. Figure 2a shows an example of two solitons. In the limit where the backbone is considered to have continuous symmetry, the twistons displayed in figures 2b and 2c are the lower lying excitations. A pair of solitons may have a relatively low creation energy, but such pairs do not destroy the long range order, as they consist of an equal amount of clockwise and counterclockwise twist. The twist mode present depends on the boundary conditions at the two ends.

The basic consideration for linking the winding property and the protein folding problem may be addressed in two conjectures: *(a)* the path of the segments of the polypeptide chain trace out a "simple motion", *(b)* the backbone does not rotate unnecessary. The physical reason for this is that proteins do not fold in vacuum but rather in a viscous aqueous medium. Often they also interact with carbohydrates, other proteins and membranes.

The twistons are long range collective excitations over an entire protein folding domain. The polypeptide backbone will begin to bend at a certain amplitude of the local twist. The twist mode will involve non-zero values for the dihedral angle ω and therefore be rather stiff. The characteristic time involved can be much shorter than the characteristic time associated with the random motion of the unfolded backbone.

Part of the polypeptide backbone depicting the dihedral angles ψ, ϕ, and ω. The shaded area is approximately planar.

In general a change in the dihedral angles, ϕ and ψ will lead to a change in the path of the backbone [30]. In contrast, a change in the twist maintain the path of the backbone. A twist mode therefore involves strained chemical bonds. However, the dihedral rotations ψ_{i-1} and ϕ_i are almost coaxial (see fig. 3). However, such a rotation, and counter rotation, does not have any long range implication for the twist of the backbone and thus does not interfere with long wavelength twist modes.

4 Resonator driven transition

We hypothesize that the phase transformation of a protein from the unfolded structure to the folded structure is initiated by excitations of long wavelength twistons of the backbone which become unstable in favor of curvature. The nature of the transition may be characterized as being catastrophic rather than entropic. By this we mean that the primary reason for the transition is not a change in entropy. The excitation of the twist mode is pumped to a higher and higher level. A resonator is responsible for this pumping of the twist mode. The resonator must continuously be re-energized such; for thermal fluctuations by contact with the thermal bath. Almost literately, the initial folding of the protein can be thought as being analogous to the famous collapse of The Tacoma Narrows Bridge in Seattle, 1940. Twist modes of the bridge are excited by strong winds. Eventually, the amplitude of the twist modes becomes so large that the bridge fractures. The bridge did not have the option, that proteins do, to form folded structure.

Figure 1
Resonator driven protein folding
J. Bohr, H. Bohr and S. Brunak

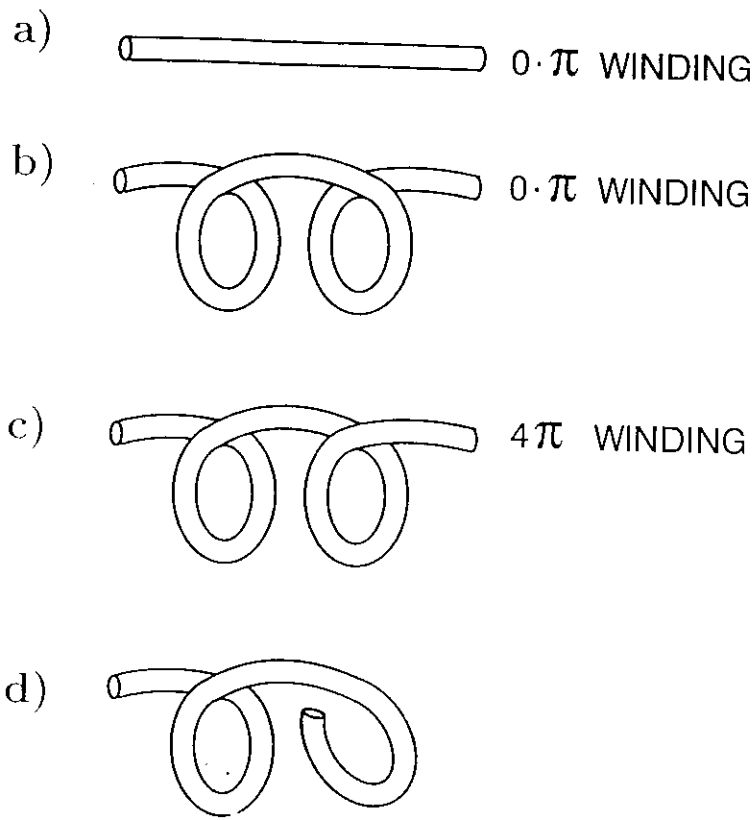


fig12.eps

Figure 2
Resonator driven protein folding
J. Bohr, H. Bohr and S. Brunak

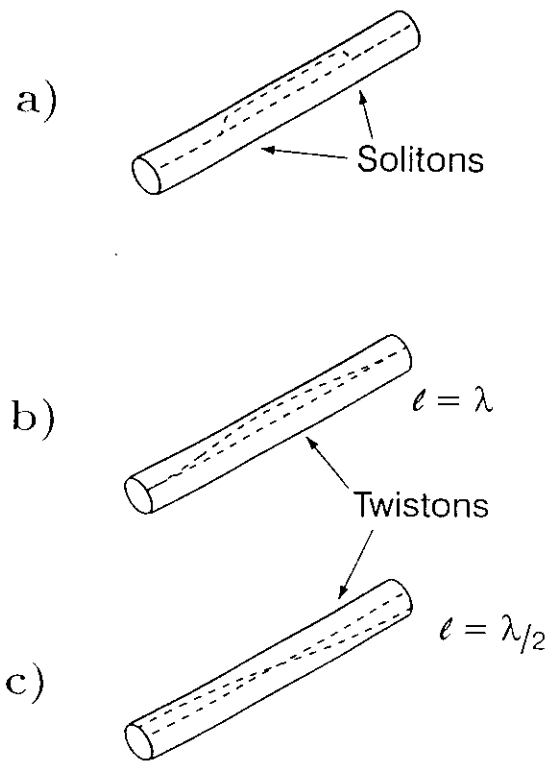


fig13.eps

Figure 3
Resonator driven protein folding
J. Bohr, H. Bohr and S. Brunak

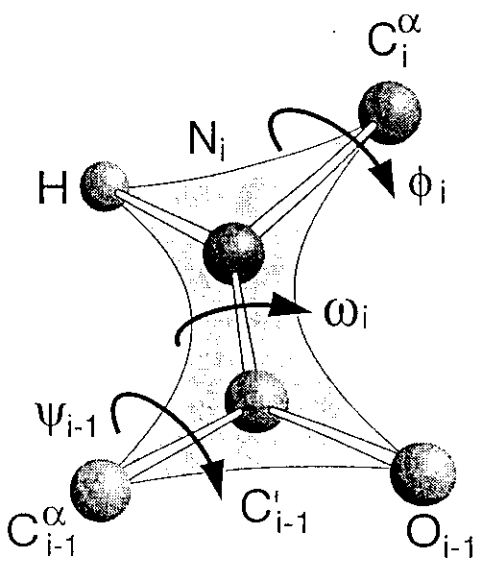


Figure 4
Resonator driven protein folding
J. Bohr, H. Bohr and S. Brunak

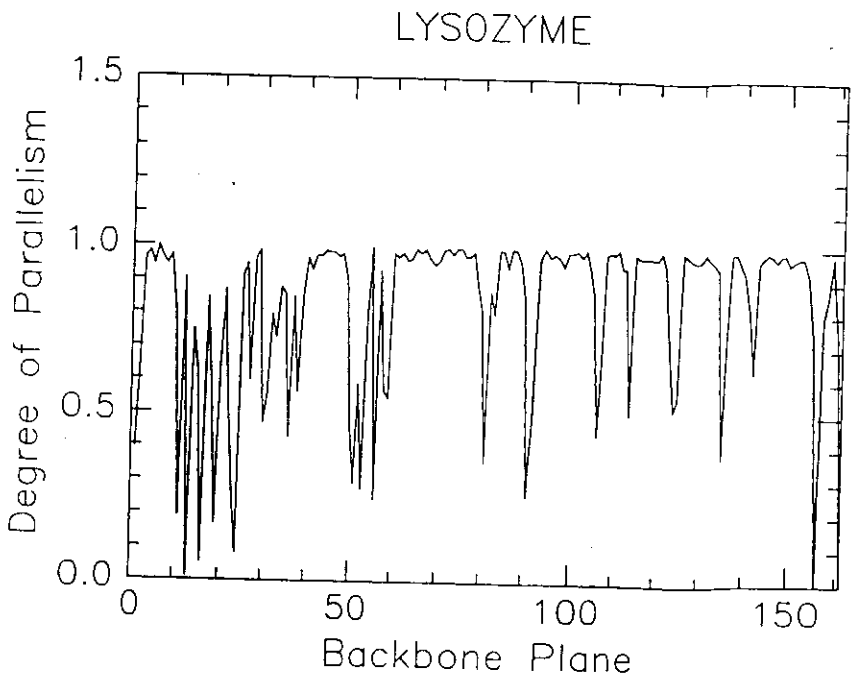


Figure 5
Resonator driven protein folding
J. Bohr, H. Bohr and S. Brunak

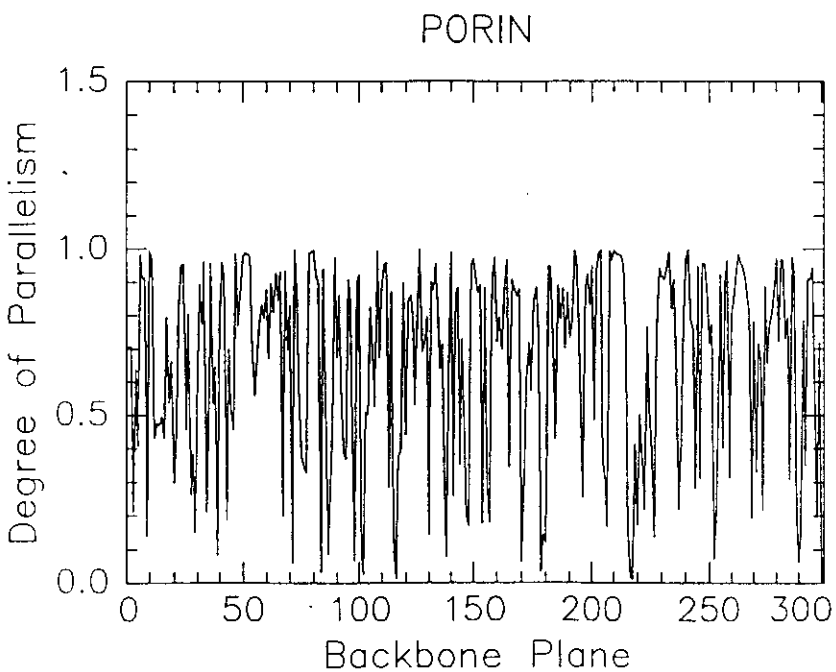
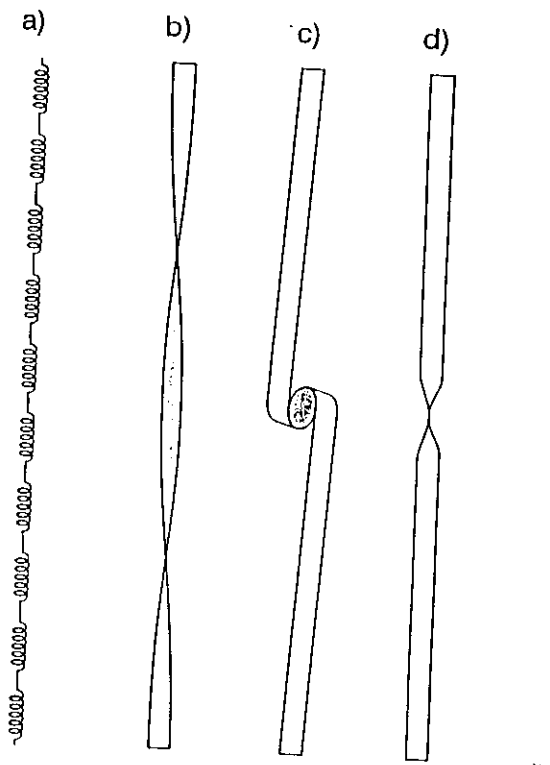


Figure 6 Schematic drawing of a part of a chain molecule: a) as a chain of springs. In this case the energy threshold for inducing breaking is, $(N-1)E_{max}^{spring}$. In b) as a twisted ribbon. In this case the molecule will undergo conformational changes when $2k_r < k_s$, and the excess energy can be concentrated in a relatively small part of the molecule as shown in c) and d).



The degree of parallelism in lysozyme between the geometrical frame associated with the backbone path and the planes defined by the backbone atoms, see fig. 3. The frequently observed values of nearly 1 reflect the almost prefect parallelism found in α helices.

What is the "pump"? At this point we shall restrict the discussion to making it plausible that such a resonator exists by mentioning a set of alternative possibilities. It could originate from thermal vibrations or rotations in the sidechains. It could originate from the rhythm of the ribosome translational process, from fluctuating carbohydrates, or from external molecules such as chaperonins [31].

A number of unique features follow from the scenario of a pumped transition. The resonance would be sensitive to the length of the polypeptide chain. This is consistent with protein folding domains all being of roughly the same length [32, 33]. For example, insulin is synthesized as a single chain which is folded before it is cleaved into its constituents being 21 and 30 amino acids long [34]. Carbohydrates with large masses can define folding domains by constraining the twist modes. Protein folding can be promoted by pauses in the ribosome translation process if halts [35, 36, 37, 38] are engineered to come after a particular part of the peptide chain has been completed. It is natural to interpret this phenomenon as a means of selection of length, see below in the summary. The part of the protein which would first begin to fold will depend on the twist mode, and on the detailed location of the different amino acids (the bridge broke at the 1/4 point — corresponding to maximum torsion). This means that the protein folding path to a large extend would be deterministic and that folding could be a fast process. Finally, winding symmetries and conservation laws will exist, depending on the twist mode.

5 Linking and writhing

The heuristic description of the winding state of a protein can be extended to a more well founded continuous measure using differential geometry and knot theory [39]. The topological conserved winding is given by the linking number of a closed curve. Define a ribbon as the geometrical frame where the twist is equal to the torsion of the curve. A linking number of zero means that if one cut the ribbon (with a scissor) along a central line, one will end up by two non-interwoven ribbons. If the linking number is ± 1 you will end up by two ribbons that are linked as links in a chain.

The linking number, L, is related to the writhing number, W, and the total twist, T, through the White theorem

$$L = W + T \tag{2}$$

The vectors $\vec{t}$, $\vec{v}$, and $\vec{v}^{\perp}$ define a right-handed frame of the curve; $\vec{t}$ being a unit vector parallel to the velocity $\vec{r}'$. The total twist can then be calculated as

$$T = \frac{1}{2\pi}\int_0^l \vec{v}^{\perp}\cdot d\vec{v} \tag{3}$$

and the writhing as

$$W = \frac{1}{4\pi}\int\int_{c\times c} \frac{\partial\vec{e}}{\partial s_1}\times\frac{\partial\vec{e}}{\partial s_2}\cdot\vec{e}\, ds_1\, ds_2 \tag{4}$$

where

$$\vec{e} = \frac{\vec{r_2}-\vec{r_1}}{|\,\vec{r_2}-\vec{r_1}\,|} \tag{5}$$

Writhing can be formed on the expense of twist. This may commonly be observed in double twisted telephone cords. A concept that was introduced in discussions of circular

supercoiled DNA [15, 16, 17, 29], and has been suggested to be associated with protein folding as well [23, 26, 27].

How to calculate the writhing along the backbone of the protein? Writhing must be calculated as a double integral [23], a fact that otherwise have been ignored in studies of proteins [26, 27]. Two fundamental problems arise. One being the issue of how to extend the backbone into a closed curve, and the other being concerned with the fact that the position of the curve is known at discrete points as given by X-ray or NMR measurements. In the limit $\omega_i = 0$, the polypeptide backbone consists of planar plates joined together at the C_α atoms by the two other dihedral angles. Due to the relatively free rotations of these planes, which in particular is the case when glycine is involved, one cannot calculate the winding state solely by inspecting the planes. Firstly, between each pair of planes there is a right/left twist ambiguity. And as the rotations often are large it is not adequate to let the right/left question be settled simply by which angles are smallest. Secondly, the unfolded state is unknown. A given physical representation of the ribbon is not necessarily untwisted in the unfolded state. But, in contrast, the geometrical frame described above can provide us with a way of calculating meaningful linking numbers. Notice, that the sign of torsion (1) and the chirality is determined by the right/left handedness of the local frame. However, for backbone geometries where the chirality is a global property, the chirality becomes associated with the sign of the writhing (see fig. 1).

The degree of parallelism in porin between the geometrical frame associated with the backbone path and the planes defined by the backbone atoms (see fig. 3).

It is interesting to compare the orientation of the planes of atoms with the geometrical frame of the folded structure as it provides us with information about the preferences of the folded structure. For this purpose the backbone is characterized by a set of points that are midpoints between the C_α atoms, and the successive planes defined by normal vectors given by $\vec{O_iC_i'} \times \vec{N_iC_i'}$. The coarse grained mesh of points on the backbone leads to some sign errors in calculating the geometrical frame. A 1 Å random motion leads to about 4 sign changes in every 100 residues. In figure 4 the absolute value of the scalar product of the vector of the geometrical frame and the normal of the physical plane is shown for lysozyme (7LZM). The absolute value of the scalar product is not sensitive to sign errors. For lysozyme it is very close to 1 for the majority of the residues; there is a high degree of parallelism between the geometrical frame and the physical planes. The reason for this is due to the relative rigidity of the α-helix structures which are abundant in lysozyme. Figure 5 shows the result obtained for porin (2POR). The tendency to parallelism between the geometrical frame and the physical frame is less clear, because of the lower degree of rigidity of the β-strands, which are abundant in porin. In porin the sign is alternating most of the time. This reflect the alternating orientations of atomic planes making up the β-strands. It shows an inherent problem with winding since such a behaviour is consistent with a collapse of a spiral where the period is reduced to two residues. As such, it could explain the frequently observed alternating motif between α-helices and β-strands as seen in tim-barrels [40]. We will not further develop this point of view here, but stress the ambiguity it leaves.

Eigenfrequencies

An upper limit for the energy stored in a twist mode can be estimated by considering the energies of the chemical bonds of the backbone. We take a rotation of $\pi/2$ of one bond to correspond to about 1 eV/Å, in accordance with typical bond energies, although, the energy involved in strained chemical bonds is often smaller. Hence, the torsion constant per inverse unit length, y, is limited to about 0.4 eV/Å. The moment of inertia of the backbone (per unit length). i, is about 100 a.u.Å, depending on the degree to which the sidechains are involved in the twisting. The eigenfrequency can be estimated to be

$$\nu = \frac{1}{2\pi L}\sqrt{\frac{y}{i}}$$

where L is the length of the backbone. Equation (10) can be derived as the classical solution to a torsional Lagrange equation, or by modifying the equation describing the Young's modulus. For a typical folding domain, *e.g.* 125 amino acids, L is about 475Å, and the cyclic frequency ν becomes about 2.1 GHz. This frequency corresponds to rotational frequencies of sidechains, and is slightly higher than the rotational frequency of short peptides. The numeric estimate given above assumes a linear torsion term. However, the torsion term is not linear, and it is therefore possible that for small amplitudes of the wring modes the above estimate of ν is as much as a factor of 10 too high.

The relaxation time for the rotational motions of the atoms in the backbone has been measured by ^{13}C nuclear magnetic resonance, and it was found that the relaxation time is much longer than the approximately 10^{-12} s that should be expected for individual atoms. Instead, values typically measured are about $5.4 \cdot 10^{-10}$ for *Lys* C^{γ} in ribonuclease (Glushko et al., 1972). This five-hundred fold enlargement of the relaxation time corresponds well with equation (10) where it was shown that the relaxation time would scale proportionally to the length of the polypeptide chain. The corresponding frequency, 185 Mhz is about a factor of five lower than equation (10) gives. It is worthwhile to notice that this is a frequency range (100 – 600 MHz) where microwave absorption has been observed for proteins [Buchanan et al., 1952; Harvey and Hoekstra, 1972; Miura et al., 1994). This has been interpreted as rotational spectra of *bound water molecules*. It would require that the rotation of bound water molecules is damped to such a degree when compared to free water that the frequency is reduced by a factor of about one hundred. Such a damping must be almost *critical*, and it seems therefore somewhat peculiar that microwave absorption can be observed over a very broad temperature range of about one hundred degrees. It seems more plausible that the damping would go critical and the corresponding frequency therefore to zero. The reason that the signal was interpreted as coming from bound water is that it is unambiguously linked to the presence of the first monolayer of water on the protein. We suggest a possible alternative explanation, namely, that without the first monolayer of water the protein will not be sufficiently topological constrained to support wring modes. This possibility that the microwave absorption is caused by wring modes in the protein, rather that by the water molecules needs to be addressed in more detail experimentally.

Summary

One of the paradoxes in our current understanding of protein folding is how fast it really goes. For example it has been estimated that the time required to sample all possible conformations would be 10^{77} years (Levinthal, 1968; Creighton, 1993). It has been suggested that the existence of a bias for secondary structure formation can reduce the folding time sufficiently to explain the fact that many proteins folds in a fraction of a

second. The hypothesis for protein folding presented in this paper is consistent with fast folding as the phase space is not searched for possible conformations.

The geometrical winding of polypetide backbones has been considered and it has been shown that long wavelength excitations exist. Simple conjectures for the process of protein folding lead to constrains for the winding of the backbone of the folded state of proteins. But as the unfolded structure of the polypeptide chain is unknown no unique winding can be prescribed. It is hypothesized that the transition from the unfolded state to the folded state of a protein is due to a catastrophic transition, when a twist mode of the protein backbone becomes unstable to curvature. Energy is driven into the twist mode by a resonator.

References

[1] T. E. Creighton, *Protein, Structure and Molecular Properties*, 2nd edition, Freeman, New York, 1993.

[2] M. L. Mansfield, Are there knots in proteins?, *Nature Structural Biology*, 1, 213-214, 1994.

[3] J. Richardson, The anatomy and taxonomy of protein structure, *Adv. Prot. Chem.*, 34, 167-339, 1981.

[4] C. Chothia, Proteins: One thousand families for the molecular biologist, *Nature*, 357, 543-544 1992.

[5] C. Chothia and A. V. Finkelstein, The classification and origins of protein folding patterns, *Ann. Rev. Biochemistry*, 59, 1007-1039 1990.

[6] J. U. Bowie, R. Luthy and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, 253, 164-170, 1991.

[7] C. B. Anfinsen and H. A. Scheraga, Experimental and theoretical aspects of protein folding, *Adv. Prot. Chem.*, 29, 205-300, 1975.

[8] T. E. Creighton, Conformational restrictions on the pathway of folding and unfolding of the pancreatic trypsin inhibitor, *J. Mol. Biol.*, 113, 329-341 1977.

[9] D. B. Wetlaufer and S. Ristow, Acquisition of 3-dimensional structure of proteins, *Annu. Rev. Biochem.*, 42, 135-158, 1973.

[10] R. L. Baldwin, Intermediates in protein folding reactions and the mechanism of protein folding, *Ann. Rev. Biochem.*, 44, 453-475, 1975.

[11] P. G. Wolynes, Spin glass ideas in the protein folding problem, in *Spin glasses and biology*, D. L. Stein (ed.), World Scientific Press, New York 1990.

[12] R. L. Balwin, How does protein folding get started?, *Trends Biochem. Sci. Pers. Ed.*, 11, 6-9, 1986.

[13] P. E. Wright, H. S. Dyson and R. A. Lerner, Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding, *Biochemistry*, 27, 7167-7175, 1988.

[14] J. P. Waltho, V. A. Feher, G. Merutka, H. S. Dyson and P. E. Wright, Peptide models of protein folding initiation sites. 1. Secondary structure formation by peptides corresponding to the G- and H- helices of myoglobin, *Biochemistry*, 32, 6337-6347, 1993.

[15] F. H. Crick, Linking numbers and nucleosomes, *Proc. Natl. Acad. Sci. U. S. A.*, 73, 2639-2643, 1976.

[16] F. B. Fuller, The writhing number of a space curve, *Proc. Natl. Acad. Sci. U. S. A.*, 68, 815-819, 1971.

[17] F. B. Fuller, Decomposition of the linking number of a closed ribbon: A problem from molecular biology, *Proc. Natl. Acad. Sci. U. S. A.*, 75, 3557-3561, 1978.

[18] W. Helfrich and W. Harbich, in *Physics of Amphiphilie layers*, D. Langevin and N. Boccasa (eds.), 58, Springer, Berlin 1987.

[19] H. Bohr and J. H. Ipsen, Differential geometry and topology of biological membranes, in *Characterizing Complex Systems*, H. Bohr (ed.), World Scientific, Singapore 1989.

[20] J. H. White, Self-linking and the Gauss-integral in higher dimensions, *American J. of Math.*, 91, 693-728, 1969.

[21] W. Helfrich, Elastic properties of lipid bilayers: Theory and possible experiments, *Z. Naturforsch.*, 28c, 693-703, 1973.

[22] R. E. Goldstein and S. Leibler, Model for laminar phases of interacting lipid membranes, *Phys. Rev. Lett.*, 61, 2213-2216, 1988.

[23] M. Levitt, Protein folding by restrained energy minimization and molecular dynamics, *J. Mol. Biol.*, 170, 723-764, 1983.

[24] S. Rackovsky and H. A. Scheraga, Differential geometry and polymer conformations I, *Macromolecules*, 11, 1168, 1978.

[25] S. Rackovsky and H. A. Scheraga, Differential geometry and polymer conformations II, *Macromolecules*, 13, 1440, 1980.

[26] P. De Santis, S. Morosetti and A. Palleschi, Topological aspects of the conformational transformations in polypeptides and proteins, *Biopolymers*, 22, 37-42, 1983.

[27] S. Chiavarini, P. De Santis, S. Morosetti and A. Palleschi, Topological aspects of conformational transformations in proteins, *Biopolymers*, 23, 1547-1563, 1984.

[28] E. I. Shahknovich, Proteins with selected sequences fold into unique native conformation, *Phys. Rev. Lett.*, 72, 3907-3910, 1994.

[29] W. F. Pohl, DNA and Differential Geometry, *Math. Intelligence*, 3, 20-27, 1980.

[30] A. M. Lesk, *Protein Architecture*, Oxford University Press, Oxford 1991.

[31] M. Gething and J. Sambrook, Protein folding within the cell, *Nature*, 355, 33-45, 1992.

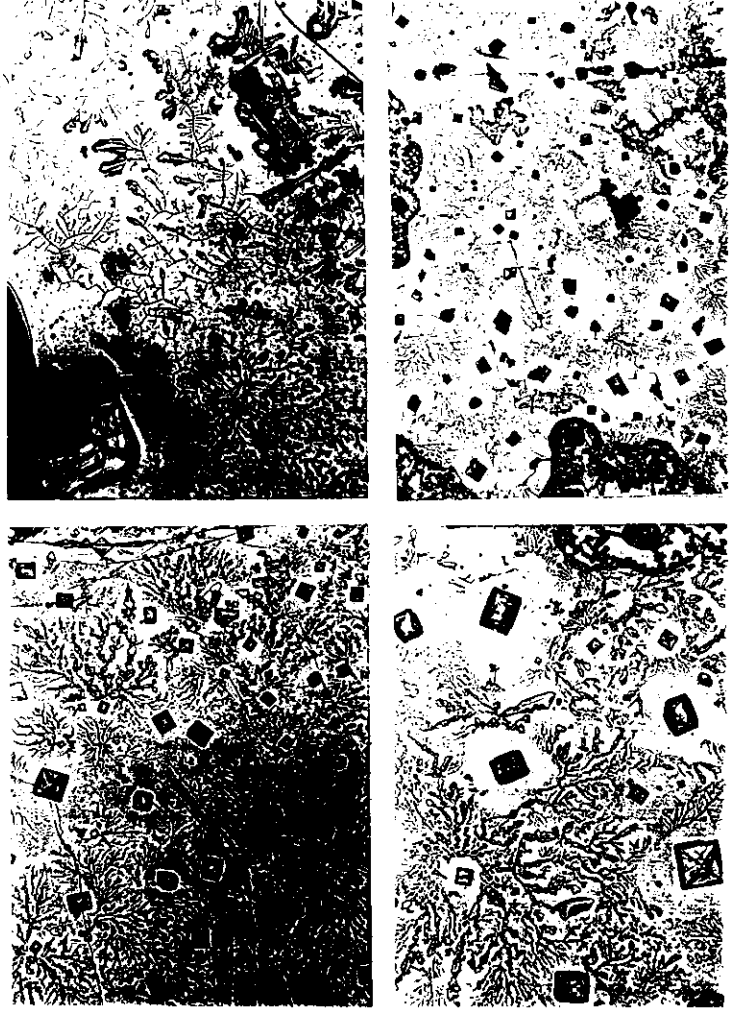
[32] J. Janin and S. Wodak, Structural domains in proteins and their role in the dynamics of protein function, *Prog. Biophys. Molec. Biol.*, 42, 21-78, 1983.

[33] J. R. Garel, Large multi-domain and multi-subunit proteins, in *Protein Folding*, T. E. Creighton (ed.), Freeman, New York 1992.

[34] J. Kendrew, *The Encyclopedia of Molecular Biology*, Blackwell Science, Oxford 1994.

[35] P. M. Sharp, and W. Li, Codon usage in regulatory genes in *Escherichia coli*, *Nuc. Acids Res.*, 10, 7737-7749, 1986.

LIGHT MICROSCOPE PICTURES
OF AGGREGATED RIBONUCLEOPROTEIN

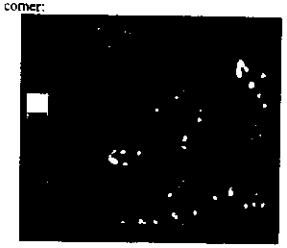


10.

14

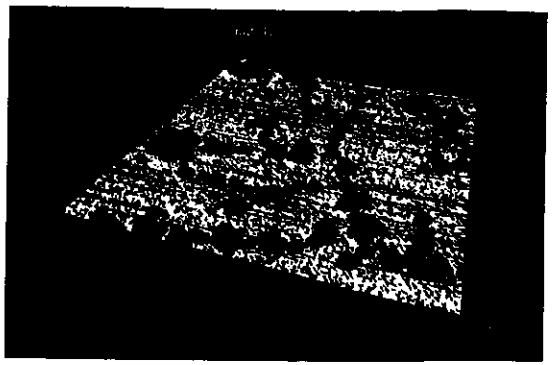
Zoom-in on upper left corner:

15

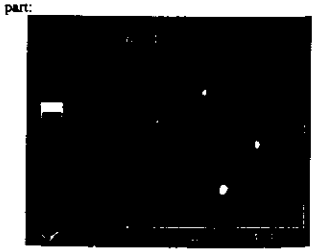


Notice presumably pronounced tip-convolution: particles have high aspect ratios.

16



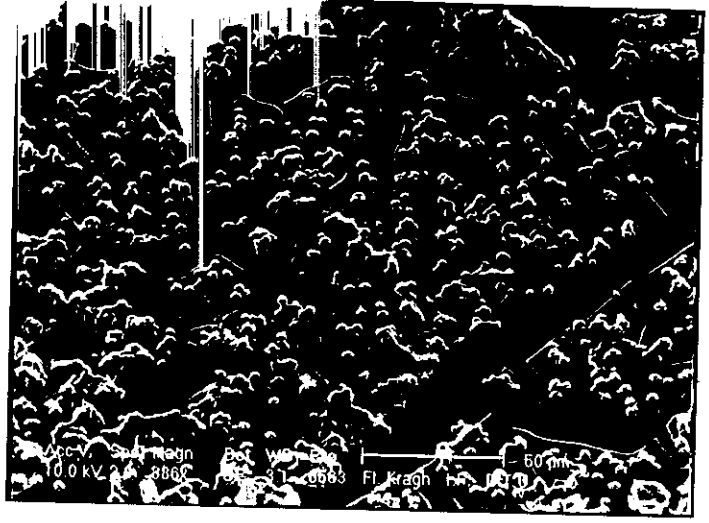
Zoom-in on central part:

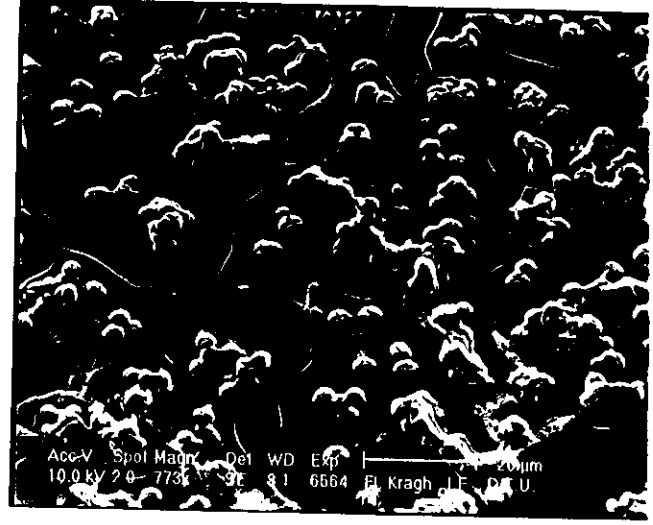


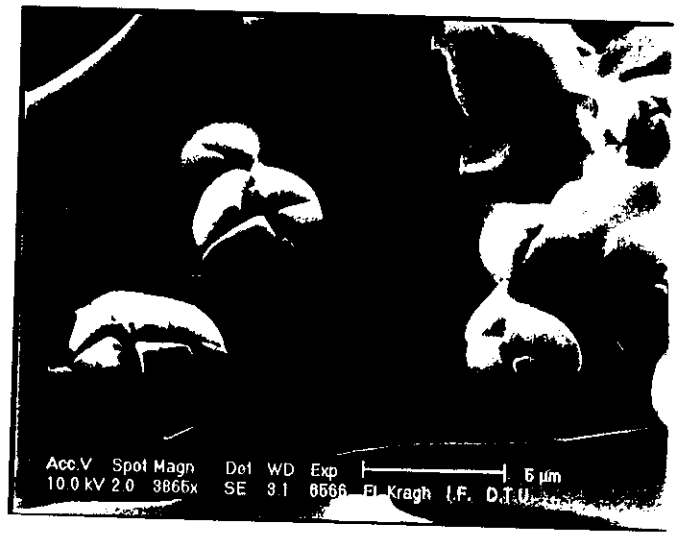
Last update Feb. 20. 1996 by
Anders Kühle: anders@carbon84.fysik.dtu.dk.

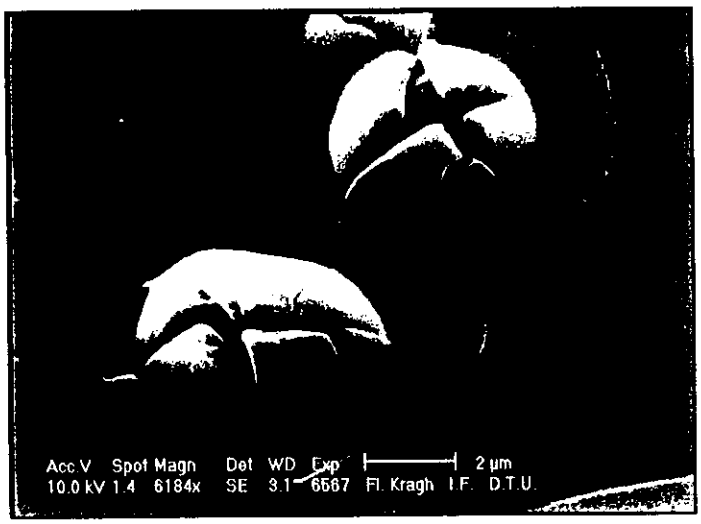
17

ATOMIC FORCE MICROSCOP. OF AGGREGATED RIBONUCLEATE
(PRECIPITATED ON GRAPHITE









ELECTRON MICROSCOPY OF LYSOZYME AGGREGATES

II

[36] P. M. Sharp, T. M. F. Tuohy and K. R. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, *Nuc. Acids Res.*, 14, 5125-5143, 1986.

[37] I. A. Purvis, J. Bettany, T. C. Santiago, J. R. Coggins, K. Duncan, R. Eason and A. J. Brown, The Efficiency of folding of some proteins is increased by controlled rates of translation *in vivo* — A hypothesis, *J. Mol. Biol.*, 193, 413-417, 1987.

[38] S. L. Wolin and P. Walter, Discrete nascent chain lengths are required for the insertion of presecretory proteins into microsomal membranes, *J. Cell. Biol.*, 121, 1211-1219, 1993.

[39] L. H. Kauffman, *Knots and Physics*, World Scientific, Singapore 1991.

[40] C. Branden and J. Tooze, *Introduction to protein structure*, Garland Publishing Inc, New York 1993.

6. MICROSCOPY OF PROTEIN AGGREGATES
(SEE PICTURES)

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 [1]. 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.

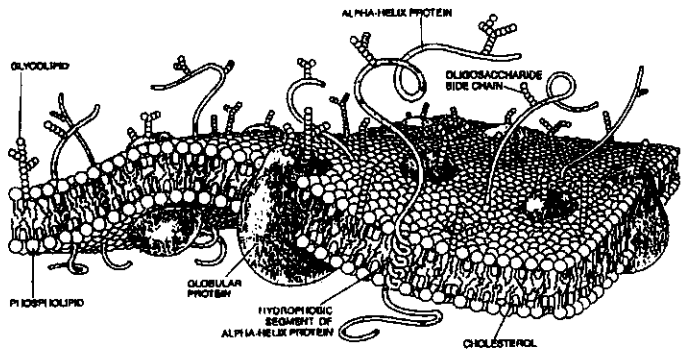


Figure 15. A popular picture of the biological membrane with proteins etc..

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 [2]. 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 [3].

A popular phenomenological theory for the description of shapes of individual lipid bilayers is the Canham-Helfrich model [4]. This model has even proven to be successfull

in the description of shape transformations of biological membranes like erythocytes [5]. Further this model has served as the basis for recent studies on the effects of thermal ondulations on the forces between bilayers [6] and the possibility of order-disorder transitions in macroscopic conformations of membranes [8, 9]. 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 [10].

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 [1] 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 [11] and freeze-fracture electron microscopy [12]. 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 [13, 14].

A considerable effort has been directed toward characterization of the phases in terms of surface geometries [20]. 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 [15]. 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 $\mathbf{R^3}$.

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 [16]. 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 [17]. 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 [18] and fourier analysis of the thermal membrane undulations [19].

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 $W(\tilde{\Sigma}) = W(\Sigma)$ [21]. 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 [22] we write $\beta_g = inf W(\Sigma)$, where inf is taken over compact genus g sufaces without self-intersections. The following inequality is fundamental:

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

Equality holds on the left if and only if $g = 0$ and Σ is a round sphere [21]. The right hand side inequality was observed independently by U. Pinkall and R. Kusner, see [23]. Simon then showed [22], that if we put $e_g = \beta_g - 8\pi$, then

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

for any integers $q \geq 2$ 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 $g_o \leq g$) such that $W(\Sigma_k) \rightarrow W(\Sigma_o) = \beta_{g_o}$ for $k \rightarrow \infty$. Thus for a given surface the minimization of W may cause a drop in genus number. However, the fundamental conjecture [25] 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.

We are now in the position to set up the partition function for a single closed membrane made out of A lipids. Note, by introduction of a fixed membrane area we break the conformal invariance explictly.

$$\begin{aligned} Z(A) &= \sum_{g=0}^{G_A} \mathrm{Tr}_g(exp(-\beta\mathcal{H})) \\ &= \sum_{g=0}^{G_A} exp(-4\pi\bar{\kappa}(1-g))\mathrm{Tr}_g(exp(-\beta\frac{\kappa}{2}W)) \end{aligned} \tag{132}$$

Tr_g represents the summation of all physical distinct surfaces Σ_g with fixed area corresponding to A. Surfaces which are identical apart from a reparametrisation of the surface are not physically distinct. G_A represents a cut-off in the number of genus for a membrane of size A. G_A exist if the diameter and the sectional curvatures have upper bounds [24]. This is indeed the case for a closed membrane due both to a limited number of lipids A

involved and to material parameters determined by size and physical properties of the molecular constituents. A first approximation to $Z(A)$ can be obtained by restricting Tr_g to surfaces which realize the the minimum of W. This is according the previous considerations the case for one surface for each g. The struture of this surface is not known for general g at present, which makes it impossible to go beyond this simple thermodymanic level of description. We will further reduce the number of degrees of freedom and assumes that the membrane is motionally hindered, so g is the only available degrees of freedom and the minima of $W(\Sigma)$ are non-degenerate. The further evaluation of the partition function lies, strictly speaking, in the case 1 where we sum over the number of handles g

Case 1.:

$$Z(A) \approx Z^{\mathrm{SP}}(A) = \sum_{g=0}^{G_A} exp(-4\pi\beta\bar{\kappa}(1-g))exp(-\beta\frac{\kappa}{2}inf_g(W)) \tag{133}$$

By use of Eq.(4) and Eq.(7)

$$Z_1(A) < Z^{\mathrm{SP}}(A) < Z_2(A) \tag{134}$$

where

$$Z_c(A) = exp(-4\pi\beta(2\kappa+\bar{\kappa})) + exp(-8\pi\beta c\kappa)\frac{1-exp(4\pi\bar{\kappa}\beta G_A)}{1-exp(4\pi\beta\bar{\kappa})} \qquad c = 1,2 \tag{135}$$

For $G_A \to \infty$, Z_c and Z^{SP} are analytic for $\bar{\kappa} < 0$, and Z_c is solely controlled by the second term in Eq.(10) for $\bar{\kappa} \to 0^-$. From Eq.(9) it is evident that Z^{SP} is also governed by a singularity of this nature as $\bar{\kappa} \to 0^-$. Z_c can thus be considered as a good approximation for Z^{SP} under these conditions. Note that Z_1 in this limit has the form of the partition function for a quantum harmonic oscillator. However $1 \ll G_A < \infty$ is the regime of interest in the description of the physical system. Here Z_c and Z^{SP} are analytic for $\bar{\kappa} \neq 0$. The inequality in Eq.(9) holds term by term in an expansion of Z_c and Z^{SP} in $\exp(4\pi\bar{\kappa}(1-g))$ like Eq.(8). It is then trivial to show that the inequality hold term by term in an expansion in $\bar{\kappa}$. The expansion coefficients are thermal expectation values which will be considered in the following.

The free energy derived from Z_c, $F_c(A) = -\beta^{-1}ln(Z_c(A))$, and its thermal behaviour may be analyzed. The obvious orderparameter in this problem is the averaged genus $< g >$ or the average Euler-characteristic $< \chi >=2(1- < g >)$.

$$\begin{aligned} < \chi > &= -\frac{1}{2\pi}\frac{\partial F_c(A)}{\partial(\beta\bar{\kappa})} \\ &= -\frac{1}{\beta\cdot 4\pi}\frac{1}{Z_c(A)}\frac{\partial Z_c(A)}{\partial(\beta\bar{\kappa})} \end{aligned} \tag{136}$$

The fluctuations in $< \chi >$ can be expressed as

$$\begin{aligned} \sigma(\chi) &= -\frac{\partial^2 F_c}{\partial(\beta\bar{\kappa})^2} \\ &= \beta^{-2}(\frac{1}{Z_c(A)^2}\frac{\partial^2 Z_c(A)}{\partial^2(\beta\bar{\kappa})} - (\frac{1}{Z_c(A)}\frac{\partial Z_c(A)}{\partial(\beta\bar{\kappa})})^2) \end{aligned} \tag{137}$$

where

$$
\begin{aligned}
Z_c(A) &= \frac{a}{x} + b\frac{1-x^{G_A}}{1-x} \\
\frac{\partial Z_c(A)}{\partial(\beta\bar{\kappa})} &= 4\pi\beta^{-1}(-\frac{a}{x} + b\frac{-G_A x^{G_A} + (G_A-1)x^{G_A+1} + x}{(1-x)^2}) \\
\frac{\partial^2 Z_c(A)}{\partial(\beta\bar{\kappa})^2} &= (4\pi\beta^{-1})^2(\frac{a}{x} + b\frac{(G_A-1)^2 x^{G_A+2} - (2G_A^2 - 2G_A - 1)x^{G_A+1} + G_A^2 x^{G_A} - x^2 - x}{(1-x)^3})
\end{aligned}
\tag{138}
$$

We have here used the notation $x = \exp(4\pi\beta\bar{\kappa})$, $a = \exp(-8\pi\kappa\beta)$ and $b = \exp(-8\pi\kappa\beta c)$. $< \chi >$ display a sudden, but continous change at $\bar{\kappa} = 0$ from $< \chi > \approx 2$ corresponding to a sphere for $\bar{\kappa} < 0$ to $< \chi > \approx -2G_A$ for $\bar{\kappa} > 0$, for the limit $c = 1$, see Fig. 2a. This transition is governed by strong fluctuations in g, which is manifested by a peak in $\sigma(\chi)(\approx (2\pi G_A)^2)$ at $\bar{\kappa} = 0$. For the limit $c = 2$ the transition occurs at $\bar{\kappa} \approx \frac{1}{G_A} 2\kappa$. This limit approaches the transition point for $c = 1$ ($\bar{\kappa} = 0$) when $G_A \to \infty$.

The solutions of *case 1* are pictured in the phase diagram of fig.16a, where closed surfaces assume a large number of handles. The phase transition line shown in the figure is for $g = 100$ which is small in reality. From this one can conclude that around $\bar{\kappa} = 0$, the system transforms from $g = 0$ to $g = G_A$ (which is a large number). The precise nature of the phase transformation from $g = 0$ to large g and its position is dependent on the exact values of the minima of the Wilmore functional for different g.

In Fig.17a is shown a profile of the genus number g along a line of constant κ of the phase diagram of *case 1*. Both the order parameter $< g >$ and the fluctuation are shown exhibiting a softer transition in the case of $c = 1$ than $c = 2$.

Case 2.

The simplest extension of *case 1* is to consider a collection of immobile closed membranes which can exchange lipids giving rise to a vesicle size distribution. The fundamental degrees of freedom in the partition function is now the number of vesicles and handles and the lipids can redistribute. The partition function takes the form:

$$
Z = \sum_{\{N_A\}} \frac{e^{-\beta\mu\sum_A N_A A}}{\prod_A (A!)^{N_A}} \prod_A Z(A)^{N_A} \tag{139}
$$

where the sum runs over all possible vesicle size distributions $\{N_A\}$. The total amount of lipids in the system $\sum_A N_A A$ is controlled by the chemical potential μ. A has a lower cut-off A_{min} which is determined by molecular details [2]. This is of course a very approximate description where the aqueous solvent has been ignored, ideal exchanged of lipid between the vesicles is assumed and all the effects of molecular interactions are rationalized into the model parametres κ, $\bar{\kappa}$ and μ_o (which then is absorbed into μ). The free energy takes under these conditions the form:

$$
\bar{F}_c = -\beta^{-1} \sum_{A > A_{min}} ln(\sum_{N_A=0}^{\infty} \exp(-\beta(F_c(A) + \beta^{-1} A ln(A) + \mu A)N_A) \tag{140}
$$

where the factorial products in Eq.(14) have been approximated by Stirlings formula. For negative arguments of the exponential function in Eq.(15) $\bar{F}_c$ can be expressed:

$$\bar{F}_c = -\beta^{-1} \sum_{A>A_{min}} ln(\frac{1}{1-\exp(-\beta(F_c(A)+\beta^{-1}Aln(\frac{A}{A_o e}))}) \tag{141}$$

where A_o parametrize the chemical potential through:

$$\mu = \beta^{-1} ln(\frac{A_o}{e}) \tag{142}$$

The vesicle size distribution can in principle be evaluated and will have a peak at the value of A at which the argument takes its minimum. The only size dependence of F_c is through G_A. We will assume that for $G_A \gg 1$ a simple relationship $G_A = \gamma A$ is approximatively valid. We are now in position to derive the phase diagram for *case 2* in the parametres κ and $\bar{\kappa}$, where the most probable vesicle size is used as the orderparameter. In Fig. 16b is shown the various phase transition lines deviding the parametre space into three phases. The midle area II represents a phase with large g and area III a phase with $g = 0$ and a large number of vesicle. The phase transition line dividing these two phases from area I represents a breake down of the evaluation of F in Eq. (16) since F becomes divergent and μ can no longer control the total lipid content in the system. Region I could also contain non-bilayer structures, e.g. lamella and inverse hexagonal structures which cannot be reached but foretold by this theory, that relies on bilayer structure. The profile of g along the line $\kappa = 3.0$, transversing all the areas, is also shown in fig. 17a.

Case 3.

This case represents a collection of closed membranes, which can exchange lipids and are free to move in space. Only the translational degrees of freedom will be taken into account, because handling of the rotational degrees of freedom require detailed information about the stuctures of the Willmore surfaces for all g, which is not avialble. Aggregates of lipid molecules are considered as free particles in plenty of water e.g. excluded volumen effects can be neglected.

Such an ensemble behaves like a Boltzmann gas of particles in contrast to the previous Bose-Einstein distribution in case 2, and the correspondent partition function gets an extra factor $V^{N_A}/N_A!$ due to a volumen factor for every particle that is to be integrated over all translational degrees of freedom (the factor ϕ arises from integration over the momenta $\phi = \phi(A) = (\frac{1}{2\pi k_B T m_A})^{3/2}$ where m_A is the mass of the vesicle with size A, and its A-dependence is written as $\phi = \phi_o(T)A^{-3/2}$):

$$\tilde{Z} = \sum_{N_A} \phi \frac{V^{N_A}}{N_A!} \frac{e^{-\mu \sum_A N_A A}}{\prod_A (A!)^{N_A}} \prod_A Z(A)^{N_A} \tag{143}$$

Similarly the free energy becomes:

$$\tilde{F} = -\beta^{-1} \sum_{A>A_{min}} ln(\sum_{N_A=0}^{\infty} \frac{V^{N_A}}{N_A!} \exp(-\beta(F_c(A)+\beta^{-1}Aln(A)+\mu A)N_A) \tag{144}$$

and with similar approximations as in case 2:

$$F = -\beta^{-1} \sum_{A>A_{min}} exp(-\beta(F_c(A)+\beta^{-1}Aln(A))) \tag{145}$$

with the size distribution of vesicles $< S_A >$ being

$$< S_A >= \frac{< N_A >}{sum_A < N + A >} \tag{146}$$

where

$$< N_A >= V\phi_o exp(-\beta(F_c(A)) + \beta^{-1}(A + 3/2)ln(A) - \beta^{-1}Aln(A_o e))) \tag{147}$$

For fixed volumen and temperature $< N_A >$ and $< S_A >$ are determined by A_o which is related to the total lipid content A_{total} through:

$$\sum_A < N_A > A = A_{total} \tag{148}$$

corresponding to the equation of state $\frac{\partial \tilde{F}}{\partial \mu} = A_{total}$.

We can write $\tilde{F}$ in a more compact form carrying out the sum in Eq. (19):

$$\tilde{F} = -\beta^{-1} \sum_{A>A_{min}} (\phi_o V) exp(-\beta(F_c(A) + \beta^{-1}(A + 3/2)ln(A) + \beta^{-1}Aln(\frac{1}{A_o e}))) \tag{149}$$

The phase diagram arising from this expression is far easier to evaluate than in case 2. It is pictured in Fig. 16c. It looks similar to the one of Fig. 2b but here the full lines representing the border to the forbidden areas are replaced by punctured lines representing just limits of well-defined domains. The area I that in fig. 16b was a forbidden zone is now, either merging with area II representing vesicles with a high genus number or being a phase of small vesicles. The midle area III represent a phase of big vesicles. The transition between these physically different phases is a smooth transition of infinite order and denoted by a dashed line (the only real phase transition) separating the high genus number domains and $g = 0$ domains so actually there excist a distinct phase transition between high genus number domains and $g = 0$ domains. The phase diagram is shown for constant μ. The figure can of course be extrapolated down to $\kappa = 0$. In the regions III and IV the mean size of the vesicles is to a large extend constant but the mean number of vesicles of size A_o will increase when the dotted line is passed from III to IV. This also happens when going from II to I, where the corresponding vesicle size increases. The dashed line in Fig. 2c represents a phase transformation where there is an abrupt change between the phase consisting of vesicles of size A_o and zero g and the phase with high genus g and a larger surface area for each vesicle.

A particle size distribution can be evaluated directly and is also shown in Fig. 17b for a certain value of μ. The distribution is $< A >= exp(\beta^{-1}Aln(A/A_0 e))$, and if approximated to a Gaussian leads to a determination of μ when the number of lipids is assumed constant. The size distribution is shown for different values of $\bar{\kappa}$ and a certain value of $A_{total}/(V\phi_o)$ and κ. The distributions turn out to be very narrow around well-defined values of vesicle sizes. The vesiclesizes are obviously decreasing as $\bar{\kappa}$ decreases in accordance with the observations in ref.10. The values of $\beta\kappa$ and the vesicle sizes are chosen very small compared to the experimental lipid bilayer systems, due to numerical conveniences.

We could now try to unify case 1 and case 2 in a theory that contains an interaction term between the vesicles that can be turned on in going from case 3 to case 2. We define two partition functions Z_1, Z_2. Z_1 is basically of the form of the expression in eq. (14) and contains further an interaction term e^{-EAN_A}. It represents a situation of an aggregate of vesicles (as in case 2), but binding another single free vesicle involving an energy $-EA$. Z_2 represents the situation of case 3 with freely moving vesicles. Hence we write:

$$Z_1 = \sum_{N_A^1} \frac{e^{-\mu \sum_A N_A^1 A}}{\prod_A (A!)^{N_A^1}} \prod_A Z(A)^{N_A^1} e^{EAN_A^1} \tag{150}$$

and

$$Z_2 = \sum_{N_A^2} \phi \frac{V^{N_A^2}}{N_A^2!} \frac{e^{-\mu \sum_A N_A^2 A}}{\prod_A (A!)^{N_A^2}} \prod_A Z(A)^{N_A^2} \tag{151}$$

and in a full description comprising both situations we thus write (under the assumption of weakely interacting systems):

$$Z_{total} = \sum_{N_A^1, N_A^2} \frac{\phi V^{N_A^2}}{N_A^2!} \frac{exp(-\mu(\sum_A N_A^1 + N_A^2))}{\prod_A (A!)^{N_A^1 + N_A^2}} Z_1^{N_A^1} Z_2^{N_A^2} \tag{152}$$

and the total free energy:

$$\begin{aligned} F_{total} &= \beta^{-1} \sum_{A > A_m} exp(-\beta(F_c(A) + \beta^{-1} A ln(A) + \mu A)) \\ &\quad -\beta^{-1} \sum_{A > A_{min}} ln\left(\frac{1}{1 - exp(-\beta(F_c(A) + \beta^{-1} A ln(\frac{A}{A_{oe}}) - EA))}\right) \end{aligned} \tag{153}$$

We shall not try to elaborate more on these expressions but in principle one now has an aggregation parameter or an interaction energy to ones disposal, parametres that can be turned more or less on, bringing case 3 into case 2.

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 $\frac{\partial^n F}{\partial(\beta\bar{\kappa})} \propto (G_A)^n \rightarrow \infty$ for $n \rightarrow \infty$ 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.

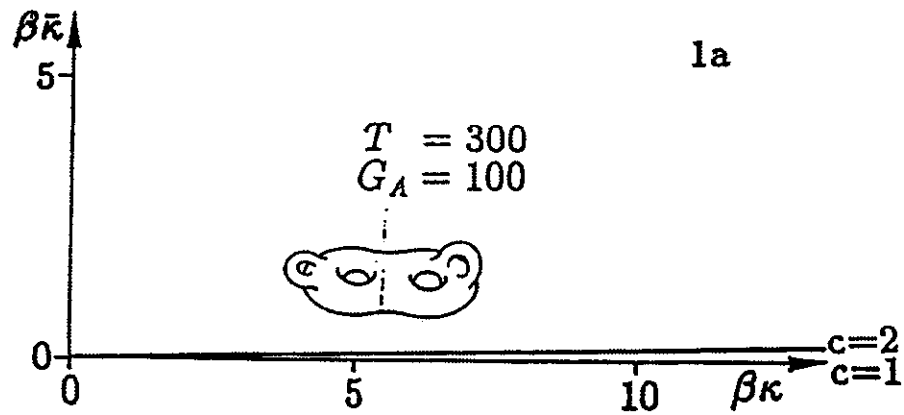
In *case 3* the membranes can move around in space. In this case the description is sensitive in the whole parameter space, i.e. all regions of the parameter space can be structurally determined when the chemical potential μ is kept fixed. There appear phases of vesicles with high genus number for large 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.

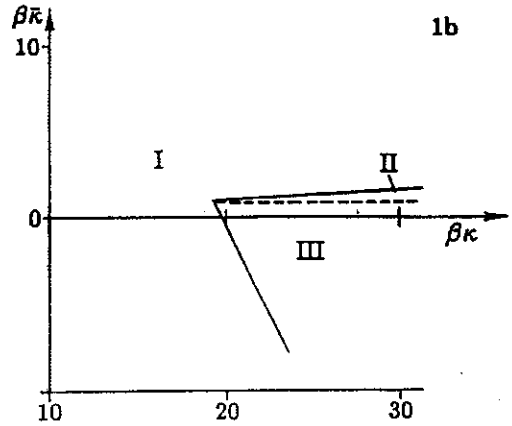
An important result of this study has been the prediction of a topological phase where structures with high genus (many handles) are formed. As we saw, this was due to the fact that the energy would be lowered (at least $2\pi\bar{\kappa}$) by for mig handles. Furthermore the energy distribution (from 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 [28]) 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 [26]. 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 [27], [14] (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.





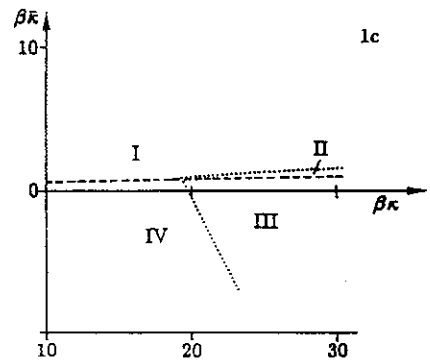


Figure 16a,b,c. Figure 16 shows the phase diagrams in the three cases mentioned in the text.
The phase diagram contains many topological phases as functions of κ and $\bar{\kappa}$.

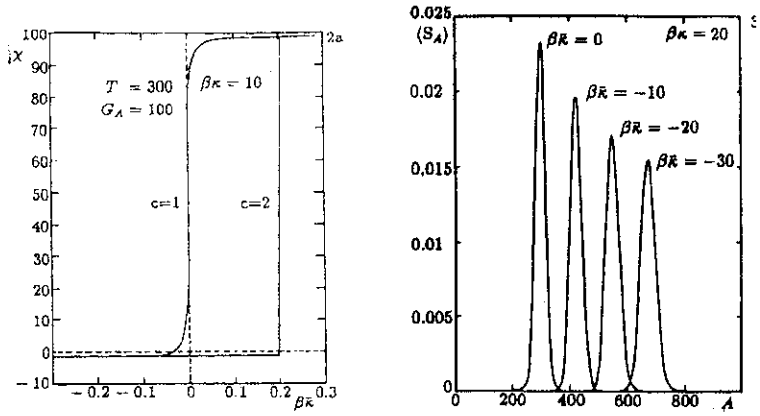


Figure 17a,b.: Figure 17a shows what happens when transversing through genus=0 in the phase diagram.
17b: shows the vesicle distribution at different values of κ.

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 [28] .

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!