CONF-9507246--3

# ISMB-95
## ROBINSON COLLEGE, CAMBRIDGE

Tutorial Programme
Sunday 15 July 1995

TUTORIAL T3

# Computational Tools for Experimental Determination and Theoretical Prediction of Protein Structure

(Sean O'Donoghue & Burkhard Rost)

## DISCLAIMER

EMBL

European Molecular Biology Laboratory
Laboratoire Européen de Biologie Moléculaire
Europäisches Laboratorium für Molekularbiologie

Biocomputing

Postfach
Meyerhofstraße 1
69012 Heidelberg
Germany
Telex: 461613 (embl d)
Fax.: +49-6221-387-517
+49-6221-387-306
Tel.: +49-6221-387-534

Internet: rost@EMBL-Heidelberg.DE
odonoghue@EMBL-Heidelberg.DE

# Computational tools for experimental determination and theoretical prediction of protein structure

Séan O'Donoghue & Burkhard Rost

Tutorial for

## Third International Conference on Intelligent Systems for Molecular Biology

July 16, 1994; Robinson College, Cambridge, U.K.

# Table of Contents

**3. Prediction of protein structure**



**Appendix: Additional material**

## Summary of tutorial

In the first part of the tutorial, we will briefly review what is known about protein structure. Due to advances in sequencing methods, the number of proteins for which the amino acid sequence is know is currently over 40,000 and rapidly increasing. In principle, the tertiary structure of proteins is determined by the amino acid sequence. Currently, the relationship between sequence and structure is unknown: we cannot in general predict structure from sequence. However, from the growing database of experimentally-determined protein structures, some rules are emerging. First: the number of unique protein folds is quite limited. Second: there are many proteins with the same fold, but no similarity of sequence. Third: 'neutral' mutations not altering the protein structure are relatively unlikely. Hence naturally evolved proteins are a record of the unlikely, since most neutral mutations are probably realised. These rules suggest that a key to understanding protein structure lies in the patterns of neutral amino acid exchanges.

Experimentally determining the tertiary structure of a protein is still far more difficult than sequencing; however, the situation has improved greatly in the last few years, and over 2,000 atomic-resolution tertiary structures are now known. Part of this improvement is due to the recent development of computational methods for the determination, and the availability of computers powerful enough to run them. An understanding of the philosophies and assumptions behind these methods is needed in order to assess the accuracy and limitations of experimentally-determined structures. We will briefly cover the basic experimental methodology behind the two main techniques for atomic-resolution structure determination - nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography (XRC). For NMR, structures are calculated from a set of short (<5Å) distances using either distance-geometry (DG) or dynamical simulated annealing (DSA). We will focus on several NMR methods which have also been applied to tertiary structure prediction. For XRC, the initial problem is determining the phase of the reflections in the diffraction pattern. We will discuss briefly several computational approaches: direct methods, maximum entropy, density modification, and molecular replacement. Once the phases are determined, structure refinement is normally done using DSA methods. Due to the rapid pace at which the NMR and XRC computational methods have been developed, most have been proposed based on prototype, single-case studies; there are currently no adequate measures for comparing methods.

How far can theory bridge the growing gap between the data bases of sequence and structure? For a sequence with significant similarity to a protein of known structure, homology modelling can be used to construct a 3D model with correct fold, but inaccurate loop regions. Homology modelling effectively raises the number of 'known' 3D structures to about 10,000. In absence of significant sequence identity, threading techniques can potentially detect remote homologies. For most proteins neither homology modelling nor threading is applicable: the prediction problem has to be simplified. We will discuss generic methods for prediction at three different levels of simplification, namely one, two, and three dimensions. We will emphasis the importance of measuring the accuracy of the methods. Prediction in 1D (secondary structure, solvent accessibility and transmembrane helices) can be improved significantly through the use of evolutionary information. Prediction in 2D (inter-residue contacts, inter-strand contacts, disulphide bonds) can also, to a certain extent, profit from evolutionary information, but so far, is of only limited accuracy. Some progress in 3D prediction has been made: incorrect structures can now be detected with remarkable accuracy (mean-force potentials) and technical improvements and data base growth have made alignments, threading, and homology modelling increasingly powerful.

**Notes about the tutorial (duration, audience, goals, time schedule)**

*Duration:* half day = 4 hours

*Audience:* The tutorial will be addressed to both computer scientists and biologists.

*Goals:* We intend to review the state of the art in the experimental determination of protein 3D structure (focus on nuclear magnetic resonance), and in the theoretical prediction of protein function and of protein structure in 1D, 2D and 3D from sequence (focus on methods that are being applied by biologists).

All the atomic resolution structures determined so far have been derived from either X-ray crystallography (the majority so far) or Nuclear Magnetic Resonance (NMR) Spectroscopy (becoming increasingly more important). We shall briefly describe the physical methods behind both of these techniques; the major computational methods involved will be covered in some detail. We shall highlight parallels and differences between the methods, and also the current limitations. Special emphasis will be given to techniques which have application to ab initio structure prediction.

Large scale sequencing techniques increase the gap between the number of known proteins sequences and that of known protein structures. We shall describe the scope and principles of methods that contribute successfully to closing that gap. Emphasis will be given on the specification of adequate testing procedures to validate such methods.

*Time schedule:*

| | |
|---|---|
| Introduction: proteins the complex machinery of life | 20 min |
| Experimental determination of protein structure | 90 min |
| Prediction of protein structure | 90 min |
| Overview; Evaluation of prediction methods | 10 |
| Prediction of protein structure in 1D | 40 |
| Prediction of protein structure in 2D | 20 |
| Prediction of protein structure in 3D | 20 |

## Notes about the handouts (contents, materials, structure)

*Summaries for tutorial:*

Introduction: proteins the complex machinery of life
Experimental determination of protein structure
Prediction of protein structure

*Materials for handouts:*

Abbreviations used
Sources of Figures
References

Structure of handouts:

For each of the three main parts (Introduction; Determination; Prediction) we shortly summarise the main points touched (pages labelled, e.g., *1S-n* ) and collect all transperencies used (pages labelled, e.g., *1T-n* ). At the end of each summary, we list some of the relevant literature. The appendix (pages labelled *Appendix-n* ) contains some of the abbreviations used, lists titles and sources of all figures and all references.

# Introduction: proteins the complex machinery of life

## Contents



## Summary

The basic principles of protein structures are shortly introduced. Protein structure is determined by sequence. However, there are many proteins which have strong structural similarity, but no similarity of sequence. In other words, structure is more conserved than is sequence. Naturally evolved proteins are a record of the unlikely in that all mutations not altering the structure are probably realised, although the likelihood to find a neutral mutation is small. The patterns of amino acid exchanges not changing structure are highly informative about a given structure. It is commonly assumed that the number of unique protein folds is quite limited.

### What is a protein?

*Building blocks: amino acids.* Proteins are built up from 20 different types of amino acids that are joined by peptide bonds to form a linear chain. The information is coded in the DNA and translated into protein sequences. The basic information about life is coded in a sequence of four different nucleotide bases in the genes. There are two types of nucleic acids: the permanent storage system of the more stable deoxyribonucleic acid (DNA), and the intermediate blueprint tool of the less stable ribonucleic acid (RNA). The information is translated from the genes into the sequences of macromolecules which are involved in every process that keeps life going in an organism, the proteins. Proteins are build up by sequences of amino acids. Known protein sequences contain from some 30 to 10,000 amino acids.

*Formation of peptide bonds.* In general, there are some one hundred different natural amino acids, but only 20 are usually found in proteins. They all have in common the same basic tetragonal structure (Fig. 1.1). The amino acids differ in their side chains (Fig. 1.2). Transcribing the four base alphabet of the RNA on the ribosomes into the 20 letter alphabet of amino acids, the protein sequences are build up residue by residue by joining the amino acids with peptide bonds (Fig. 1.3). The atoms along the line connecting the $C^{\alpha}$ atoms are referred to as the main chain of the protein or as its backbone.

### What determines protein structure?

*Hierarchy of protein structure terminology.* The following hierarchy is often used: primary structure = amino acid sequence; secondary structure = regular patterns of the main chain atoms, like α-helices or β-strands; tertiary structure = the arrangement of all atoms in a protein chain in three dimensions; quaternary structure = the arrangement of all atoms of the whole protein possibly consisting of multiple chains.

Sequence determines structure. A fully unfolded amino acid sequence diluted in the appropriate solvent (under proper conditions in terms of pH value and temperature) folds into a unique tertiary (3D) structure (Anfinsen, et al. 1961, Epstein, et al. 1963, Anfinsen 1973, Anfinsen & Scheraga 1975). The process is reversible (Creighton 1984, Creighton 1991). Consequently, it is assumed that folding is determined exclusively by the information contained in the amino acid sequence (Ewbank & Creighton 1992). Recent experiments suggest that the formation of some secondary structure precedes tertiary organisation (Ewbank 1992). A possible exception to the Anfinsen-hypothesis constitute molecular chaperones, i.e., proteins which assist or hinder folding (Fig. 1.6) (Hubbard & Sander 1991, Hartl, et al. 1994).

*Secondary structure facilitates dense packing.* The main driving force for folding water-soluble globular protein molecules is the need to pack hydrophobic side chains into the interior of the molecule, thus creating a hydrophobic core and a hydrophilic surface. But how can that be realised with the main chain being highly polar (with NH as hydrogen donor and C'=O as hydrogen acceptor, Fig. 1.1)? The simple trick is to neutralise the NH and C'O groups by a formation of hydrogen bonds (Ptitsyn 1992). These bonds effect the formation of the regular patterns of secondary structure like

α-helix (Fig. 1.7) and β-strand (Fig. 1.68). Any region of the protein that is not in either helix or strand will be termed 'loop' in this work (some authors use the term 'random coil' based on the helix-coil model (Zimm & Bragg 1959)). Helices and strands form dipoles (Hol, et al. 1981). The existence of secondary structure elements was first proposed by Pauling and Corey on theoretical grounds prior to their discovery in protein structures (Pauling & Corey 1951, Pauling, et al. 1951, Pauling & Corey 1953a, Pauling & Corey 1953b).

*Function-specific motifs of secondary structure.* Combinations of a few secondary structure segments with a specific geometric arrangement occur frequently in protein structures (3D structure). Such combinations are termed super-secondary structure or motifs. Some of these motifs are associated with particular functions. Examples are the helix-loop-helix DNA binding motif (Gibson, et al. 1993), the calcium binding motif (Fig. 1.9), or the Greek key or β-meander motif (Fig. 1.10) (Hutchinson & Thornton 1993)

*Classification of proteins into structural classes.* Motifs can be used to classify proteins (Richardson 1981, Richardson 1985, Richardson & Richardson 1989, Murzin & Chothia 1992, Orengo, et al. 1993, Wodak & Rooman 1993, Murzin 1994, Murzin, et al. 1995), a more simple classification is based purely on the content of secondary structure (Chothia 1976, Richardson 1981). A protein can be classified as, e.g., all-α, if it contains almost no strand structure and a high content of helix (Fig. 1.11).

**Protein folding: a problem solved only by nature?**

*Variety of protein structures.* Protein structures show a fascinating variety. Structure is more conserved by evolution than sequence. This is mainly explained by the fact that the 3D structure is closely related to the function of the protein. Although the mutation of a few residues in a protein are likely to destabilise the fold (Dao-pin, et al. 1990, Dao-pin, et al. 1991a, Dao-pin, et al. 1991b), evolution has created a record of sequence variation not changing the 3D structure. Two natural protein sequences can differ by 75% of their residues and, yet, have the same 3D structure (Sander & Schneider 1991).

> *"When the first structures of proteins were solved by X-ray crystallography biochemists were struck by the beautiful topologies of their backbone folds and soon researchers in the field became eager to collect structures, and much like zoologists and botanists in past centuries they developed systematic schemes and looked for common features among the various families of folds hoping to unravel the underlying theme responsible for their bizarre structures." (Wu, et al. 1992)*

*Cracking the code.* Solving the protein folding problem means deciphering the code according to which the 3D structure is encrypted in the amino acid sequence. Can we crack the code, i.e., can we unboil the egg (Perutz 1940)? Many researchers successfully fail in doing the neat trick (which is why the issue of predicting protein structure is so interesting...). Prediction methods can be distinguished according to the principle they start from: physics or statistics. The prediction success of methods based on physical principles is still very limited.

*Marginal entropy differences determine protein stability.* What determines protein stability? The hypothesis of Anfinsen is that the folded state of a globular protein is characterised by a minimum in free energy (Anfinsen 1973). The folding transition is largely a two-state process: unfolded non-native chain (U) -> folded native structure (N). As a first approximation, intermediate states can be neglected (though recent exceptions have been found (Ewbank & Creighton 1991, Ewbank, et al. 1995)), and the difference in free energy between unfolded and native state (Δ G) can be approximated by (Lattman & Rose 1993)

$$\Delta G_{U\text{->}N} \propto - RT \ln K$$

with R being the gas constant, T the absolute temperature, and the equilibrium constant K = number of chains in U / number of chains in N. Typical values for ΔG are -5 to -15 kcal/mol (Lattman & Rose 1993).

*Hydrophobic forces drive folding stability.* Why do proteins fold? The driving force for folding has been established to be the reduction of solvent accessible surface (Kauzmann 1959). Folding is driven by the attempt for dense packing (Jaennicke 1987, Stigter, et al. 1991, Pickett & Sternberg 1993). Globular proteins are known to have mean packing densities reminiscent of solids (Lattman & Rose 1993). This density can possibly be explained by the complementarity between interior side chains, fitting together like pieces of a jigsaw puzzle (Fig. 1.12) (Taylor 1992).

*Dense packing determines the conformational specificity?* What determines the specific conformation of a fold? One explanation could again be the density of packing, i.e. only very specific conformations allow the residues to pack into the jigsaw puzzle. However, there is evidence that such a grouping can readily be done, i.e. does not require one particular conformation (Lattman & Rose 1993). This suggests that dense packing is not the primary source of conformational specificity. What then determines the fold? One attractive candidate is the stereo-chemical code: "It

is plausible that conformational specificity is imposed through a redundant stereo-chemical code that arises from the interplay between the shape and polarity of residue side chains and secondary structure conformation." (Lattman & Rose 1993).

**Evolution creates a record of the unlikely!**

*A single mutation can destabilise a protein.* The mutation of a single residue typically causes an approximate reduction of the free energy difference between native and unfolded state of about 1 kcal/mol (Lattman & Rose 1993). Thus, the exchange of a few residues can already destabilise a protein of more than 100 residues (Dao-pin, et al. 1990, Dao-pin, et al. 1991a, Dao-pin, et al. 1991b, Zabin, et al. 1991). Does this imply that two proteins with some different residues have a different 3D structure? And if, are all potential 3D structures realised in nature, i.e. are there some $20^N$ different folds for proteins with N residues realised in nature? The fact that a single mutation can destabilise a protein implies only that the majority of the $20^N$ possible sequences adopt different structures. But, has evolution created such an immense variety?

*Only mutations not altering the structure survive.* Random errors in the DNA lead to the wrong translation of the information coded in the genes into sequences of amino acids. These errors are the basis for evolution (Darwin 1859, Monod 1970). Are all such errors carved into fossils, or do only the fittest survive? The function is determined by the structure and the environment of the protein. Mutations resulting in a structural change are not likely, since the protein cannot perform its task. Thus, only those errors are likely to be accepted which do not alter the structure. Of course, this is only one side of the coin, would it not be possible to accept changes of the structure and consequently of the function, there would not be much room for evolution. Indeed, one of the evolving pictures is that proteins consists of functional modules, which are combined in various ways to yield different properties for the proteins (Bork 1992, Bork, et al. 1992c, Doolittle & Bork 1993, Green, et al. 1993).

*How much variation in sequence is possible?* Mutations of amino acids survive if they do not change the 3D structure of the folded protein. The known proteins are a record of exploration for variation of sequence with no effect to structure. Structure is more conserved than sequence (Chothia & Lesk 1986). But, how much variation of the sequence can exactly be accepted without changing the structure? The surprising result is quite some (Fig. 1.13). Evolution has realised pairs of proteins which have the same 3D structure, although they have only 25 of their 100 residues alike. Of course not any two residues can be exchanged anywhere in the sequence. Instead, the possible exchanges depend on the details of the structure and on the physico-chemical properties of the amino acids involved. Thus, the pattern of residue substitution - the record of the unlikely - carries information rather specific for a particular protein structure (Zuckerkandl & Pauling 1965).

**How many different protein folds exist?**

Speculations are that the number of different protein folds realised by nature is fairly limited (Chothia 1992, Finkelstein & Reva 1992, Finkelstein, et al. 1993). However, the concept of 'similarity' between folds is not clear-cut (Sippl 1982). The number of unique chains is > 300 (Hobohm & Sander 1994). Based on this number and recent analyses of entire chromosomes (Bork, et al. 1992b, Bork, et al. 1992a, Bork, et al. 1994) the estimate for the number of folds appears to confirm the notion of 1,000 folds (factor of 3 possible).

*Literature on protein structure*

*Introductions to protein structure and folding (books):* (Schulz & Schirmer 1979, Fasman 1989b, Brändén & Tooze 1991, Lesk 1991, Rees, et al. 1992)

*Introductions to protein structure and folding (reviews):* (Richardson 1985, van Gunsteren 1988, Fasman 1989a, Richardson & Richardson 1989, Brünger & Nilges 1993, Dill 1993, van Gunsteren 1993, Murzin 1994)

# Computational tools for experimental determination and theoretical prediction of protein structure

- Introduction: proteins the complex machinery of life
- Experimental determination of protein structure
- Prediction of protein structure



# Introduction: proteins the complex machinery of life

- What is a protein?
- What determines protein structure?
- Protein folding: a problem solved only by nature?
- Evolution creates a record of the unlikely!
- How many different protein folds exist?

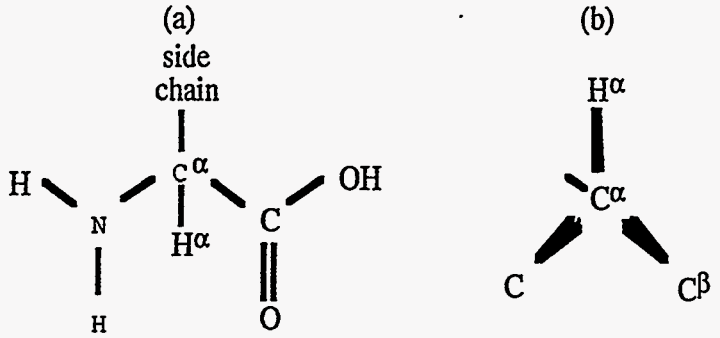
# What is a protein?

- Proteins are the machinery of life
  - Rosetta stone


  - 30 - 10,000 amino acids
  - alphabet = 20 letters of amino acids
  - common: basic tetrahedron
    - » Fig 1.1
    - » Fig 1.2
  - biosynthesis of amino acids into polypeptides
    - » Fig 1.3
  - flexibility of chain: the dihedral angles
    - » Fig 1.4

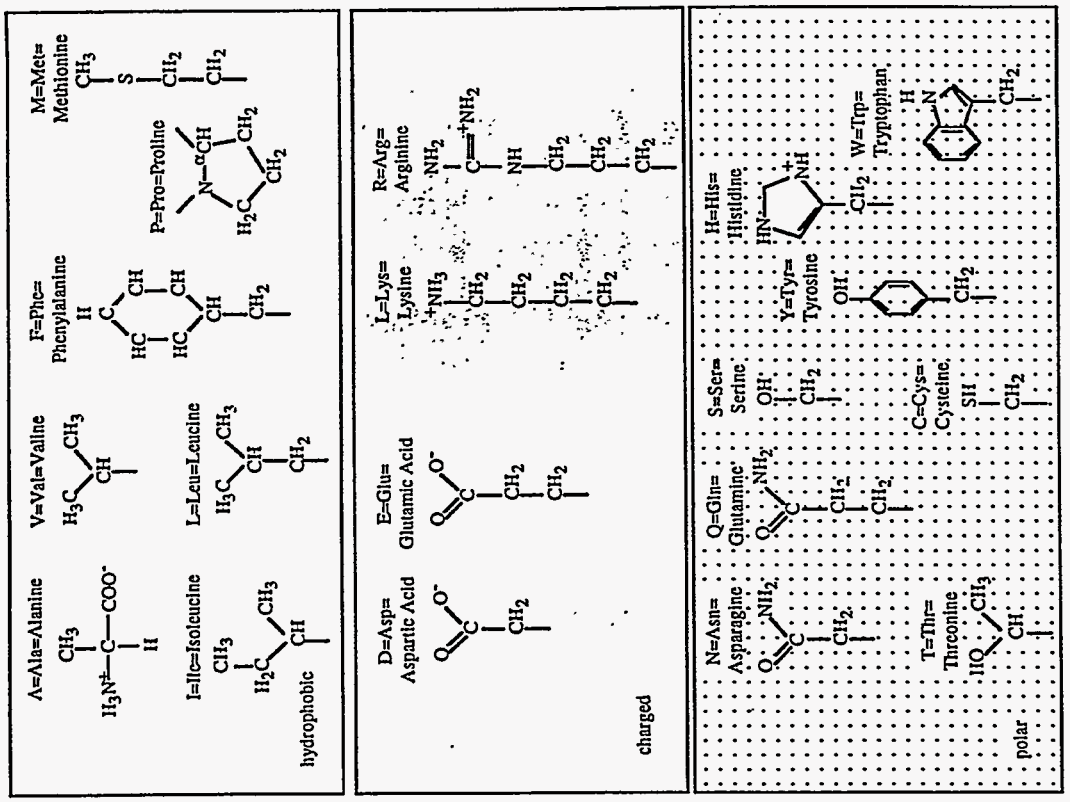


**Fig. 1.1: Basic tetrahedron of all amino acids**




(a) The atoms around the $C^{\alpha}$ atom all amino acids have in common. The convention is to label the carbon atoms in the side chains with Greek letters starting from the central $C^{\alpha}$ (IUPAC-IUB, 1970). (b) In nature generally only the left-handed L-configuration of an amino acid is found. The reason for the symmetry breaking is eventually a random initial event (Schulz & Schirmer, 1979).

**Figure 1.2: The 20 amino acids**

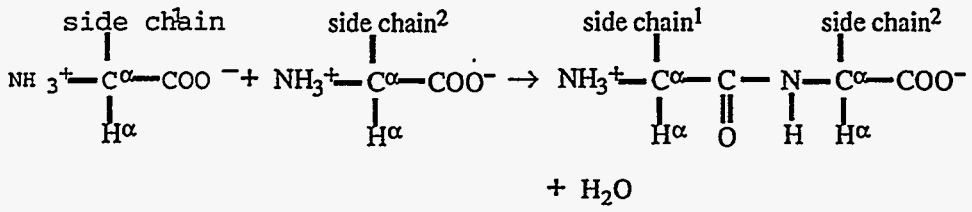



Only for Alanine are the central $C^{\alpha}$, and the amino group ($NH_2$), the carboxyl group (COOH) shown. For the other 18 amino acid only the side chains are given. The 20th amino acid is G=Gly=Glycine which has only a hydrogen as side chain. The amino acids can be grouped according to the physico-chemical properties as shown above.

Figure taken from (Rost, 1993)



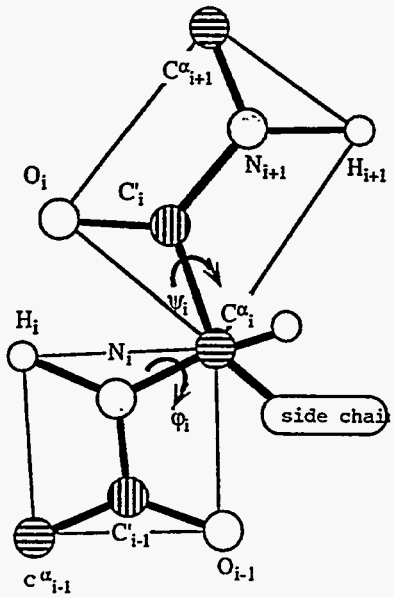
**Fig. 1.3: Biosynthesis of amino acids to polypeptides**




Amino acids are joined end-to-end during protein synthesis by the formation of peptide bonds. According to the characteristic ends, the first residue of the left hand side in a protein is termed the N-terminal end, the right hand side as the C-terminal end.

Figure taken from (Rost, 1993)

Fig. 1.4: The dihedral angles




The peptide bond (CO-NH) has a partial double bond character. As a consequence, the surrounding 4 atoms lie in a plan (indicated by quadrangles). Rotation along the polypeptide chain is possible around the angles $\varphi$ and $\psi$ on both sides of the $C^{\alpha}$ atoms.

Figure taken from (Rost, 1993)



# What determines protein structure?

- Hierarchy:
  - » primary structure — amino acid sequence
  - » secondary structure — e.g.: α-helix; β-strand;
  - » tertiary structue — arrangement in 3D
  - » quaternary structure — grouping protein chains

- 3D structure is determined uniquely by sequence
  - – unfolded sequence folds into unique 3D structure
    (Epstein et al., 1963, Anfinsen & Scheraga, 1975)
  - – folding reversible
    (Creighton, 1992)
  - – thus, information contained in sequence
    (Ewbank & Creighton, 1992)
  - – formation of secondary structure first
    (Ewbank, 1992)
    - » Fig. 1.5

- Exception: chaperones

*Fig. 1.5: Simplified view of protein folding*

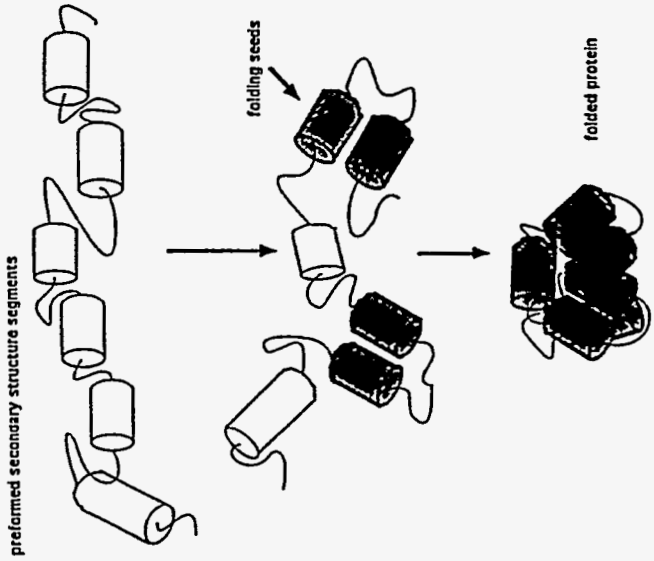



**Figure 4:**
**Simplified view of protein folding**
The process of protein folding takes about 1 second, or about $10^{12}$ microscopic molecular events. In this simplified view, first helical segments about 10-15 polymers units in length form locally. Later, higher structures evolve at certain 'seed' positions. In the final, cooperative, stage, the protein interior condenses and the internal packing of molecular groups is optimized. This process is much too complicated to be fully simulated by molecular dynamics for realistic times, even on a Teraflop machine. The problem of folding a protein correctly on a computer is one of the major unsolved problems of molecular biology [7].

Figure taken from (Sander et al., 1992)



# Chaperones

- "the art of avoiding sticky situations"
  (Hartl et al., 1994)
  - prevent aggregation of newly synthesized polypeptides
    - » Hsp70 (heat shock protein 70kDa)
  - folding to the native state
    - » Hsp60 / GroEL
    - » Fig. 1.6

  - further literature:
    (Hubbard & Sander, 1991, Saibil & Wood, 1993, Hartl et al., 1994)

*Fig. 1.6: Chaperone mediated protein folding*

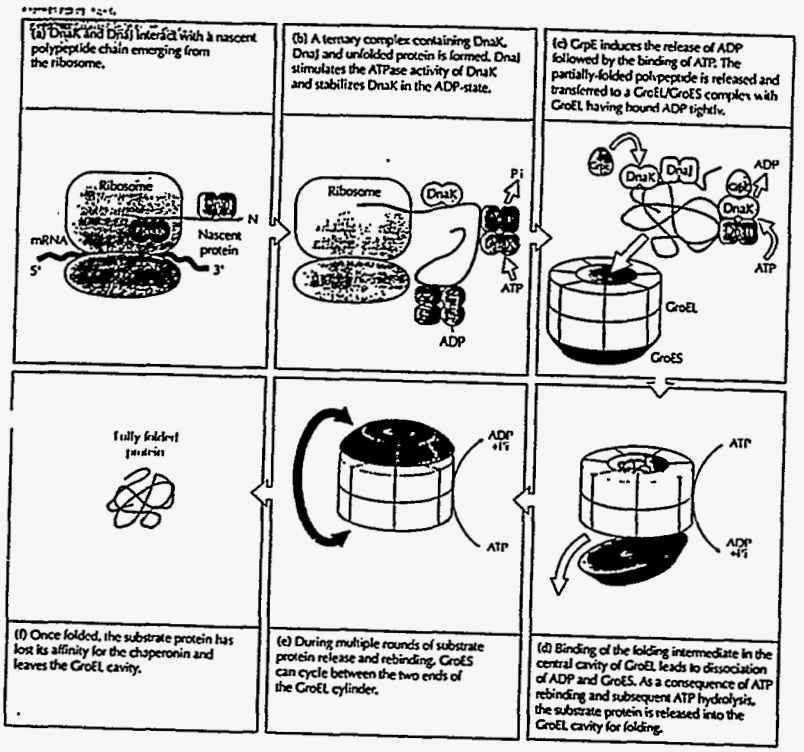



Figure taken from (Martin & Hartl, 1993)



## Secondary structure facilitates dense packing

- driving force for folding globular water-soluble proteins: hydrophibic side chains into the interior => hydrophobic core, hydrophilic surface

  (Kauzmann, 1959, Lesk, 1991, Creighton, 1992, Lattman & Rose, 1993)
- but, main chain highly polar (NH donor, C'=O acceptor)
- trick: neutralise polarity by forming hydrogen bonds

  (Ptitsyn, 1992)

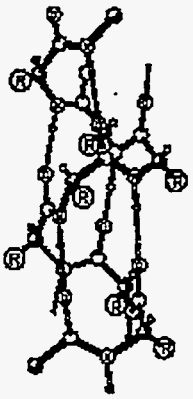
  » Fig. 1.7/1.8
- H: α-helices and E: β-strands form dipoles

  (Hol et al., 1981)
- third class: termed L: loop (often called: random coild, based on helix-coil model)

  (Zimm & Bragg, 1959)
- secondary structure formation proposed before first X-ray structures were solved

  (Pauling & Corey, 1951, Pauling et al., 1951, Pauling & Corey, 1953a, Pauling & Corey, 1953b)

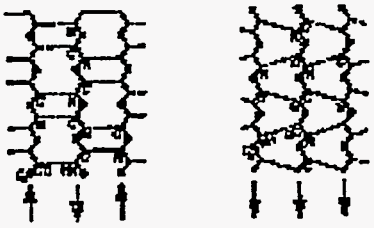
**Fig. 1.7: Hydrogen bond pattern of helix**



Helices of polypeptide chains with internal hydrogen bonds (dashed). Hydrogen bonds are between the amide and carbonyl groups of residues i and i+3. Thus one helix turn covers 3.6 residues extending over some 1.5Å per residue. The side chains point outwards (circles marked with R). Figure taken from Schulz & Schirmer (1979).



**Fig.. 1.8: Hydrogen bond patterns of strand**



Hydrogen bonds are indicated by dashed lines and chain directions by arrows. $C^{\alpha}$ are marked by dots. (a) antiparallel three-stranded β-sheet, (b) parallel three-stranded β-sheet. The side chains point alternatively above and below the sheet. The distance between two neighbouring strands is about 5Å. It has been noted recently that often the larger of the two holes formed in an antiparallel sheet (a) cannot be filled by side chains, thus, effecting a majority of the defects of close packing in protein globules (Finkelstein & Nakamura, 1993). Figure taken from Schulz & Schirmer (1979).

# Secondary structue patterns form function-specific motifs: HTM

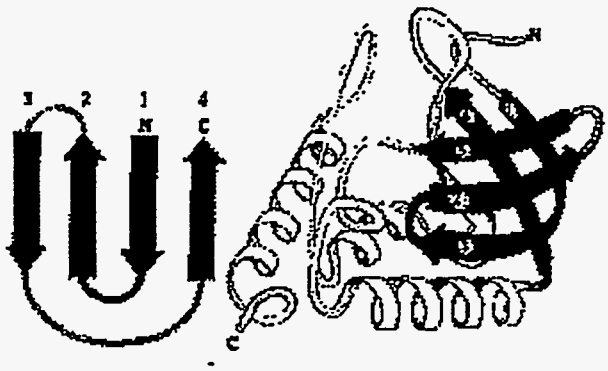
**Fig. 1.9: Calcium binding motif: helix-loop-helix**


In ribbon diagrams, helices are usually drawn as spirals or cylinders, and strands as arrows. Calcium binding motif: the two helices (from the muscle protein parvalbumin) give the scaffold for binding and releasing the calcium ligand (shown as a sphere). Figure taken from Brändén & Tooze, (1991).



# Secondary structue patterns form function-specific motifs: greek-key

**Fig. 1.10: Greek-key motif: four strands**




Greek key (or β-meander) motif: four adjacent antiparallel β-strands are arranged in a pattern similar to ornamental patterns used in ancient Greece. The structure is that of staphylococcus nuclease, an enzyme that degrades DNA. Figure taken from Brändén & Tooze, (1991).

# Classification of proteins into structural classes

- **classification based on motifs**

  (Richardson, 1981, Richardson, 1985, Richardson & Richardson, 1989, Johnson, 1991, Murzin & Chothia, 1992, Orengo et al., 1993, Wodak & Rooman, 1993)

- **classification based on structural alignments and domains**

  (Holm et al., 1993, Holm & Sander, 1993, Holm & Sander, 1994a, Holm & Sander, 1994b)

- **classification based on content in secondary structure**

  (Chothia, 1976, Richardson, 1981, Zhang & Chou, 1992)

  - all-α: $\% \alpha \geq 45\%$ ; $\% \beta < 5\%$
  - all-β: $\% \alpha < 5\%$ ; $\% \beta \geq 45\%$
  - a/b: $\% \alpha \geq 30\%$ ; $\% \beta \geq 20\%$
  - rest

» Fig. 1.11



*Fig. 1.11: Percentage helix vs. percentage strand in known 3D structures*

**Figure 3: Content of helix vs. content of strand**

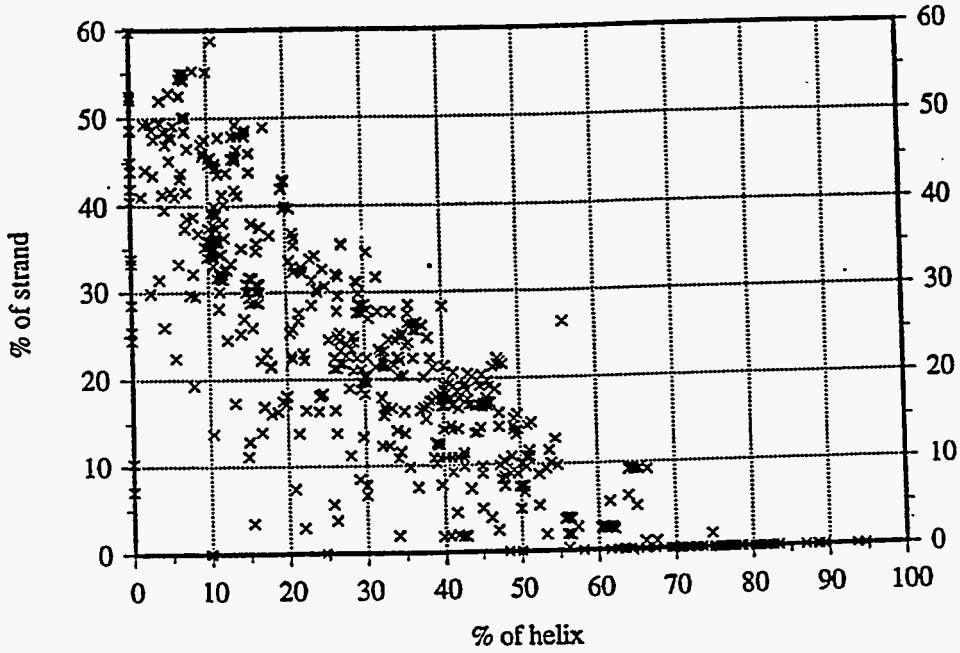



Figure taken from (Rost & Sander, 1994b)

# Protein folding: a problem solved only by nature?

- What drives folding?
- What determines conformational specificity?



# What drives folding?

- folding largely two state transition:
  - unfolded nonnative U -> folded native N
  - free energy:

$$\Delta f_{U\text{->}N} \propto -RT \ln K$$

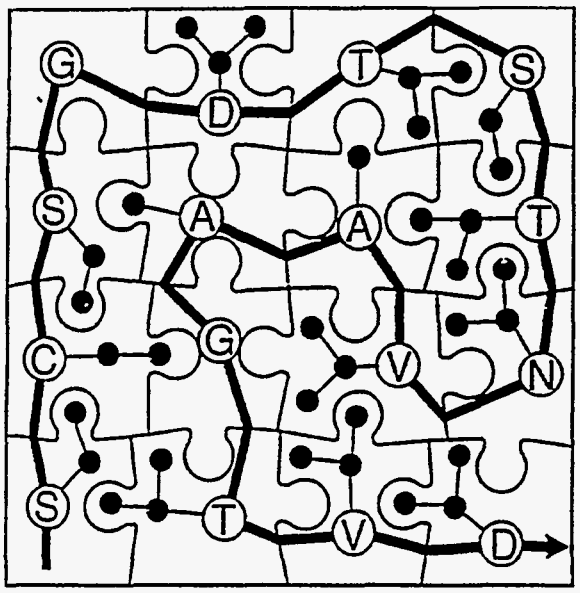
» R, gas constant; T, absolute temperature; K, equilibrium constant = #U/#N

» typical values -5 to -15 kcal/mol (Lattman & Rose, 1993)

- packing densities reminiscent of solids (Jaennicke, 1987, Lattman & Rose, 1993, Pickett & Sternberg, 1993)
- possible explanation for density: jigsaw puzzle (Taylor, 1992)

» Fig. 1.12

*Fig. 1.12 Protein jigsaw puzzle*




The protein jigsaw puzzle. At first sight the solution is easy because there is a known backbone structure (green) to copy. But packing the side-chains (small red and black circles) is difficult, because for each piece there are a number of alternatives (rotamers) only one of which will appear in the completed picture at any position. The approach of Desmet *et al.* can be explained, in simplified terms, by considering the options for the residue (C) at the second position. If there are three rotamers for C and two rotamers for S, then each C is tried with each S at the first and third positions. If there is a rotamer of C that will not fit with any S at either adjacent position (or with G at the thirteenth position), then that piece cannot be part of the final picture and can be thrown away. This test is applied to all positions, so reducing the number of pieces that need to be considered when it comes to the final (combinatorial) assembly stage.

Figure taken from (Taylor, 1992)



## What determines conformational specificity?

- **one candidate: density of packing:**

  only very specific conformations fit into the jigsaw

- **another: stereochemical code:** $\Delta f_{U\text{->}N} \propto -RT \ln K$

  "It is plausible that conformational specificity is imposed through a redundant stereochemical code that arises from the interplay between the shape and polarity of residue side chains and secondary structure conformation."

  (Lattmann & Rose, 1993)

- **Can we model protein folding?**

# Variety of protein structures

"When the first structures of proteins were solved by X-ray crystallography biochemists were struck by the beautiful topologies of their backbone folds and soon researchers in the field became eager to collect structures, and much like zoologists and botanists in past centuries they developed systematic schemes and looked for common features among the various families of folds hoping to unravel the underlying theme responsible for their bizarre structures."

(Sippl et al., 1994)

- **first structures: myoglobin and hemoglobin (oxygen binding)**

  (Kendrew et al., 1960, Perutz et al., 1960)

- **today more than 2,000 structures known**

  (Berstein et al., 1977; Abola et al., 1988)



- Myohemerythrin (2mhr)
  four helix bundle with 118 residues. The molecule binds oxygen in muscle cells (source: sipunculan worm). The helices are shown as spirals, the loop regions as thin lines. The fifth helix is a $3_{10}$ helix, spanning only over three residues. The other helices extend over 16-24 residues.
- Myoglobin (1mba)
  seven helix bundle with 146 residues. This molecule was one of the two first experimentally solved structures (Kendrew et al., 1960). It is used for oxygen storage (source sea hare). The oxygen is stored in form of the heme group shown in the centre (green with blue centre). The heme is enclosed by the seven helices (shown as cylindres) like in a pocket. The helices span over 5-16 residues. The $3_{10}$ helix shown on the left hand side (red, above the heme) spans over 6 residues.
- Bence-Jones immunoglobulin (1bjl)
  dimer (two distinct chains) with 247 residues (source: human). The strands are shown as arrows with the head pointing towards the end of the protein (C-terminal end). The hydrogen bonding partners of the residues in a strand are those at the strand nearest by (bonds not shown). The antiparallel β-sheets extend over 2-10 residues. Immunoglobulins act as antigen receptors on the surface of B cells in the immune system. All immunoglobulin domains have similar 3D structure.
- Satellite tobacco necrosis virus coat protein (2stv)
  dominantly an α / β structure with 184 residues. The virus RNA is embedded in the pockets formed by the β-sheets. The structure is typical for most virus coat proteins.
- Flavodoxin (4fxn)
  mixture of helices and strands with 138 residues (source: clostridium MP). It is involved in electron transport (flavin mononucleotide-binding redox protein). The binding of the ion is illustrated by the aromats at the right hand side (green). The β-sheets form a pocket in the core of the protein, whereas the helices lie on the surface. The ion is bound on the loop regions at the N-terminal ends (arrowheads) of the parrallel β-strands.
- TIM barrel triose phosphate isomerase (6tim)
  barrell structure with 249 residues (source: trypanosoma Brucei). It functions as an enzyme to transform ATP (adenosine tri phosphate) into ADP (adenosine di phosphate). The molecule is built up from four β–α–β–α motifs that are consecutive both in sequence and structure. The motifs are arranged such that in the centre a barrel is formed.

## Unboiling the egg?

- Boiling an egg implies unfolding proteins
  (Perutz, 1940; Perutz, 1980)
- Can this procedure be reversed in theory?
  Can the Rosetta stone of protein folding be decrypted?
- Two schools: statitics and physics
- Obstacles to modelling from first principles:
  - marginal free energy difference between folded and unfolded state
  - one residue exchange can destabilise a structure
  - given complexity -> too much CPU-time
  - inaccuracy in knowledge of physical constants, i.e., potentials
- How far do we come with today's molecular dynamics?
  - refining structures
  - modelling the interactions between protein and ligands
  - modelling of short (some residues) loop regions
    (Abagyan & Totrov, 1994, Abagyan et al., 1994)
- BUT not distinction: native fold and grossly misfolded structure
  (Novotny et al., 1984, Novotny et al., 1988)

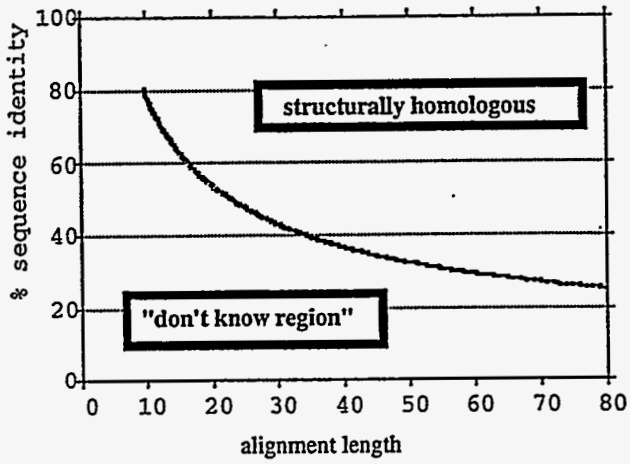


## Evolution creates a record of the unlikely!

- A single mutation can destabilise a protein
  - free energy marginally different (≈ 1kcal/mol) (Lattman & Rose, 1993)
  - thus single residue exchange can destabilise (Zabin et al., 1991)
  - => $20^N$ different folds of proteins with N residues?
  - in principle yes, but are they all realised?
- Only mutations not altering structure survive
  - random errors in DNA basis for evolution (Darwin, 1859; Monod, 1970)
  - all errors carved into fossils of protein structures?
  - no, function has to be maintained -> structure
  - more complex: evolution by shuffling domains or modules
    (Bork, 1992, Doolittle & Bork, 1993, Green et al., 1993)
- How much variation in sequence is possible?
  - structure evolutionarily more conserved than sequence
    (Chothia & Lesk, 1986, Schneider & Sander, 1991)
  - 75% of the sequence can be exchanged without changing the structure
    (Sander & Schneider, 1991)

» Fig. 1.13

**Figure 1.13: Relationship between structural homology and sequence identity**




For about 1,000 pairs of fragments from proteins with known 3D structure, alignments are made. The percentage of identical sequences in this fragment is plotted versus the length of the fragment (alignment length). The homology threshold divides the graph into a region in which all pairs are structurally homologous (root mean square deviation of backbone < 2.5Å), and a region where homology is unlikely ("don't know region"), i.e. where some fragment pairs are structurally similar and some are not. Figure kindly provided by Reinhard Schneider.



# How many different protein folds exist?

- **only 1000 folds?**

  (Chothia, 1992)

- **how similar are similar folds?**
- **root-mean square deviation of backbone:**

$$D(S, S') = \min_{k} \left\{ \frac{1}{N} \sum_{i}^{N} (r_i - r_i')^2 \right\}^{\frac{1}{2}}$$

where $r_i$ is the vector pointing to residue i of structure S, N the number of residues, and the minimum is taken over all k possible orientations, i.e. the optimal solution. A reasonable cut-off to regard two structures as homologue is $D \leq 3$ Å.

(Sippl, 1982)

- **how many folds = where setting the cut-off**

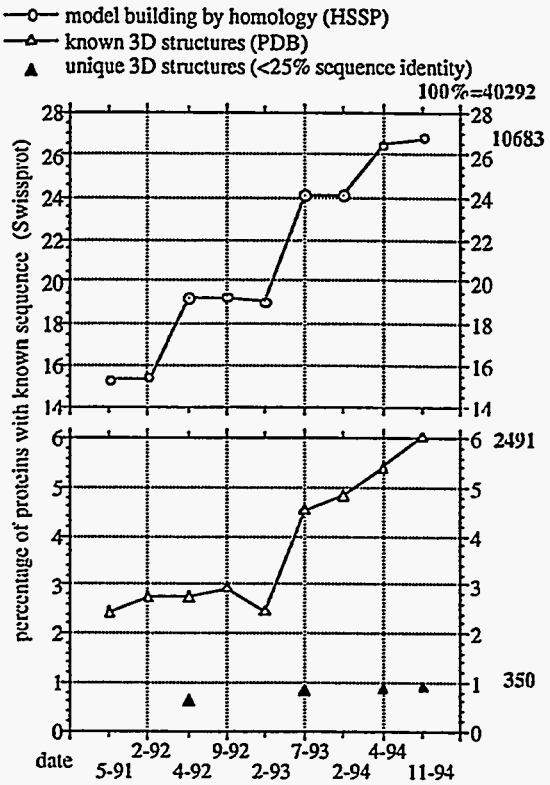





*Fig. 1.15 Number of unique protein folds*

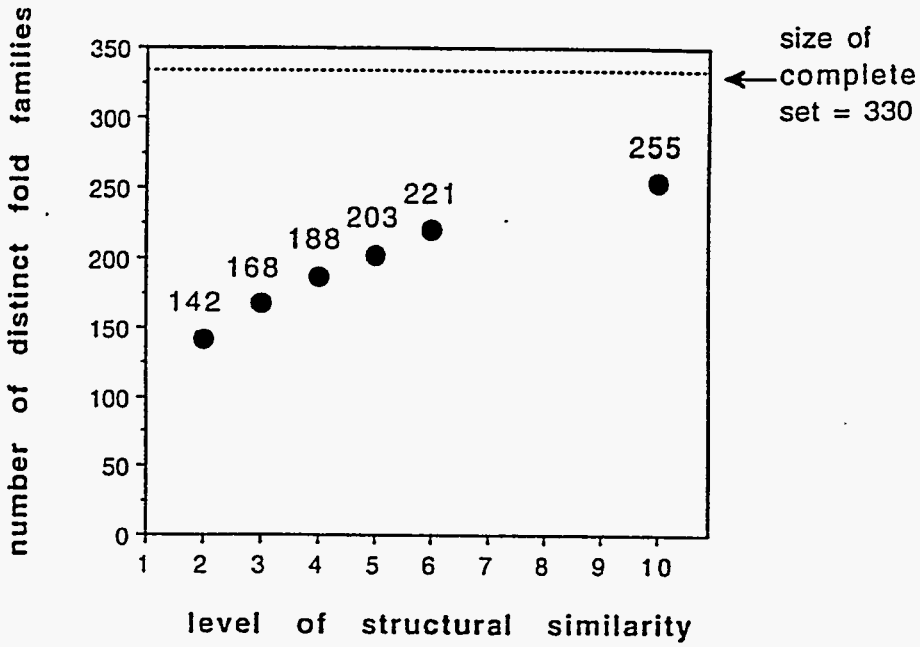



Figure taken from (Holm & Sander, 1994a)

# The number of folds is limited

- Given a certain cut-off:
  How many folds exist?
- Again: not as simple
  - secondary structure level? => 3
  - level of all residues? => currently >2,000
  - level of domains? => currently > 400

  (Holm & Sander, 1994b)
- Thus: how many?

| | |
|---|---|
| human sequences | 100,000 |
| now implicitly known 3D | 10,000 |
| unique( <25% pairwise seq. ident.ity) | < 400 |
| unique (most stringent cut-off) | < 150 |
| => unique human folds | < 1,500 ?? |

# Calculating protein structures from experimental data

**Synopsis of talk**



## Summary


The experimental determination of protein structure is blooming. Part of the reason is the recent development of computational methods for the determination, and the availability of computers powerful enough to run them. In spite of the fundamental role of this methods in determining the accuracy of the protein structure, none have been rigorously evaluated.

We will briefly cover the basic experimental methodology behind the two main techniques for atomic-resolution structure determination - nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography (XRC). In this tutorial, the NMR methods will be emphasised. For NMR, structures are calculated from a set of short (<5Å) distances using either distance-geometry (DG) or dynamical simulated annealing (DSA). These initial structures are then refined with a number of methods: currently, there is no consensus on which methods are best.

For XRC, the central problem is determining the phase of the reflections in the diffraction pattern. We discuss briefly several computational approaches: direct methods, maximum entropy, density modification, and molecular replacement.


## Introduction

*The state of the art.* Look in any recent volume of *Nature* or *Science* and you are guaranteed to find at least one important protein structure which has recently been solved at atomic-resolution. For more structures, you could glance at the newly created *Nature Structural Biology*, designed to handle the overflow of structures from *Nature*, or several other newly created journals loaded with protein structures e.g. *Proteins*, *Protein Engineering*, *Protein Science*, or *Structure*. Driven by advances in molecular biology, data acquisition, and computer power, the experimental determination of protein structures is blooming.

*What is the rle of computational methods?* Which computational method is used is of fundamental importance to the accuracy and precision of structures obtained, or even to the success or failure of the determination. Recent advances in these methods have increased the scope of structures which can be determined. However, due to rapid development, new methods have been introduced based only on prototype, single-case studies; there is currently no adequate measures for comparing methods.

*What are the major techniques? Which are the limitations?* For any protein given we want to know the structure of, we have two major approaches. If the protein less than about 250 residues, we can use NMR spectroscopy to examine the solution structure - this will almost always work, and the process takes from two months to two years, sometimes forever. The NMR technique is still emerging - in the near future, the current limitations of protein size and speed of determination will improve; also, NMR will yield more detailed information on the dynamics of proteins in solution. Other breakthroughs are likely, although the direction is less clear. The other approach to structure determination, with no theoretical limit on the protein size, is to try to convince the protein to crystallise. The crystals must be large and well-ordered. This is a question of luck and patience, but the success rate is nowadays very high. However, not every protein structure can yet be solved. Most of the extremely large proteins are simply too irregular to form adequate crystals; here, we have to be content with breaking the protein into smaller domains and solve the structures of them. Some classes of proteins, for example membrane proteins, still present a challenge that we have no established method for dealing with, although even here we are making progress.

*How to access experimental protein structures?* For convenient browsing through all protein structures, it is worth looking at the *Macromolecular Structures* series published annually by Current Biology press which covers all proteins solved in the last year . The central public computer database is the Brookhaven protein databank - PDB (http://www.pdb.bnl.gov/). Currently, there are over 3000 structures; of these, 400 have homologies less than 25%; about 150 are unique folds. About 20% of the structures in the database are determined by NMR; this proportion of NMR structures is increasing.

The current format of a PDB entry is widely recognised as outdated; note the card number at the last line - helpful if your stack of paper cards (one per line) falls on the floor! There are efforts to come up with a new format, but they are not expected to come to fruition for at least two years more.

## Structures of proteins in solution: NMR spectroscopy

### Basic experimental methodology

*Sample preparation.* There are a few important requirements for studying a protein by NMR spectroscopy. One is that the sample must be sufficiently concentrated (around 0.5mg/ml or more); for many proteins, this is actually close to physiological concentration, however *in vitro* aggregation is often a problem at these concentrations. This must be prevented as it will raise the effective molecular weight of each molecule above what can be studied by NMR. Another requirement is the production of $^{2}$H-, $^{13}$C-, or $^{15}$N - labelled protein samples - larger proteins, labelling is essential. Fortunately, this is not so difficult with modern cloning techniques.

*Collecting spectra.* Not a trivial step - first, the NMR spectrometer: it costs at least $US1/3 million - requires special housing in a building without iron; the liquid nitrogen cooling the superconductor coils must be renewed weekly; most metals must be kept at least 3-4 metres distant from the spectrometer. Such a valuable instrument is usually purchased for the use of several (or many) research groups - unfortunately, to collect adequate (3D and 4D) spectra to determine a large protein can take one, two, or more weeks of uninterrupted measurement time - this means fighting with other users for exclusive access.

Then comes the measurement techniques, or pulse sequences, which determine the type of spectra obtained; NMR is a rapidly evolving technique - there is no standard set of techniques; instead, there are many standard techniques, all constantly being improved by different groups. Different proteins require different pulse sequences to be used, depending on the size, type of labelled compound that can be made, etc. In addition to a background in biochemistry or molecular biology, the protein NMR spectrometrist must have a good grasp of the mathematics and physics behind biological NMR.

*1D NMR spectra.* The sample is placed in a high intensity magnetic field (7 Tesla or more) - nuclei with a net magnetic moment tend to align with the field creating a macroscopic magnetisation pointing in the same direction as the spectrometer's field (up, or usually along the $z$ axis).

In continuous-wave (CW) spectrometry, we apply radio-frequency waves, and slowly scan through a range of frequencies. Different nuclei in the sample have different resonant frequencies; when the frequency scan passes through a resonant frequency, a small absorption peak can be detected.

CW spectrometry, although conceptually simpler, has been superceeded by pulse spectroscopy pioneered by R. Ernst (for an excellent review, see his Nobel lecture: Ernst, 1994). Here, the radio frequency is applied as a pulse, rather than a continuous wave. The pulse is timed to rotate the magnetisation 90°, from the $z$ to the $x$ axis. The pulse is then stopped, and the magnetisation precesses around the $z$-axis, slowly decaying towards it. This precession causes a (much weaker) radio frequency electromagnetic signal which is detected. The different frequencies corresponding to the different nuclei all contribute to the signal. By calculating the Fourier transform (FT) of the signal, one can then construct the same spectrum measured by the CW method, however the entire spectrum is obtained at once. The main advantage is a tremendous gain in signal-to-noise.

*Spectra of two or more dimensions. The NOE spectrum.* The 1D NMR spectra of proteins is not in itself very useful; it is simply too crowded, since the dispersion or frequency differences between nuclei is often smaller than the line width. But to calculate structures from these spectra, the first thing we must do is to assign each peak with a specific atom in the protein (in the future, this may not always be necessary - see below). Normally we do proton NMR - we see one or more peaks for each proton - some protons are easy to assign. For example, if we have only one tyrosine in our protein, the protons of the tyrosine will have a clearly distinguishable chemical shift (frequency) due to the effect of the ring-current magnetic field. What we then need to do is to connect these assigned protons to neighbouring protons. This is the principle behind 2D spectra.

The most important 2D spectra are the nuclear Overhauser effect (NOE) spectra: it consists of cross-peaks which we can normally assign as arising from pairs of protons (*a* and *b*). The volumes of the cross-peaks, $V_n$, can be related to the distances between the protons to a first approximation (Macura & Ernst, 1980):

$$V_n = c\, d\,(a_n, b_n)^{-6} \qquad (1)$$

where *c* is a constant determined once for each spectrum. Thus for each volume we can assign in the NOE spectra, we obtain a distance restraint $D_n$ - the set of all such restraints for a given protein is denoted *D*. Due to the inverse-sixth power of this relationship, only small distances (<5Å) can be detected. Thus the calculation problem of NMR is to find the structure given only *D*.

*Structure calculation from interproton distances.* First, we must use *ab initio* methods, i.e., those which begin with no prior knowledge of the structure - hence random starting structures are used, either random x,y, and z coordinates, or random phi-psi coordinates (i.e. structures with correct geometry). These methods then generate 'well-defined' structures which have reasonable geometry and agreement with *D*. The next step is to use refinement methods (see section below) which start from these well-defined structures, and attempt to improve them.

**Major problems**

The major open questions with NMR structure determination are:

P1 How to increase the molecular size limitation?

P2 How to automate the assignment process?

P3 Which structure calculation procedure to use?

P4 How to handle the dynamic nature of the data?

We will discuss new approaches to P2-P4.

***Ab initio* structure calculation in distance space**

Distance space is where each interatomic distance is considered a coordinate: hence we have $A(A-1)/2$ dimensions for a molecule of *A* atoms. Methods which work in this space are called distance geometry (DG) methods. DG methods were the first to be used to calculate structures from NMR data; they are still in wide use, although the molecular dynamics methods (next section) are better. We will cover DG in some detail.

*Aim.* Beginning only with *D*, search in distance space to find sets of complete distance matrices (which correspond to 3D structures) which simultaneously satisfies (as closely as possible) all restraints in *D*, covalent geometry restraints, and also some non-bonded term to prevent spatial overlap of the atoms.

*Method:* From *D*, construct upper- and lower-bound distance matrices reflecting the initial knowledge about *all* interatomic distances in the molecule. Most distances in these matrices will have initial lower bounds of the van der Waals radius, and initial upper bounds of infinity. These bound matrices are then smoothed by repeated application of the triangle inequality - $d(a,c) \leq d(a,b) + d(b,c)$. From these smoothed bound matrices, we then use some procedure (either random selection, or meterisation (Havel & Wüthrich, 1984)) to choose unique values for each interatomic distance, $d_{ij}$. Such a distance matrix is called **embeddable** if a 3D structure exists which is consistent with the matrix. From the distance matrix, we construct a metric matrix $g_{ij}$ - if the distance matrix is embeddable, the metric matrix has exactly three Eigenvalues which give the coordinates of the structure. Normally, however, due to inaccuracies in the data, and difficulties with the method, *d* is not embeddable; hence *g* has more than three Eigenvalues. In these cases, the three largest Eigenvalues are chosen. The resulting structure is normally quite poor, and require further refinement.

*Variations*: (Havel *et al.*, 1983; Kuntz *et al.*, 1989); DG with meterisation, (Havel & Wüthrich, 1984); substructures, (Havel & Wüthrich, 1984); linearized embedding, (Crippen, 1989).

*Results.* Provided *D* has a sufficient number of distances, the method works well. When *D* is sparse, there are sampling problems (Melzler *et al.*, 1989).

*Discussion* Initially, this approach was used for all NMR structure calculations. It is still widely used for the initial *ab initio* calculations, although DSA (see below) has better sampling and is more efficient. The method is sometimes used to generate substructures of about 1/3 of all the atoms; these are then refined with other methods. A major limitation is the requirement that all input data be expressed as distances: many restraints derived from NMR spectra cannot be expressed in terms of distances, e.g. ambiguous constraints. Hence this method is limited in application.

### *Ab initio* structure calculation in Cartesian space

In 3D Cartesian space, each atom is described by three coordinates; hence the total dimension for a molecule of *A* atoms is 3*A* - about *A/6* times fewer dimensions than for distance space. The method we will discuss in some detail is called dynamical simulated annealing (DSA) (Griewank, 1981). This method combines the simulated annealing principal (Metropolis *et al.*, 1953) with molecular dynamics techniques (Verlet, 1967).

*Aim.* Beginning from random structures, search in 3D Cartesian space to find sets of structures which satisfy *D*, covalent geometry, and non-bonded term.

*Method:* Start with either an extended polypeptide chain (Brünger *et al.*, 1986; Nilges *et al.*, 1988c), a random chain (i.e. random $\phi$ and $\psi$ angles) (Nilges *et al.*, 1991b), or with atoms in the gas phase (i.e. random Cartesian coordinates) (Nilges *et al.*, 1988a). Then calculate the dynamic trajectory of the system using a molecular dynamics (MD) force field, plus a 'soft' potential energy term (Nilges *et al.*, 1988c) which directs the motion toward structures which satisfy *D*; the soft potential switches between flat, square, and asymptotic behaviour:

$$E_{NOE}=k_{NOE}\sum_n \begin{cases} 0 & :\bar{D}_n<D_n \\ (\bar{D}_n-D_n)^2 & :\sigma>\bar{D}_n\geq D_n \\ \alpha(\bar{D}-\sigma)^{-1}+\beta(\bar{D}-\sigma)+\chi & :\bar{D}_n\geq\sigma \end{cases}$$

where the sum is over each NOE distance restraint $D_n$, $\bar{D}_n$ is the corresponding distance in the current model structure, and the parameters $\alpha$ and $\chi$ are set by the constraint that the function is continuous and differentiable at the switching distance $\sigma$.

In the DSA method, the temperature of the dynamical system is controlled by coupling to a heat bath. By setting the initial bath temperature to 1000K, and reducing the temperature gradually throughout the simulation, ending at or near zero, we anneal towards low energy structures. Essentially, we are simulating the condensation of the molecule from the liquid or gas phase to the solid phase.

*Variations.* DSA with DG-generated substructures (Nilges *et al.*, 1988b); solving symmetric multimers (Nilges, 1993; O'Donoghue *et al.*, 1993); PEACS (van Schaik *et al.*, 1992); RUSH (Li *et al.*, 1992; Byrne *et al.*, 1994); Monte Carlo approaches have also been tried, however in MD, motion is restricted to the physically plausible steps, effectively reduceding the dimensionality of the search space. Thus MD is expected to be more effienient than Monte Carlo (Griewank, 1981).

*Results.* The method is faster than DG, and has better sampling that the DG methods (Brünger *et al.*, 1987; Nilges *et al.*, 1991b).

*Discussion.* Currently DSA is the method of choice for *ab initio* structure generation. The method is very general and flexible, and is still being actively developed. An additional advantage over the DG method is the possibility of including ambiguous distance data. The approach is also potentially applicable to 3D structure prediction, provided that 2D distance information can be obtained (see section on 2D structure prediction).

### *Ab initio* structure calculation in torsion-angle space

In torsion-angle space, each torsion angle is considered a coordinate: every residue has two free backbone torsion angles ($\phi$ and $\psi$), and an average of about three side-chain torsion angles ($\chi_i$); thus, for a molecule of *R* residues we have a total dimension of about 5*R*, nine times less than for Cartesian space. We will discuss in detail the most popular implementation of these methods which is in the program DIANA (Güntert *et al.*, 1991).

*Aim.* Beginning from random structures, searches in torsion-angle space to find sets of structures which satisfy *D*, covalent geometry, and a non-bonded term.

*Method.* Beginning with random chains, a variable target function is used: in the first stage, only restraints between sequentially close residues are used. Later, all distance restraints are used. Minimisation is done with a gradient decent algorithm.

*Variations.* DISMAN, (Braun & Go, 1985); Monte Carlo methods (Bassolino *et al.*, 1988).

*Results.* Due to the reduced number of dimensions, these methods are fast. Some difficulties handling $\beta$-sheets, although work-around methods have been proposed.

*Discussion.* Currently shares the equal most popular position with DSA as a method for the *ab initio* structure generation. The high speed makes it useful for quick testing of distance data. Major limitation is the use of gradient minimisation - simulated annealing would probably give better performance. Not useful for further refinement. One disadvantage of these methods compared with DSA is the assumption of perfect geometry - real structure have occasional violations. So far, no MD methods have been used in torsion angle space because of the difficulty in solving Newton's equations with so many holomonic constraints, although the technique is being developed.

### Distance-based refinement

*Aim.* These algorithms start with a well-defined structure - by which we mean a structure with reasonable geometry, no serious van der Waals overlap, which also agrees reasonably well with $D$. The starting structure could come from any of the three *ab initio* techniques discussed above. The aim is to improve the agreement to the data, also possibly to fit the structure to a more sophisticated force field.

*Method.* All these methods use constrained MD in Cartesian space. Most popular is refinement with DSA algorithms (Nilges *et al.*, 1988b). A newer class tries to focus on the dynamic nature of proteins by generating ensembles of structures which have average distances that satisfy the NOE data: time-averaged distances (Torda *et al.*, 1989; Torda *et al.*, 1990); ensemble-averaging (Sheek *et al.*, 1991); exclusion potential (see talk).

*Results.* It is clear that many of these methods do improve the structures, and the computational requirement for these calculations can be easily met.

*Discussion.* As yet there is no consensus as to which is the best method. The most popular is to use DG-generated initial structures with DSA distance-refinement, ending the structure-determination at that point. The success of these methods is evident in the number of solution structures now being produced. In many cases, where similar structures are available from XRC, the agreement between the two independent methods is very good (usually better than 1Å RMSD). However, there still remains the question of the intrinsic dynamic nature of proteins: the calculation procedures do not sufficiently address this issue.

### Relaxation-matrix refinement

Here we describe a relatively new type of refinement procedure that promises to improve the accuracy NMR structures - but at a cost!

*Aim.* NMR does *not* measure distances directly - from the NOE spectra we obtain a set of cross-peak volumes, $V$, which we normally interpret as distances. But this interpretation has several assumptions which systematically fail. The main problem is called spin diffusion - the magnetisation transfer that we observe between two protons may have transferred *via* another proton. Thus we have some systematic errors in the distance set $D$. These algorithms attempt to address this problem. The algorithms start with structures already refined against $D$; the aim is to refine the structure to fit $V$.

*Method.* We calculate the volumes from the atomic coordinates, we calculate the matrix of magnetic relaxation rates between all possible proton pairs - this is called the complete relaxation matrix (Keepers & James, 1984). We also require the gradient of this matrix (Yip & Case, 1989). Unfortunately, both these are $O(N^3)$ algorithms, hence requiring significant computation time. However, several recent algorithms have been developed to speed the calculation, and the problem can be parallelised. As with distance-refinement, we use a dynamical force field with an annealing schedule, however we replace the distance potential $E_{NOE}$ with a potential which measures agreement to $V$ (Nilges *et al.*, 1991a).

*Variations.* torsion-angle minimisation (Mertz *et al.*, 1991); ensemble-averaging (Landis & Allured, 1991; Yang & Havel, 1993; Bonvin *et al.*, 1994; Forster & Mulloy, 1994).

*Results.* The method has not been widely used as yet, however in all cases tried so far, relaxation matrix refinement has changed the structures by about 1Å from the distance-refined structures, and has moved closer towards the crystal structure by about 0.4Å.

*Discussion.* Although computationally challenging, RMA refinement has been achieved at least for some small proteins. Probably the biggest initial barrier is the integration of all the NOE peaks - in the past, researchers usually just counted counter lines - the idea of going back and integrating several thousand peaks manually it not pleasant. However, new methods of assignment which increasingly involve computers from the beginning stages means that peak integration is increasing automatic. Thus, we are likely to see more RMA refinements. There is also currently a feeling of uncertainty about the application of this method, before the problem of multiple conformations has been adequately addressed.

### New calculation methods for assignment

We will discuss several new computational methods - all currently in progress - aimed at helping the assignment problem, which is the major bottle neck in the structure determination process. Methods covered include techniques for making assignments *after* the structure calculation floating assignment (Nilges, 1993; O'Donoghue *et al.*, 1993; Nilges, 1994); structure calculation in the absence of *any* initial assignments; also, attempts to use homologous structure to aid assignment.

## Protein structure in the crystalline state: X-ray diffraction

### Basic experimental methodology

*Crystallisation.* Proteins don't form crystals *in vivo*; convincing them to do so *in vitro* is a black art - but then, so is structure calculation. It is relatively easy to find conditions for a given protein to form a crystal, but forming the *right* crystal takes at lot of playing around with different solvents and conditions. Finding the right crystals can take one week, 15 years (the case of actin), or forever. Average time: about six months.

The smallest parallelepiped from which the whole crystal lattice can be constructed is called the **unit cell**. From group theory we know that there are only 65 possible types of crystal lattices in three-dimensions.

*Data collection.* Having got the right crystal, the next step is to put it into the path of a X-ray (generally 1.5Å wavelength) and record the resulting diffraction pattern. The diffraction results from scattered of the X-rays by interaction with the electrons in the structure; hence for each atom, scattering is proportional to its atomic number. Due to the crystal lattice, which acts like a grating, the diffraction pattern is made up of discrete *reflections.*

### Phasing

*Aim.* From the diffraction pattern, we can construct an image of what each unit cell looks like; we add a sine wave for each reflection, with frequency determined by its position in the diffraction pattern. But one thing is missing - we need to know the phases of these waves - this information is simply not in the diffraction pattern! This is the infamous phase problem of crystallography. We discuss briefly computational approaches to the problem.

*Ab initio methods.* The problem is to reconstruct the molecule from the discretely sampled FT without phase information. The problem can be solved for small molecules, using the additional constraints of atomicity and positivity. So far, however, this approach breaks down for more than about 40 atoms. *Direct methods* are those which calculate the phases automatically: these methods work up to about 40 residue (but only using very high quality data). A more promising approach involves representing the phase information in probabilities, and using maximum entropy and likelihood methods (Jaynes, 1978; Bricogne, 1984; Bricogne, 1991).

*Molecular replacement methods.* Phases can also be determined if a sufficiently similar (usually < 1.4Å) structure is available using molecular replacement (Hoppe, 1957; Rossmann & Blow, 1962). It has the theoretical danger that it uses previously determined structures to determine later structures, hence potentially adding biases. One example of this biases is that the number of structures solved with similar structures will be artefactually increased. This can affect database statistics on fold similarities.

Having obtained at least preliminary phase information, an electron density map can be constructed: the initial map is usually quite poor and requires substantial refinement. Substantial improvements can be made by imposing simple constraints on the electron-density map consistent with chemical knowledge about the molecule (atomicity, positivity, map continuity, etc.).

### Model building

*Aim.* To build an atomic model of the protein which fits into a given electron density map.

*Methods.* Hand-building is still very popular. Computer-assisted methods are available using fragment databases, however these require expert knowledge to use correctly. Several attempts have been made at developing automatic approaches (Read & Moult, 1992; Lamzin & Wilson, 1993). This is a problem suited to artificial intelligence methods (Fortier *et al.*, 1993); neural networks have also been tried (Torda, personal. comm.).

### Refinement

*Aim.* The initial structures built from the density maps can be very crude - the aim is to refine these models in Cartesian space.

*Methods.* The method of choice is DSA (Brünger *et al.*, 1987) - the minimisation procedure is similar to that for NMR DSA - molecular dynamics force field with an additional term to constrain the structure to fit the X-ray data.

## Structure verification and assessment

How can we define which structures are acceptable? How do we measure the quality of structures? There are two approaches to these questions: acceptable structures must agree with the data used to derived them (internal criteria), and they must also satisfy additional criteria derived from our knowledge about what correct structures look like (external criteria).

Regarding the internal criteria, the use of 'free' R-factors has been recently proposed for both XRC and NMR structure determination (Brünger, 1992; Brünger, 1993; Brünger *et al.*, 1993). This quantity is derived analogously to the cross-validation statistics (see Evaluation of prediction methods, next chapter). Use of this quantity promises to avoid over-refining and to recognise errors.

Regarding the external criteria: several packages are now available for checking protein structures: Procheck (Thornton, University College, London) and What if (Vried, EMBL) check covalent geometry against small molecule databases, as well as some stereochemistry, and overlap checking. These groups have combined with several others and the PDB in a project to provide comprehensive checking tools - they have a common WWW server (http://www.embl-heidelberg.de:8400/) which has a hypertext interface to their programs and can be used to submit structures for checking. Unfortunately, at the moment, support for NMR structures is only limited. A very different type of checking is done by Prosa (Sippl, University of Salzburg): it uses mean force potentials derived from the PDB to assess if the overall fold is native-like (see next chapter).

# Computational tools for experimental determination and theoretical prediction of protein structure

- Introduction to protein structure
- Experimental determination of protein structure
- Prediction of protein structure



# Computational methods for experimental structure determination

- Overview
  - *`State of the art*
  - *Importance of the computational methods*
  - *Comparison of methods*
- Solution structures: NMR
  - *Basic methodology*
  - *Ab initio calculation*
  - *Refinement*
  - *Assignment*
- Crystal structures: XRC
  - *Methodology*
  - *Phasing*
  - *Model building and refinement*

# Overview

- The State of the art
  - *unprecidented rate of structure determination - Fig. 2.1*
  - *The PDB format -showing its age*
  - *<250 residues - NMR or XRC*
  - *>250 residues - XRC*
  - *very large complexes, membrane proteins - ?*
- Importance of calculation techniques
  - *Set the scope of what structures can be solved*
  - *Accuracy and precision are interwoven with methods*
  - *How to compare methods?*



# Fig. 2.1: The growth of the protein data bank

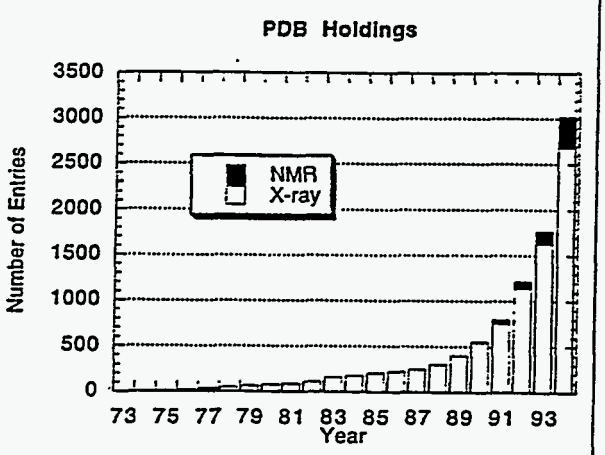



Fig. 1

## Fig. 2.2: The PDB format - showing its age

- In case you drop your stack of computer cards on the floor...

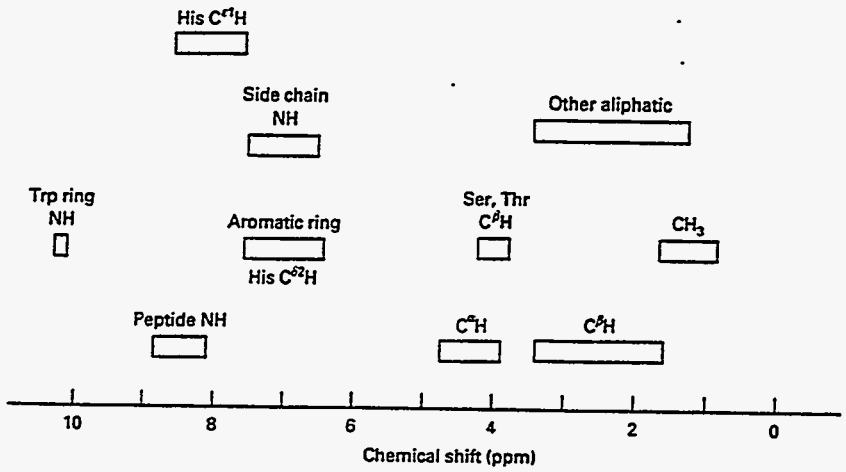
```
ATOM    1  CA   ACE A   0     105.046  51.546  40.626  1.00 72.72      1ATN 263
ATOM    2  C    ACE A   0     105.314  50.822  41.951  1.00 72.72      1ATN 264
ATOM    3  O    ACE A   0     105.220  51.451  43.013  1.00 72.56      1ATN 265
ATOM    4  N    ASP A   1     105.665  49.507  41.867  1.00 71.64      1ATN 266
ATOM    5  CA   ASP A   1     105.992  48.589  42.982  1.00 70.20      1ATN 267
ATOM    6  C    ASP A   1     107.024  49.191  43.936  1.00 69.70      1ATN 268
ATOM    7  O    ASP A   1     106.927  49.088  45.163  1.00 69.14      1ATN 269
ATOM    8  CB   ASP A   1     106.533  47.248  42.410  1.00 70.66      1ATN 270
ATOM    9  CG   ASP A   1     106.801  46.077  43.383  1.00 71.73      1ATN 271
ATOM   10  OD1  ASP A   1     107.722  46.143  44.215  1.00 71.57      1ATN 272
ATOM   11  OD2  ASP A   1     106.092  45.066  43.291  1.00 71.25      1ATN 273
ATOM   12  N    GLU A   2     107.976  49.873  43.293  1.00 69.24      1ATN 274
ATOM   13  CA   GLU A   2     109.054  50.658  43.886  1.00 69.94      1ATN 275
ATOM   14  C    GLU A   2     108.707  51.166  45.277  1.00 69.71      1ATN 276
ATOM   15  O    GLU A   2     109.454  51.029  46.250  1.00 69.74      1ATN 277
ATOM   16  CB   GLU A   2     109.372  51.861  42.969  1.00 69.58      1ATN 278
ATOM   17  CG   GLU A   2     110.164  51.624  41.669  1.00 68.60      1ATN 279
ATOM   18  CD   GLU A   2     109.564  50.572  40.753  1.00 68.20      1ATN 280
ATOM   19  OE1  GLU A   2     108.416  50.739  40.320  1.00 67.20      1ATN 281
```



## Structure of proteins in solution: NMR spectroscopy

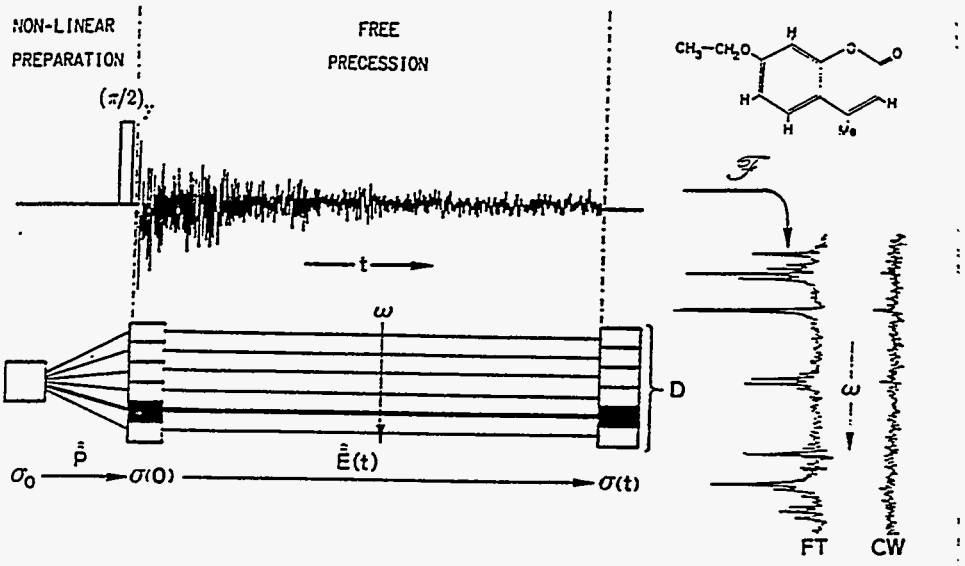
- Experimental methods
  - *Requirements for sample*
  - *Spectrometer - expensive ,superconductor*
  - *Background theory*

## Fig. 2.3: 1D NMR spectra: Chemical shifts for different hyrdogen atoms in peptides

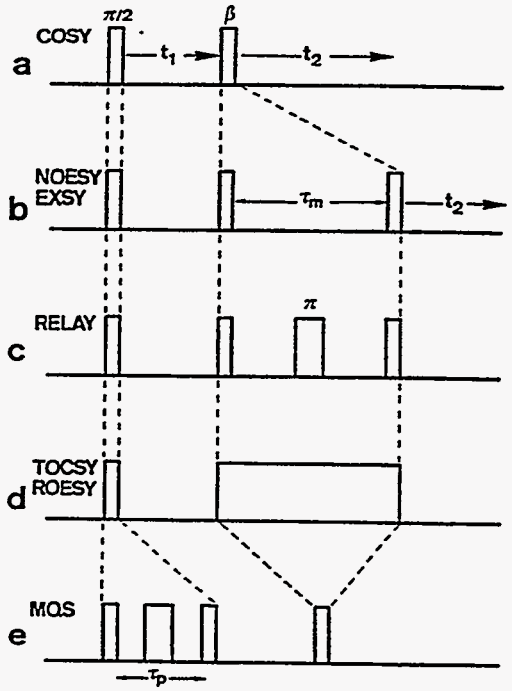





## Fig. 2.4 : Continuous wave vs Fourier transform spectroscopy

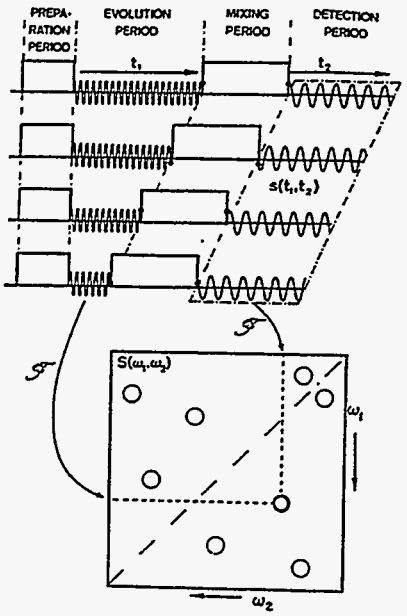


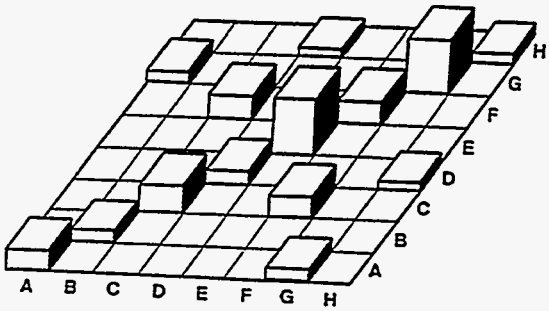
# Fig. 2.5 : Pulse sequences - a black art



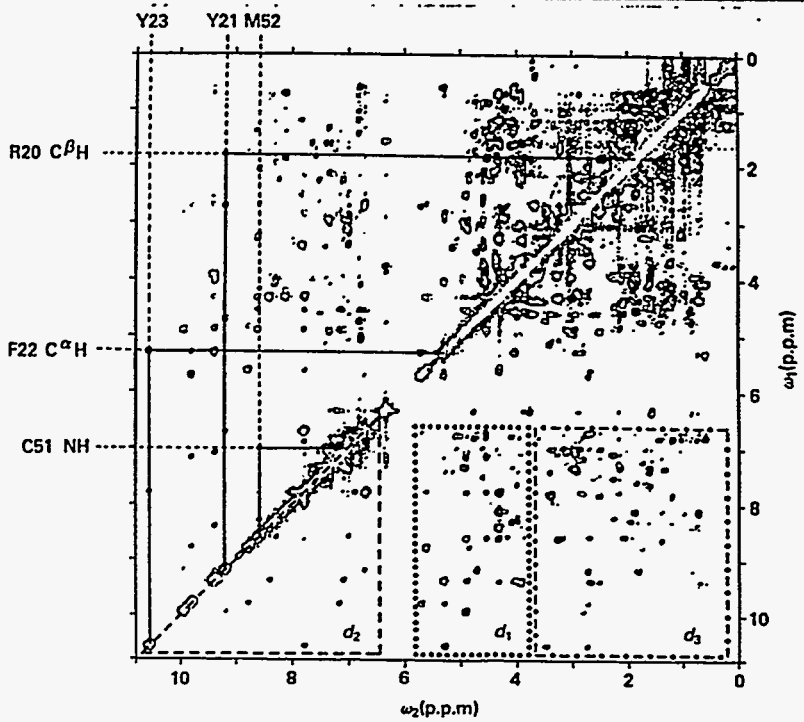



# Fig. 2.6: Schematic representations of 2D spectra



## Fig. 2.7 2D NOE spectrum of a small protein

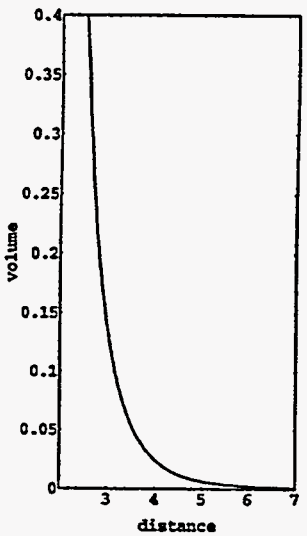





## NOE volumes vs distance - the two spin approximation

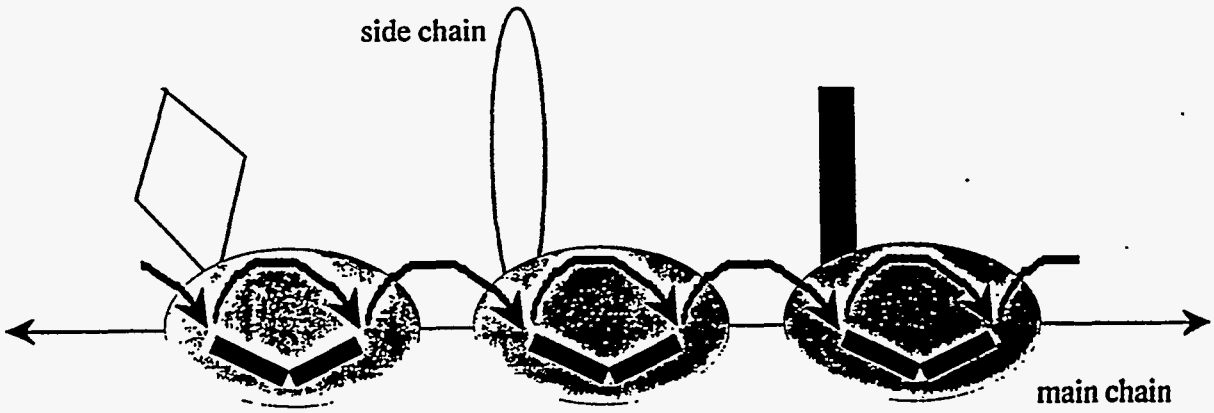
- *Under certain conditions*: The distance between a pair of protons *a* and *b* is related to the volume by:

$$V_n = c\mathrm{d}(a_n, b_n)^{-6}$$




- In practice <5 Angstrom
- spin diffusion and other evils

## Fig. 2.8 Sequential assignment: For once, Mother Nature is generous!





within unit: covalent *and* spatial connectivity
between units: *only* spatial connectivity



## Major challenges for NMR

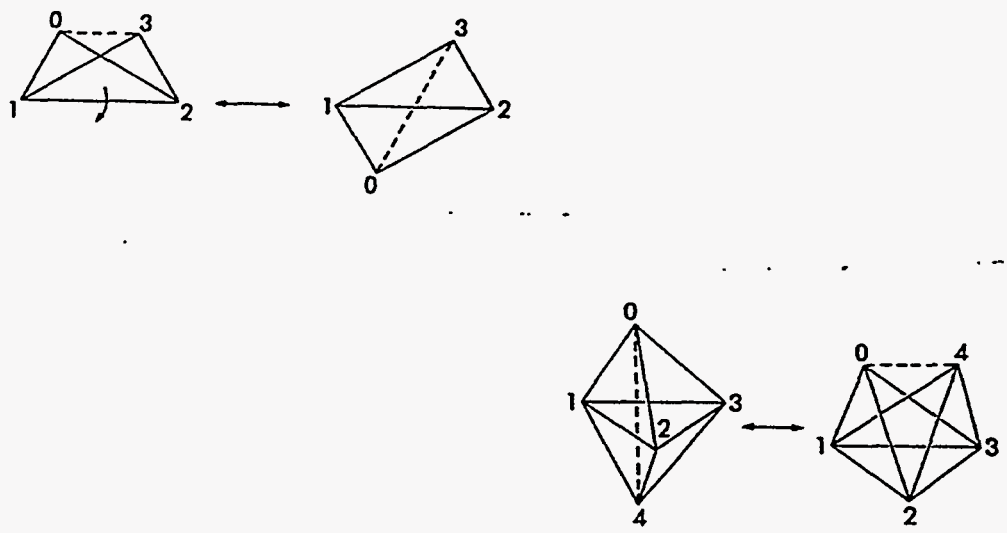
- P1 How to increase the molecular size limitation?
- P2 How to automate the assignment process?
- P3 Which structure calculation procedure to use?
- P4 How to handle the dynamic nature of the data?

# Ab initio calculation: Distance geometry

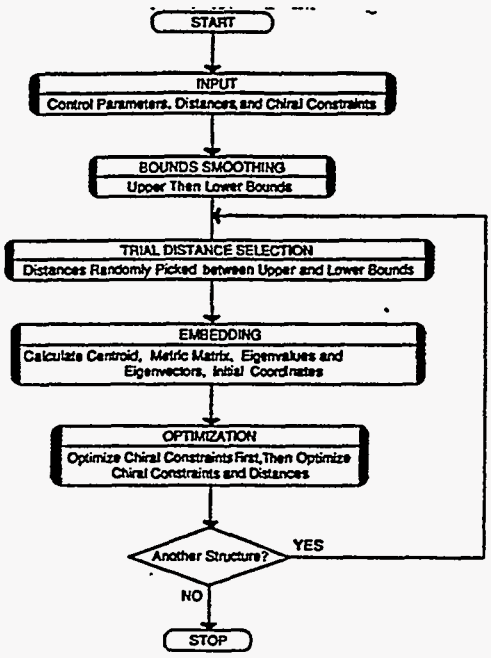
- Distance space
  - *high-dimensional space*
  - *Requires* all *interatomic distances*
- Triangle inequality- $d(a,b) \leq d(a,b) + d(b,c)$
- Also tetrangle and pentangle inequalities, but... (Fig. 2.9)
- Bounds smoothing
- metric matrix - *g*
- embedding



# Fig. 2.9: Tetrangle and pentangle inequalities



# Fig. 2.10: Distance geometry algorithm






# Fig. 2.11: Problems with metrisation

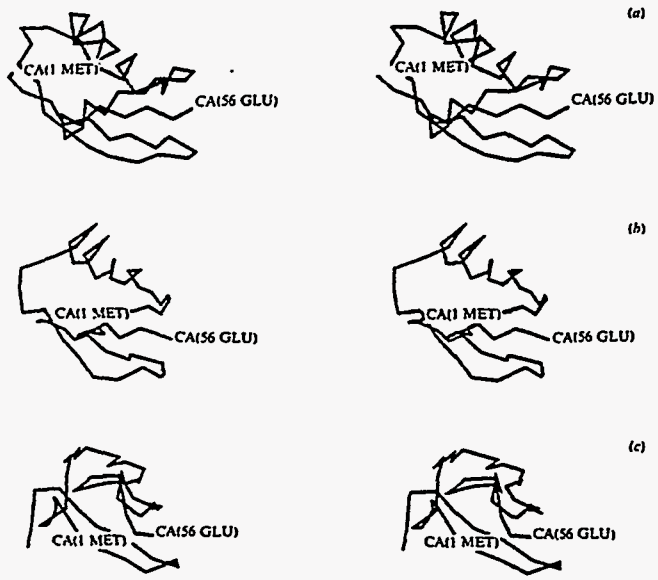
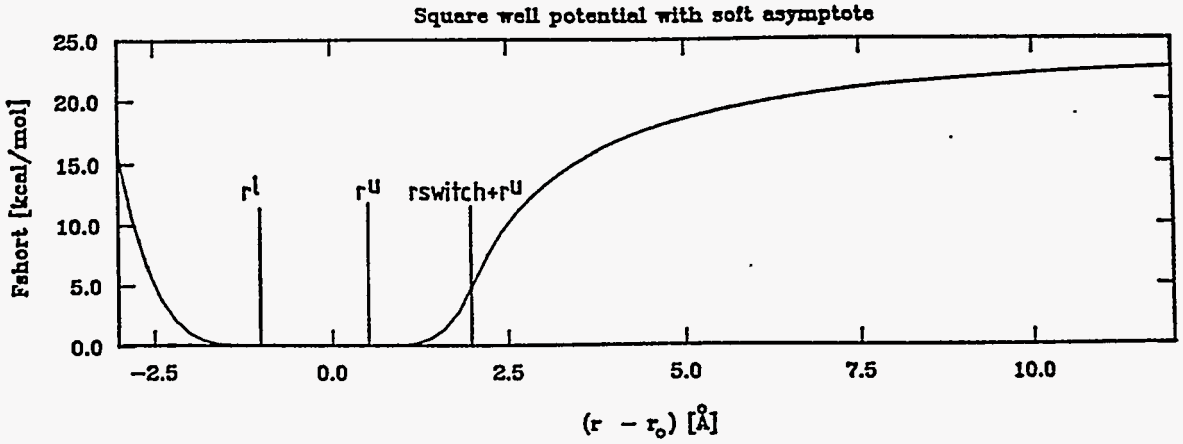



Fig. 9. $C^{\alpha}$ backbone traces of protein G. (*a*) NMR structure as described by Gronenborn *et al.* (1991). (*b*) Structure after metric matrix distance geometry without metrization. (*c*) Structure after metric matrix distance geometry with complete random metrization.

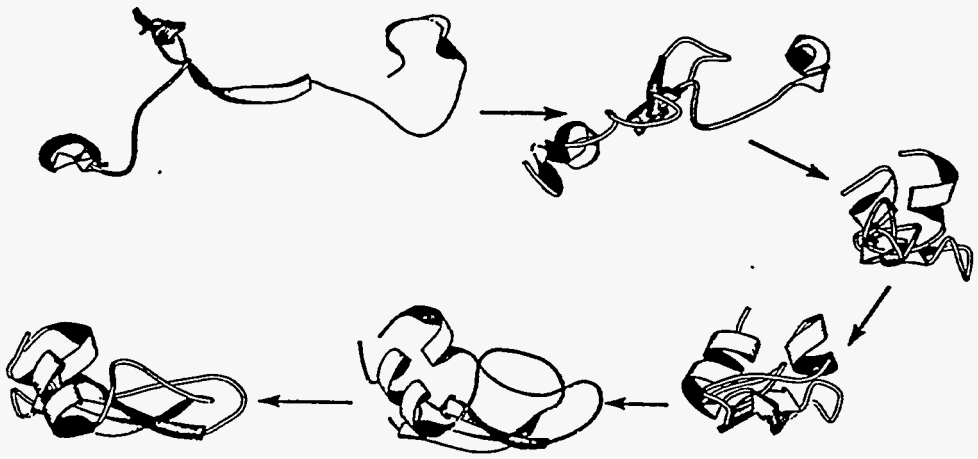
# Fig. 2.12: The soft potential fuction



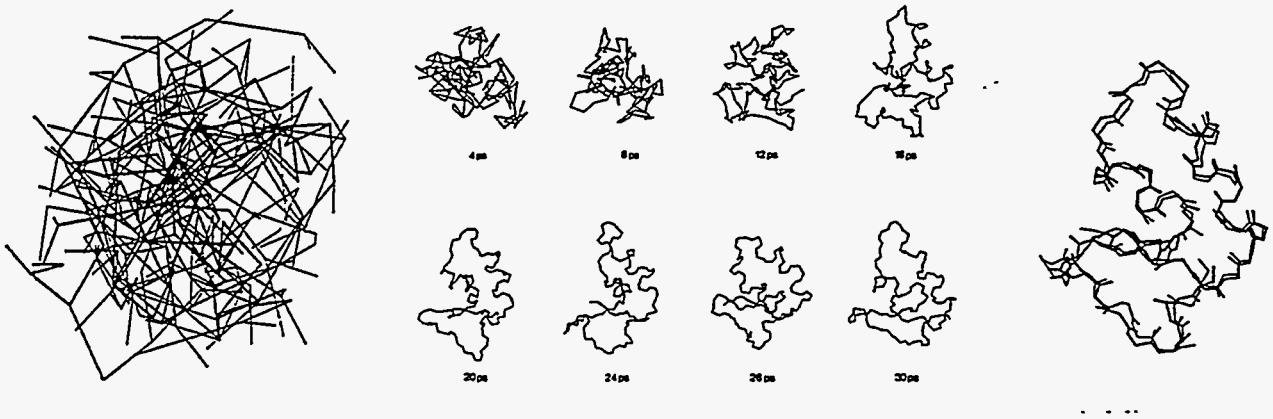

2. Potential form of $F_{short}$ (equation 6) for $r_0 = 3.0$ Å, $k_s = 1.0$ kcal/mol/Å$^2$, $r^l = 2.0$ Å, $r^u = 3.5$ Å, *rswitch* = 1.5 Å and $c = 0$. The positions $r^l$, $r^u$ and *rswitch* + $r^u$ are indicated.



# Fig. 2.13: Annealing a protein structure from the liquid to the solid phase

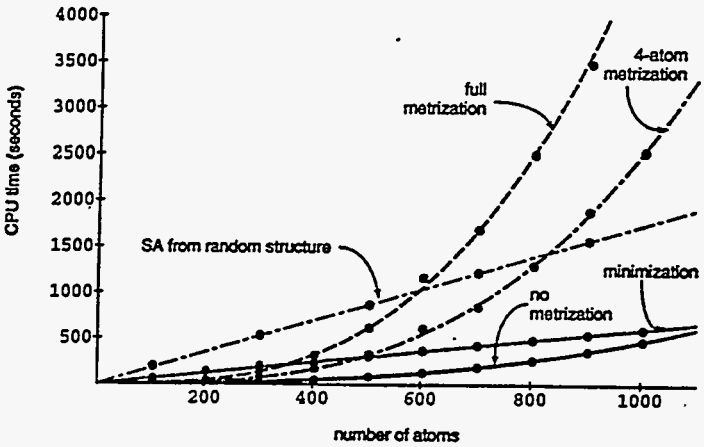


# Fig. 2.14: Annealing a protein structure from the gas to the solid phase

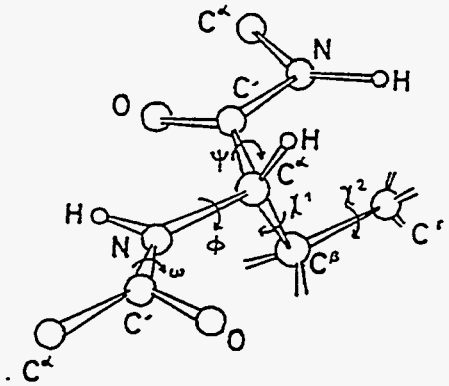





# Fig. 2.15: Comparision of CPU times for DG vs. DSA

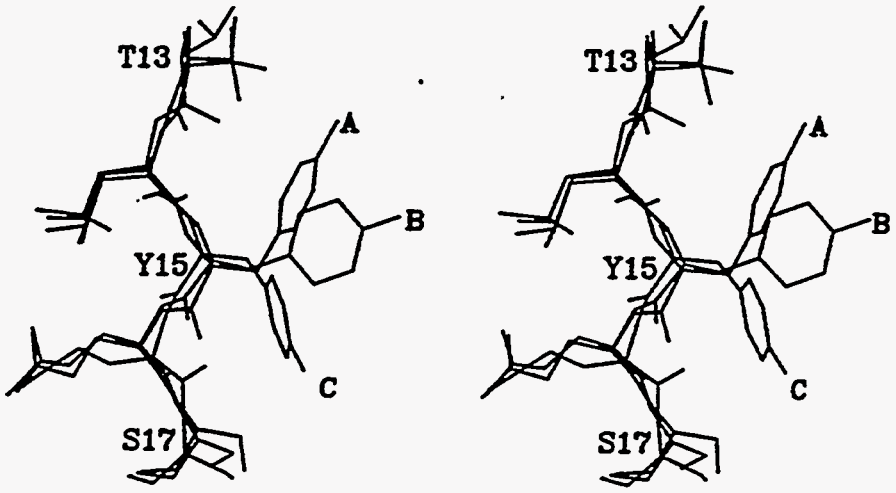


## Fig 2.16: Torsion-angles in a protein - torsion-angle space

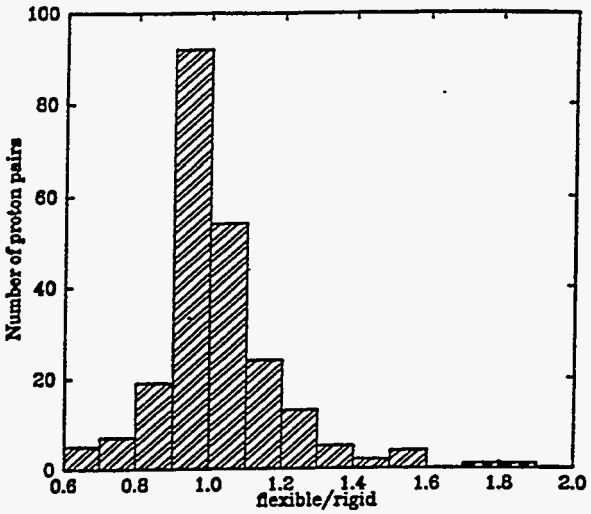





## Fig. 2.17: Methods which consider protein dynamics - time-average constraints

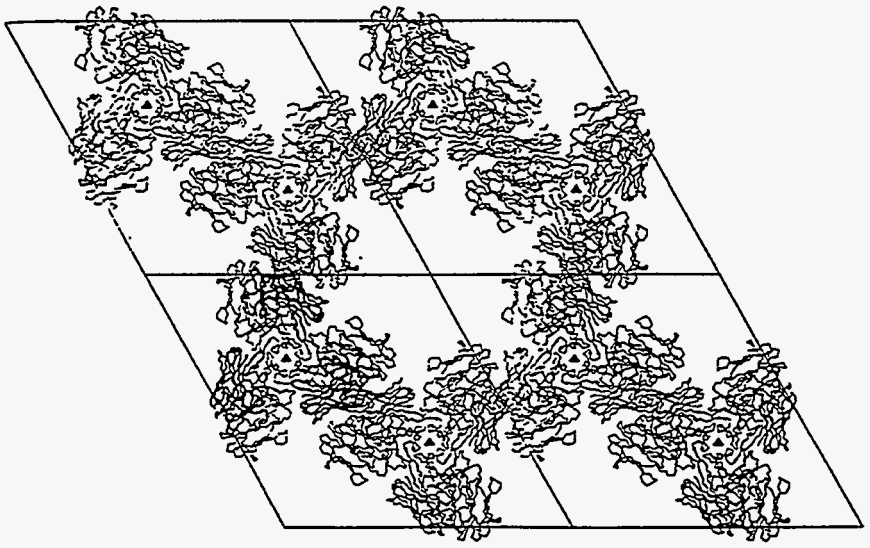


## Fig. 2.18: Effect of motion on relaxation rate

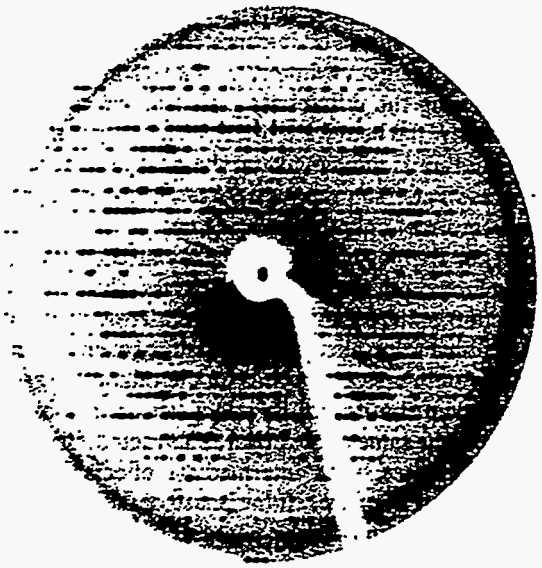





## Fig. 2.19: A protein in the crystalline state: the unit cell of an immunoglobulin Fab fragment

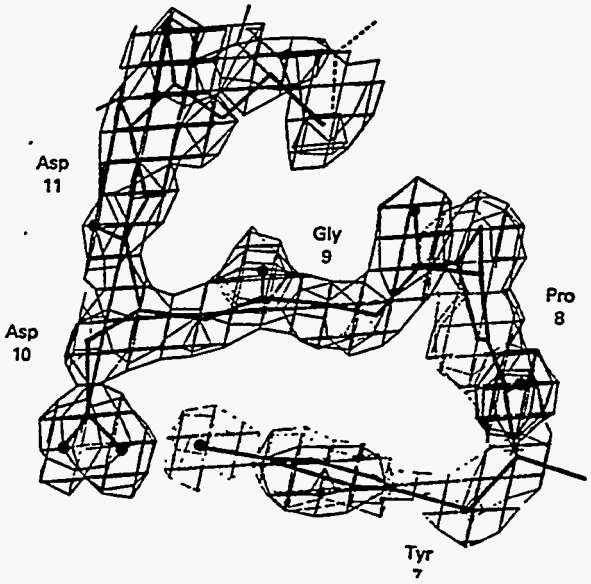


## Fig. 2.20: Example diffraction pattern

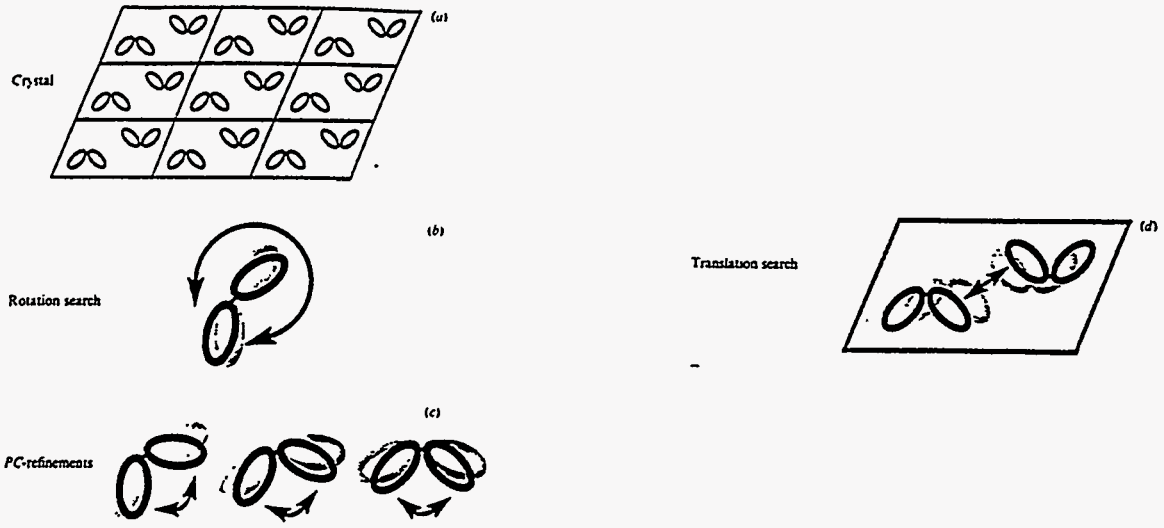




## Fig. 2.21: Fitting a protein model into a refined electron density map



# Fig. 2.22: Molecular replacement search strategy



# Prediction of protein structure

## Contents



## Summary

*Theoretical tools become increasingly demanded.* Suppose one has a protein sequence of unknown structure, say SOUS. What can be learned about SOUS before beginning an experiment? Data banks of protein sequences and structures are growing rapidly (Bernstein, et al. 1977, Abola, et al. 1988, Bairoch & Boeckmann 1994)as a result of large-scale sequencing projects (Oliver, et al. 1992, Johnston, et al. 1994) and improvements in experimental determination of 3D structure (Holm & Sander 1994c, Lattman 1994) .Can we profit from the information flood? Does the data bank teach us how to predict the 3D structure of SOUS?

*Homology modelling allows prediction in 3D.* The most successful tool for prediction of three-dimensional structure is homology modelling. An approximate 3D model (which has a correct fold, but inaccurate loop regions) can be constructed if SOUS has significant similarity to a protein of known structure, evaluated in terms of sequence similarity (i.e. alignment) or sequence-structure fitness (i.e. threading). Homology modelling effectively raises the number of 'known' 3D structures from about 1500 to 10.000 (Sander & Schneider 1991, Sander & Schneider 1994). But what if SOUS has no homologue of known 3D structure? Can 3D structure be predicted directly from sequence?

*For most proteins the prediction task has to be simplified.* Without detectable homology we are still forced to resort to simplifications of the prediction problem. In the process, we can make use of the rich diversity of information in current data bases. For this tutorial we have selected generic methods for prediction at three different levels of simplification (Fig. 3.2), namely one, two and three dimensions (for a short review (Rost & Sander 1994e). Prediction in 1D (secondary structure, solvent accessibility and transmembrane helices) can be improved significantly through the use of evolutionary information. Prediction in 2D (inter-residue contacts, inter-strand contacts, disulphide bonds) can also, to a certain extent, profit from evolutionary information, but so far, is of only limited accuracy. Lastly, incorrect 3D structures can now be detected with remarkable accuracy (mean-force potentials) and technical improvements and data base growth have made alignments, threading and homology modelling become increasingly powerful.

## Overview

*What is the state of the art in structure prediction?* We cannot predict 3D structure in general, yet (Rost & Sander 1994e). The most successful theoretical tool for the prediction of structure is homology modelling (Greer 1980, Greer 1991, Holm, et al. 1994, May & Blundell 1994). Homology detectable by significant sequence identity (>25%) to a protein of known 3D structure can be applied to some 25% of all known proteins (Sander & Schneider 1994). In absence of significant sequence identity, threading techniques can be used for remote homology modelling (Sippl & Weitckus 1992, Sippl & Jaritz 1994). (The lack of reliability of current threading techniques makes it difficult to estimate the scope of this technique. The number of proteins for which 3D structure could be predicted based on remote homology would currently probably be some thousands (Holm & Sander 1994a).) For proteins, for which neither homology modelling nor threading is applicable, the prediction problem has to be simplified (Rost & Sander 1994e).

*How can the prediction problem be simplified?* The most extreme simplification is to project the full complexity of 3D information onto 1D, i.e., secondary structure, or solvent accessibility assignments for each residue. Less information is

lost, when projecting 3D co-ordinates onto 2D maps of inter-residue distances. As explained in the experimental section, 3D structure can be generated from 2D maps.

*Which prediction is of interest for molecular biology?* Large scale gene-sequencing projects produce an overwhelming information about protein sequences (Oliver, et al. 1992, Johnston, et al. 1994). This information alone is not very useful for molecular biology. A crucial step is to associate the sequences to information about structure or function of the proteins (Bork, et al. 1992b). Given the rapid advance of sequencing techniques, such association cannot be gathered exclusively by experiments. Instead, theory has to contribute to closing the sequence-structure and sequence-function gaps. Consequently, any prediction method that contributes to closing these gaps is of help. However, not all aspects of protein structure are valuable. For example, the prediction that a protein belongs to the all-alpha class may be useful if used as input to a post-processing method that, e.g., predicts remote homology, but is hardly of any use *per se*.

## Evaluation of prediction methods

*Publishing optimistic results?* A sustained testing of the performance is a precondition for any prediction to become useful. For example, the history of secondary structure prediction has gone through a head-hunting for highest accuracy scores. Over-optimistic claims by predictors on the one hand, nourished scepticism of potential users on the other hand. Two points became clear in the first meeting for the 'State of the art in structure prediction' in Asilomar, C.A., Dec., 1994 (Defay & Cohen 1995). First, an inaccurate prediction is not as bad, as is an over-estimated one. Second, even a prediction method of limited accuracy can be useful if the user knows what to expect. In the following, some criteria will be summarised which help reducing the likelihood to fall into the trap of overestimation. As an example the prediction of secondary structure will be chosen.

*What is the goal and which limits are to be expected?* Say the goal is to predict secondary structure in three states. Which is the best current method for the prediction of secondary structure? If applicable, homology modelling. Which is the worst method? Random prediction. How accurate are existing methods for prediction?

*How to choose the data set?* Proteins used for deriving a method and for evaluating should have a pairwise sequence identity < 25%, else homology modelling can be applied (Sander & Schneider 1991). For all prediction methods the data set has to be split into a set used to set free parameters (training set) and another to estimate the expected performance on unknown proteins (test set). The criterion for separating training and test set is provided by the best alternative method (homology modelling).

*How many proteins to use for the test set?* All available unique proteins should be used for testing (currently more than 300 (Hobohm & Sander 1994)). Furthermore, results should be compared to standard sets used for the evaluation of alternative methods (Rost, et al. 1993, Rost & Sander 1994b). The reason for taking as many proteins as possible is simply that proteins have a wide spread facet of features: some are easy to predict, others harder. A criterion for a sufficient size of a test set could be the following. N proteins are enough, if (and only if) the standard deviation of a certain measure for accuracy fulfils:

$$\sigma_N \approx \sigma_{2N} ,$$

in other words, if doubling the test set would not alter the results.

*Optimising free parameter with respect to the test set?* The cross-validation described so far is still not enough. A seemingly trivial - and often violated - rule is that methods should never be optimised with respect to the test set. Instead, parameters should be optimised (if necessary) based on yet a different set (screening or optimisation set), and should be kept fixed BEFORE the final cross-validation experiment is performed.

*How many cross-validation experiments have to be performed?* Say the test set consists of 300 proteins. One extreme a two-fold cross-validation would mean to split the set into two set with 150 proteins each (A and B) and perform two experiments: first train on A, test on B, then train on B, test on A, and finally report the test results for A+B. The other extreme a 300-fold cross-validation (jack-knife) implies 300 splits into pairs of a training sets A with 299 proteins and test sets B with one protein each such that each protein is in one of the test sets. After 300 experiments the results are averaged over all 300 test sets. In practice the choice is often somewhere between two and 300. Does the number of splits have an influence on the correctness of the evaluation? The simple answer is: no! More splits tend to be better for the methods, as more proteins can be used for training. But with respect to the generality of the result there is no difference between two- and 300-fold cross-validation (as long as the previous points had been taken into account).

*Enough of testing?* Even if all those points had been taken into account, the sceptical molecular

biologist should still not be satisfied. A further necessary step is to test the method on a new set of proteins, ideally after the paper had been written. With the rapid increase in the number of known structures, it should never be difficult to find say some 50 proteins which have no significant sequence identity to any of the 300 proteins used so far. This final test helps the reader and the predictor to assess whether or not the estimated performance is likely to be correct for new proteins.

*How to measure performance accuracy?* Another rather obvious demand is that to define an appropriate measure. The measure should reflect the goal of the method. For example, if the goal is to predict 3D structure by remote homology modelling (threading), the results have to be given in e.g. root-mean square deviation. This example may appear particularly trivial, nevertheless, the current practice is the opposite. Another negative example are alignments, after more than 25 years of dynamic programming, there is still no measure for the quality of an alignment published that was tested on a large enough data set. No matter what the measure is, the predictor should always provide an estimate for the standard deviation of the expected accuracy!

**Evaluation of prediction methods: Literature**

Evaluation of secondary structure predictions: (Kabsch & Sander 1983b, Rost, et al. 1993, Rost & Sander 1994b, Defay & Cohen 1995)

Measures for secondary structure predictions: (Schulz & Schirmer 1979, Cohen, et al. 1983, Taylor 1984, Taylor & Thornton 1984, Cohen, et al. 1986, Cohen & Kuntz 1989, Benner 1992, Presnell, et al. 1992, Sternberg 1992, Thornton, et al. 1992, Benner, et al. 1993, Colloc'h, et al. 1993, Rao, et al. 1993, Rost & Sander 1993b, Rost, et al. 1993, Russell & Barton 1993, Rost, et al. 1994c)

## Prediction of protein structure in 1D

**Secondary structure prediction**

*Prediction methods.* Secondary structure prediction has been attempted even before the first X-ray structures became known (Szent-Györgyi & Cohen 1957, Kendrew, et al. 1960, Perutz, et al. 1960). Ever since the problem startled many researchers and served for many physicists, mathematicians and computer scientists as an entrance into the world of molecular biology. The principle idea of most methods is to make use of the fact that segments of consecutive residues have preferences for certain secondary structures (Fig. 3.7). Thus, the prediction problem becomes a pattern-classification problem tractable by computer algorithms (3T-18/21). In many respects secondary structure prediction reflects the principle difficulties and solutions for many prediction algorithms. Therefore, we shall cover this topic in more detail than the others. Three basic algorithms are described: information theory (3T-23); neural networks (3T-24-31); and nearest neighbour classifiers (3T-32/34). Despite improving the algorithms in detail, the real break-through came by using evolutionary information (3T-35/38). Additionally, two specialised versions of the secondary structure prediction problem will be discussed: the prediction of secondary structure content (3T-50/54) and the prediction of secondary structure in two states, e.g., helix/non-helix (3T-55).

*Necessity of sustained testing.* Useful computational are urgently demanded by molecular biologists. However, to make a prediction method useful three points have to be met. Firstly, the predicted feature of protein structure or function has to be suitable for an experiment (or post-processing prediction methods). Secondly, the method has to be made available. And thirdly, most importantly to keep theory in the game, prediction accuracy has to be estimated at a sustained level. In the wake of today's flood in literature, experimental biologists and even theoreticians from slightly different fields have no chance to critically assess the claims of predictors. Thus, the application of prediction methods requires quite a level of trust. This demands a significant level of modesty on the side of the predictors. Levels of expected accuracy should be conservative and tend to under-estimate rather than to over-estimate the abilities of tools. Given the ease of distributing software and services over current internet resources, the issue of appropriate evaluation becomes increasingly sensitive. For the predictor appropriate evaluation implies to spend much more time on testing than on developing a tool. Secondary structure predictions can serve as an example for how to appropriately test methods. We shall in detail discuss different measures for prediction accuracy (3T-38/43).

*Evaluation of impact.* For single sequences prediction accuracy is about 60%. It raises above 70% if information from multiple alignments is used (3T-45/47). These levels of expected accuracy of course are not sharp numbers, but rather averages of underlying distributions, e.g., for PHD (neural network prediction) the three-state overall per-residue accuracy for single proteins chains is 72±9% (Fig. 3.20). Of practical use is the definition of a reliability index for the prediction (Fig. 3.21). For prediction methods, secondary

structure predictions are a simple example, how expert knowledge and the wealth of growing data bases can be carved into improved methods.

*Secondary structure prediction: Literature*

*Reviews:* (Kabsch & Sander 1983b, Schulz 1988, Fasman 1989a, Garnier & Levin 1991, Rost, et al. 1993)

*Measures:* (Schulz & Schirmer 1979, Cohen, et al. 1983, Taylor 1984, Taylor & Thornton 1984, Cohen, et al. 1986, Cohen & Kuntz 1989, Benner 1992, Presnell, et al. 1992, Sternberg 1992, Thornton, et al. 1992, Benner, et al. 1993, Colloc'h, et al. 1993, Rost & Sander 1993b, Rost, et al. 1993, Russell & Barton 1993, Rost, et al. 1994c)

*Methods (only basic and new methods listed):* (Pain & Robson 1970, Nagano 1973, Chou & Fasman 1974, Lim 1974, Nagano & Hasegawa 1975, Maxfield & Scheraga 1976, Nagano 1977, Garnier, et al. 1978, Maxfield & Scheraga 1979, Cohen, et al. 1980, Cohen, et al. 1983, Ptitsyn & Finkelstein 1983, Taylor & Thornton 1983, Gibrat, et al. 1987, Zvelebil, et al. 1987, Biou, et al. 1988, Bohr, et al. 1988, Gascuel & Golmard 1988, Qian & Sejnowski 1988, Holley & Karplus 1989, McGregor, et al. 1989, Benner & Gerloff 1990, Fasman 1990, King & Sternberg 1990, Kneller, et al. 1990, Rooman, et al. 1991, Rooman & Wodak 1991, Benner 1992, Hayward & Collins 1992, Muggleton, et al. 1992, Presnell, et al. 1992, Rost & Sander 1992a, Rost & Sander 1992b, Sternberg 1992, Stolorz, et al. 1992, Zhang & Chou 1992, Zhang, et al. 1992, Asai, et al. 1993, Barton & Russell 1993, Benner, et al. 1993, Fariselli, et al. 1993, Levin, et al. 1993, Maclin & Shavlik 1993, Rost & Sander 1993a, Rost & Sander 1993c, Rost & Sander 1993b, Sasagawa & Tajima 1993, Yi & Lander 1993, Donnelly, et al. 1994, Livingstone & Barton 1994, Rost & Sander 1994b, Solovyev & Salamov 1994, Wako & Blundell 1994b, Salamov & Solovyev 1995)

*Methods (information theory):* (Pain & Robson 1970, Robson & Pain 1971, Nagano 1973, Chou & Fasman 1974, Robson 1974, Robson & Pain 1974a, Robson & Pain 1974c, Robson & Pain 1974b, Nagano & Hasegawa 1975, Robson 1976, Nagano 1977, Suzuki & Robson 1977, Chou & Fasman 1978, Garnier, et al. 1978, Levin, et al. 1986, Gibrat, et al. 1987, Biou, et al. 1988, Chou 1989, Zhang, et al. 1992, Levin, et al. 1993)

*Methods (neural networks):* (Bohr, et al. 1988, Qian & Sejnowski 1988, Holley & Karplus 1989, McGregor, et al. 1989, Bossa & Pascarella 1990, Kneller, et al. 1990, Hayward & Collins 1992, Muskal & Kim 1992, Pancoska, et al. 1992, Rost & Sander 1992a, Salzberg & Cost 1992, Stolorz, et al. 1992, Zhang, et al. 1992, Andrade, et al. 1993, Fariselli, et al. 1993, Maclin & Shavlik 1993, Presnell & Cohen 1993, Rost 1993, Rost & Sander 1993a, Rost & Sander 1993c, Sasagawa & Tajima 1993, Tchoumatchenko, et al. 1993, Rost & Sander 1994b, Rost, et al. 1994a, Chandonia & Karplus 1995, Rost 1995b)

Methods (nearest neighbour): (Kabsch & Sander 1983c, Levin, et al. 1986, Schneider 1989, Zhang, et al. 1992, Yi & Lander 1993, Solovyev & Salamov 1994, Salamov & Solovyev 1995)

**Solvent accessibility prediction**

*Prediction methods.* The principle goal is to predict to which extent a residue embedded into a protein structure is accessible to solvent. Various ways for the description of solvent accessibility are possible (3T-59). The most simple is a two-state model that distinguishes whether the residue is buried or exposed. Solvent accessibility is evolutionarily conserved (3T-60). Two prediction methods will be described: neural networks and information theory-based predictions.

*Evaluation of impact.* Prediction accuracy is > 70% in a two-state (buried, exposed) description of relative accessibility. This level is sufficient to use predictions as a seed for predicting secondary structure (Benner, et al. 1994, Wako & Blundell 1994b), but it is not accurate enough to make predictions become as useful as secondary structure predictions (Rost 1995a). Although accessibility predictions have to viewed with scepticism, they contain information that is useful for many post-processing prediction methods.

*Solvent accessibility prediction: Literature*

*Definitions and hydrophobicity scales:* (Lee & Richards 1971, Chothia 1976, Janin 1976, Richmond & Richards 1978, Wodak & Janin 1978, Cohen, et al. 1980, Wodak & Janin 1980, Kyte & Doolittle 1982, Sweet & Eisenberg 1983, Eisenberg, et al. 1984a, Eisenberg, et al. 1984b, Cornette, et al. 1987, Hubbard & Blundell 1987, Lawrence, et al. 1987, Ponder & Richards 1987, Flores, et al. 1993, Jackson & Sternberg 1993, Rost & Sander 1994c)

*Methods:* (Holbrook, et al. 1990, Benner, et al. 1994, Esposito, et al. 1994, Rost & Sander 1994c, Wako & Blundell 1994a)

**Transmembrane segment predictions**

*Prediction methods.* Even in the optimistic scenario that in the near future most protein structures will be either experimentally determined or theoretically predicted, one class of proteins will certainly be abundant in terms of knowledge about 3D structure: transmembrane proteins. The major difficulty is that integral membrane proteins do not crystallise and are hardly tractable by NMR spectroscopy. Consequently, four this class

predictions will be even more important. Fortunately, the prediction task for transmembrane proteins is easier than for globular proteins, as the lipid bilayer of the membrane reduces the degrees of freedom to such an extent that 3D structure formation is almost a 2D problem. Once the location of transmembrane segments is known for, e.g., helical transmembrane proteins, 3D structure can be predicted by exploring all possible conformations (Taylor, et al. 1994). And even the prediction of 1D secondary structure, i.e., the prediction of the locations of the transmembrane helices is a much simpler problem than is the prediction of secondary structure for globular, soluble proteins. Some principle ideas of methods based on expert rules, information theory and neural networks will be sketched (3T-69/71).

*Evaluation of impact.* All prediction methods have a comparably high accuracy of about 95% (3T-72/74). However, this level is not sustained, as reliable data for locations of transmembrane helices exists for only a handful of proteins. Data used for training, e.g., neural networks stems from experiments in cell biology. Different authors often report different locations for transmembrane regions. Despite this warning the prediction of transmembrane helices is a valuable tool to quickly scan entire chromosomes. The sorting into membrane/not-membrane proteins has an expected error rate of less than 5% and can be useful for some experimental purposes.

*Transmembrane helix prediction: Literature*
*Methods:* (von Heijne 1981, Argos, et al. 1982, Engelman, et al. 1986, von Heijne & Gavel 1988, Park, et al. 1992, Edelman 1993, Sipos & von Heijne 1993, Jones, et al. 1994, Persson & Argos 1994, Taylor, et al. 1994, Rost, et al. 1995)

## Prediction of protein structure in 2D

**Prediction of (long-range) inter-residue contacts**

*Prediction of contacts.* From the knowledge of all inter-residue contact or distances (Fig. 3.2) one can, in principle, model a 3D structure using distance geometry methods (Havel, et al. 1983, Havel & Wüthrich 1984, Braun & Gö 1985, Brünger, et al. 1986, Havel 1991, Bohr, et al. 1993, Brünger & Nilges 1993, Nilges 1993, Nilges & Brünger 1993, Saitoh, et al. 1993). Two questions surround such methods: first, can contact be predicted accurately enough; and second, are all important contact predicted? A trade-off occurs between the Scylla of predicting enough contacts and the Charibdis of predicting only correct ones (Fig. 3.36).

*Prediction methods.* In sequence alignments, some pairs of positions appear to co-vary in a physico-chemically plausible manner (i.e. a 'loss of function' point mutation is often rescued by an additional mutation that compensates for the change (Altschuh, et al. 1987, Altschuh, et al. 1988). One hypothesis is that compensation would be most effective in maintaining a structural motif if the mutated residues were spatial neighbours. A method that uses correlated mutations for prediction of inter-residue contacts will be described, along with a neural network method predicting medium-ranged distances.

*Evaluation of impact.* Applying a stringent significance cut-off in the prediction of contacts by correlated mutations, a small number of residue contacts can be predicted with reasonable accuracy. Correlated mutations may provide sufficient information to distinguish between alternative models of 3D structure, but not enough information to predict conformations ab initio (Fig. 3.37/8). The success of the neural network predictions of contacts are difficult to assess. The general conclusion is that prediction of inter-residue contacts of tremendous potential value, but so far of rather limited accuracy.

*Prediction of inter-residue contacts: Literature*
*Correlated mutations:* (Altschuh, et al. 1987, Altschuh, et al. 1988, Finkelstein, et al. 1993, Finkelstein & Nakamura 1993, Gerstein, et al. 1994, Goebel, et al. 1994, Neher 1994, Shindyalov, et al. 1994, Taylor & Hatrick 1994)
*Other methods:* (Galaktionov & Rodionov 1980, Bohr, et al. 1993, Saitoh, et al. 1993, Galaktionov & Marshall 1994)

**Prediction of contacts between beta-strands**

*Prediction methods.* One simplification of the problem to predict inter-residue contacts is to specifically predict the contacts between residues in beta-strands, i.e., to predict the conformations of sheets. The only method applied to do so is based on data based derived potentials.

*Evaluation of impact.* Prediction of inter-strand contacts is possible if the locations of the strands are known. Given the error of current prediction methods, the accuracy in predicting inter-strand contacts drops, but in some cases is still high enough to be useful for modelling 3D structure.

*Prediction of inter-strand contacts: Literature*
(Hubbard 1994, Hubbard & Park 1995)

**Prediction of disulphide bonds**

*Prediction methods.* A more extreme simplification of the problem to predict inter-residue contacts is to only predict disulphide-contacts. These give the most dominant signal for methods predicting inter-residue contacts based on mean-force potentials (Valencia et al., unpublished). Thus, a prediction of disulphide bonds may be useful for other contact prediction methods. Here, we sketch the prediction of contacts between two cysteines and cysteines and other residues by a neural network.

*Evaluation of impact.* Prediction accuracy is claimed to be in the range of 80% which appears to be rather high. However, the evaluation of the usefulness of the tool is made difficult by the too small test set used.

*Prediction of disulphide bonds: Literature*
(Muskal, et al. 1990)

## Prediction of protein structure in 3D

**Sequence alignment THE prediction tool**

*Reason for success.* At the level of protein molecules, selective pressure results from the need to maintain function, which in turn requires maintenance of the specific 3D structure (Doolittle 1986, Farber & Petsko 1990, Pastore & Lesk 1990, Doolittle 1994) This is the base for attempts to align protein sequences, i.e., to optimally superpose 1D strings of amino-acid letters. Accordingly, conservation and mutation patterns observed in alignments contain very specific information about 3D structure. How much variation is tolerated? Two naturally evolved proteins with more than 25% identical residues (length > 80 residues) are extremely likely to be similar in 3D structure (Fig. 1.11). Even so, structure may be conserved in spite of much higher divergence (Holm & Sander 1994a). Do we have enough data to detect structure-specific sequence motifs (Rooman & Wodak 1988) and to correctly align very remote homologues?

*Methods.* The basic procedure of dynamic programming is rather straightforward. Although, the principle tool requires fine-tuning of parameters such as gap-open penalty, the tool is rather robust under the variation of free parameters. For more sensitive searches, biological knowledge has to be included by basing the alignment on profiles for residue exchange probabilities.

*Evaluation of impact.* When sequence similarity is sufficient, alignment procedures are (more or less) straightforward (Sander & Schneider 1991, Jones, et al. 1992b, Flores, et al. 1993). For less similar protein sequences, however, alignments may fail (Bordo 1993, Henikoff & Henikoff 1993, Bordo, et al. 1994, Vingron & Waterman 1994). The art of sequence alignment is to accurately align related sequence segments and to avoid aligning unrelated sequence stretches (Higgins & Sharp 1988, Higgins & Sharp 1989, Altschul 1991, Sander & Schneider 1991, Deperieux & Feytmans 1992, Higgins, et al. 1992, Russell & Barton 1992, Altschul 1993, Haussler, et al. 1993, Henikoff & Henikoff 1993, Heringa & Argos 1993, Johnson, et al. 1993, Lawrence, et al. 1993, Livingstone & Barton 1993, Henikoff & Henikoff 1994, Krogh, et al. 1994, Thompson, et al. 1994). Alignment techniques can easily be improved by incorporating information derived from 3D structures (Henikoff & Henikoff 1993).

*Sequence alignment: Literature*

*Methods (only basic and recent methods listed):* (Needlman & Wunsch 1970, McLachlan 1971, Smith & Waterman 1981, Waterman 1983, Gribskov, et al. 1987, Pearson & Lipman 1988, Taylor 1988, Higgins & Sharp 1989, Vingron & Argos 1989, Altschul, et al. 1990, Bacon & Anderson 1990, Smith, et al. 1990, Smith & Smith 1990, Henikoff 1991, Sander & Schneider 1991, Vingron & Argos 1991, Alexandrov 1992, Deperieux & Feytmans 1992, Higgins, et al. 1992, Altschul 1993, Henikoff & Henikoff 1993, Heringa & Argos 1993, Johnson, et al. 1993, Lawrence, et al. 1993, Livingstone & Barton 1993, Krogh, et al. 1994, Thompson, et al. 1994, Vingron & Waterman 1994)

*Methods (hashing):* (Dumas & Ninio 1982, Wilbur & Lipman 1983, Lipman & Pearson 1985, Pearson & Lipman 1988, Altschul, et al. 1990, Karlin & Altschul 1990, Karlin, et al. 1990, Altschul 1991, Altschul 1993)

*Methods (profile based):* (Higgins & Sharp 1988, Higgins & Sharp 1989, Altschul 1991, Sander & Schneider 1991, Deperieux & Feytmans 1992, Higgins, et al. 1992, Russell & Barton 1992, Altschul 1993, Haussler, et al. 1993, Henikoff & Henikoff 1993, Heringa & Argos 1993, Johnson, et al. 1993, Lawrence, et al. 1993, Livingstone & Barton 1993, Henikoff & Henikoff 1994, Krogh, et al. 1994, Schneider 1994, Thompson, et al. 1994)

**Homology modelling**

*Prediction methods.* The principle idea is to model the structure for SOUS (protein of unknown structure) based on the template of a known homologue, say KNOWN. To make this possible,

first one has to find a known structure in the data base that has a significant level of pairwise sequence identity to SOUS. The basic assumption is that KNOWN and SOUS have identical backbones (Fig. 1.3). The task is to correctly place the side chains of SOUS into the backbone given by KNOWN. Here, we shall briefly describe methods that make use of rotamer libraries (Fig. 3.54).

*Evaluation of impact.* The accuracy of homology modelling depends on the level of pairwise sequence identity between SOUS and KNOWN (Fig. 3.53). For higher levels, homology modelling is as accurate as is experimental determination of structure. However, even down to levels of some 25-30% sequence identity, homology modelling produces relatively accurate models about the fold of proteins of unknown structure.

*Homology modelling: Literature*

*Methods:* (Greer 1980, Jones & Thirup 1986, Blundell, et al. 1988, Summers & Karplus 1989, Overington, et al. 1990, Sali, et al. 1990, Greer 1991, Johnson 1991, Vriend & Sander 1991, Holm & Sander 1992b, Lesk & Boswell 1992, Levitt 1992, Overington, et al. 1992, Overington 1992, Johnson, et al. 1993, Vriend & Eijsink 1993, Abagyan & Totrov 1994, Abagyan, et al. 1994, Holm, et al. 1994, May & Blundell 1994, May & Johnson 1994, Sali & Blundell 1994, Totrov & Abagyan 1994)

*Quick data base scan:* (Bryant 1989, Islam & Sternberg 1989, Vriend 1990)

*Rotamer libraries:* (Ponder & Richards 1987, Summers & Karplus 1989, Karplus & Petsko 1990, Summers & Karplus 1990, Berendsen 1991, Cornell, et al. 1991, Holm & Sander 1992a, Levitt 1992, Eisenmenger, et al. 1993, Vriend & Sander 1993, De Filippis, et al. 1994)

*Reviews:* (Johnson 1991, Lesk & Boswell 1992, Overington 1992, May & Blundell 1994)

**Potentials of mean-force**

*Prediction methods.* A sufficiently valid working hypothesis is that protein sequence determines protein structure. Thus, in principle structure could be determined based on quantum mechanical principles. The problem is made hopelessly difficult by the size of the search space. One way around the limitations of inductive force-fields is a deductive approach, i.e., to derive an energy from knowledge contained in the data base. Here, one such potential of mean-force will be described in detail.

*Evaluation of impact.* Mean-force based potentials were successfully applied to predict errors in experimentally determined 3D structures. The sensitivity of such potentials is far beyond the mere statement that a certain structure contains errors: stresses in certain regions can be spotted and different models derived from refinement procedures can be distinguished.

*Potential of mean-force: Literature*

*Methods (basics):* (Sippl 1990, Sippl & Lackner 1992, Sippl 1993a, Sippl 1993b)

*Methods (further):* (Hendlich, et al. 1990, Casari & Sippl 1992, Sippl, et al. 1992, Sippl & Weitckus 1992, Sippl & Jaritz 1994, Sippl, et al. 1994, Sippl, et al. 1994, Flöckner, et al. 1995)

*Other knowledge-based potentials for quality control:* (Holm & Sander 1992a, Laskowski, et al. 1993, Vriend & Sander 1993)

*Semi-empirical potentials:* (Momany, et al. 1975, Brünger, et al. 1986, Brooks, et al. 1988, van Gunsteren 1988, Karplus & Petsko 1990, van Gunsteren 1993)

**Remote homology modelling (threading)**

*Remote homology.* All naturally evolved sequences with more than 30% pairwise sequence identity are homologous. However, not all with less than 25% are non-homologous. Instead, there are some thousands of pairs of structurally homologous pairs of proteins with less than 25% sequence identity (remote homologues) known (Holm & Sander 1994a). The principle objective of threading techniques is to detect such pairs and to generate alignments accurate enough to model 3D structure based on a profile to a remote homologue of known structure.

*Methods:* The principle concept of most threading method is to derive potentials that describe the fitness of a sequence for a given structure. Some potentials will be sketched.

*Evaluation of impact.* One problem with evaluating threading techniques is that their accuracy has often been over-estimated. Furthermore, hardly any method had been tested on a larger data set. Instead, so far all methods have been evaluated on a small set of favourable cases. What makes the situation even worse, is the confusion of 'fold recognition' and '3D prediction'. The conclusion from a 'prediction experiment' summarised in a meeting in Asilomar, C.A., Dec., 1994 was that threading techniques do recognise the correct fold in less than 50% of the cases and do result in correct alignments (that could be used for 3D modelling) in some cases (Shortle 1995). As frustrating as this result may sound, threading techniques may still become one of the most successful tools in structure prediction.

*Remote homology modelling: Literature*

Methods (intra-molecular potentials): (Novotny, et al. 1984, Novotny, et al. 1988)

Methods (volume computation): (Gregoret & Cohen 1990, Gregoret & Cohen 1991)

Methods (empirical solvent accessibility terms): (Eisenberg & McLachlan 1986, Baumann, et al. 1989, Chiche, et al. 1990, Holm & Sander 1992a)

Methods (contact energies): (Tanaka & Scheraga 1975, Crippen 1977, Lifson & Sander 1979, Galaktionov & Rodionov 1981, Miyazawa & Jernigan 1985, Miyazawa & Jernigan 1993)

Methods (contact potentials optimised to place native structure in global minimum): (Crippen 1991, Maiorov & Crippen 1992, Crippen & Maiorov 1994, Maiorov & Crippen 1994)

Methods (self-consistent hydrophobic force-field): (Finkelstein & Reva 1991, Finkelstein & Reva 1992)

Methods (environment specific preferences): (Bowie, et al. 1990, Overington, et al. 1990, Bowie, et al. 1991, Eisenberg, et al. 1991, Lüthy, et al. 1991, Lüthy, et al. 1992, Overington, et al. 1992, Blundell & Johnson 1993, Ouzounis, et al. 1993, Taylor 1993, Wilmanns & Eisenberg 1993)

Methods (mean-force (or Sippl) potentials): (Hendlich, et al. 1990, Sippl 1990, Casari & Sippl 1992, Jones, et al. 1992a, Sippl & Weitckus 1992, Bryant & Lawrence 1993, Nishikawa & Matsuo 1993, Bauer & Beyer 1994, Sippl & Jaritz 1994, Sippl, et al. 1994, Flöckner, et al. 1995, Koehl & Delarue 1995)

Methods (other): (Taylor & Orengo 1989, Taylor 1991, Godzik, et al. 1992, Godzik & Skolnick 1992, Goldstein, et al. 1992, Rost & Sander 1992b, Stultz, et al. 1993, Topham, et al. 1993, Abagyan, et al. 1994, Goldstein, et al. 1994, Lathrop & Smith 1994, Rost 1995a, Rost 1995c)

Reviews: (Wodak & Rooman 1993, Shortle 1995)

(Bork & Grundwald 1990, Hirst & Sternberg 1991, Nayal & Di Cera 1994, Villar & Kauvar 1994)

(Bork, et al. 1992b, Bork, et al. 1992a, Bork, et al. 1994, Johnston, et al. 1994)



potentials of mean force: threading

# Computational tools for experimental determination and theoretical prediction of protein structure

- Introduction: proteins the complex machinery of life
- Experimental determination of protein structure
- Prediction of protein structure



# Prediction of protein structure

- Overview:
  - Prediction of structure and function, where are we now?
- Evaluation of prediction methods
  - How to choose the data set? Why cross-validation?
- Prediction of protein structure in 1D
  - secondary structure; solvent exposure; transmembrane helices
- Prediction of protein structure in 2D
  - inter-residue contacts; inter-strand contacts; disulphide bonds
- Prediction of protein structure in 3D
  - multiple sequence alignments; homology modelling; potentials of mean force; threading
- Prediction of protein function
  - sequence motifs; binding sites

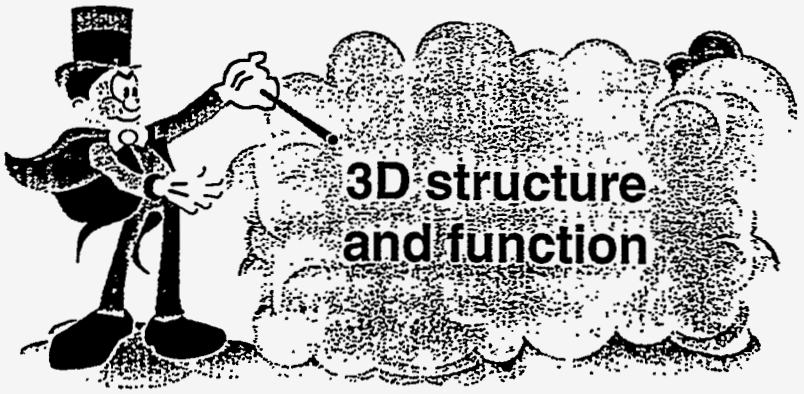
# Overview

- What is the state of the art in structure prediction?
  Fig. 3.1
- How can the prediction problem be simplified?
  Fig. 3.2
- Which prediction is of interest for molecular biology?

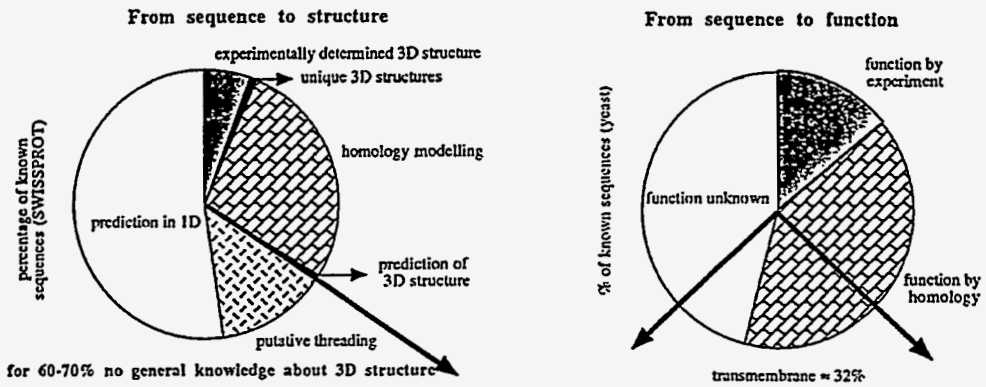


# Goal of prediction

- Epstein, Anfinsen 1961:
  sequence uniquely determines structure
  =>
- Input: protein sequence
- Output:



# Fig. 3.1: State of the prediction art






# Fig. 3.2: 3D, 2D, 1D

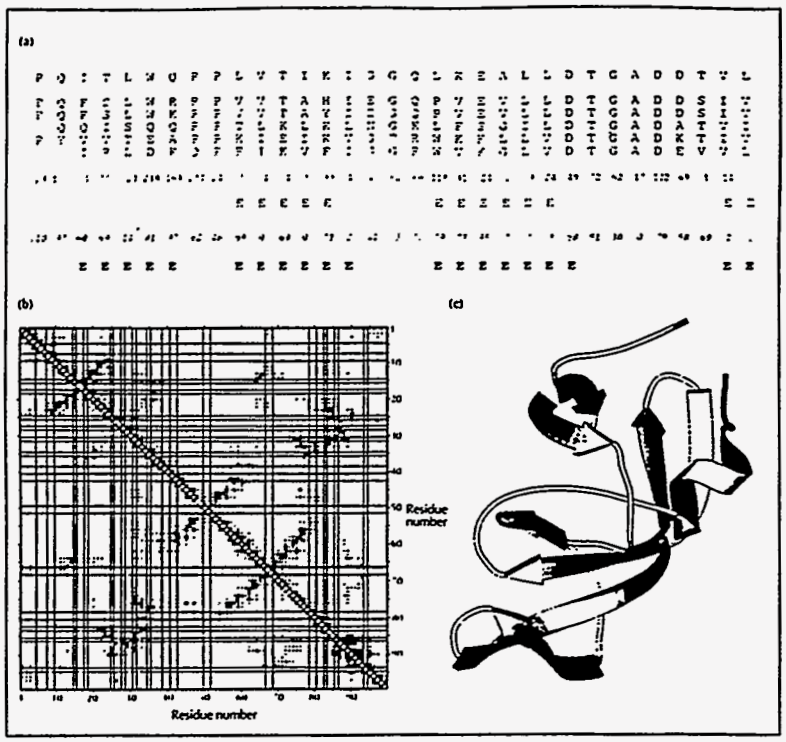



Fig. 1. Representation of HIV-1 protease monomer (Protein Data Bank code 1HHP) in one, two and three dimensions. Each of the representations gives rise to a different type of prediction. (a) One-dimensional representation, exemplifying prediction of secondary structure and solvent accessibility. The first row shows the first 33 residues of the HIV-1 protease. In the block below, the alignment exemplified by five sequences is shown. In the row below this block, solvent accessibility (measured in Å²) is shown for HIV-1 protease [108]. The next row shows secondary structure for HIV-1 (strand is indicated by an E; neither helix nor strand is indicated by a blank) [108]. The penultimate row (italic) and the last row (bold) show a typical prediction by the neural network program PHD [36] for solvent accessibility (measured in Å²) and secondary structure (indicated as above), respectively. Amino acids are given in the one-letter amino acid code. (b) Two-dimensional representation for the prediction of the inter-residue contact map. The three-dimensional structure is projected onto a two-dimensional matrix of inter-residue contacts (inter-residue distances can also be used). The contact strength at each position in the matrix is indicated by the depth of shading in each cell. Horizontal and vertical lines show borders of secondary structure segments. Matrix provided by courtesy of

## Evaluation of prediction methods

- Publishing optimistic results?
- What is the goal and which limits are to be expected?
- How to choose the data set?
- How many proteins to use for the test set?
- Optimising free parameter with respect to the test set?
- How many cross-validation experiments have to be performed?
- Enough of testing?
- How to measure performance accuracy?



## Evaluation of prediction methods

- Publishing optimistic results?
  - An inaccurate prediction is not as bad, as an over-estimated one.
  - Even a prediction method of limited accuracy can be useful if the user knows what to expect.

- What is the goal and which limits are to be expected?
  - » best alternative prediction?
  - » worst prediction (random)?
  - » how accurate are existing prediction methods?

    Fig. 3.3

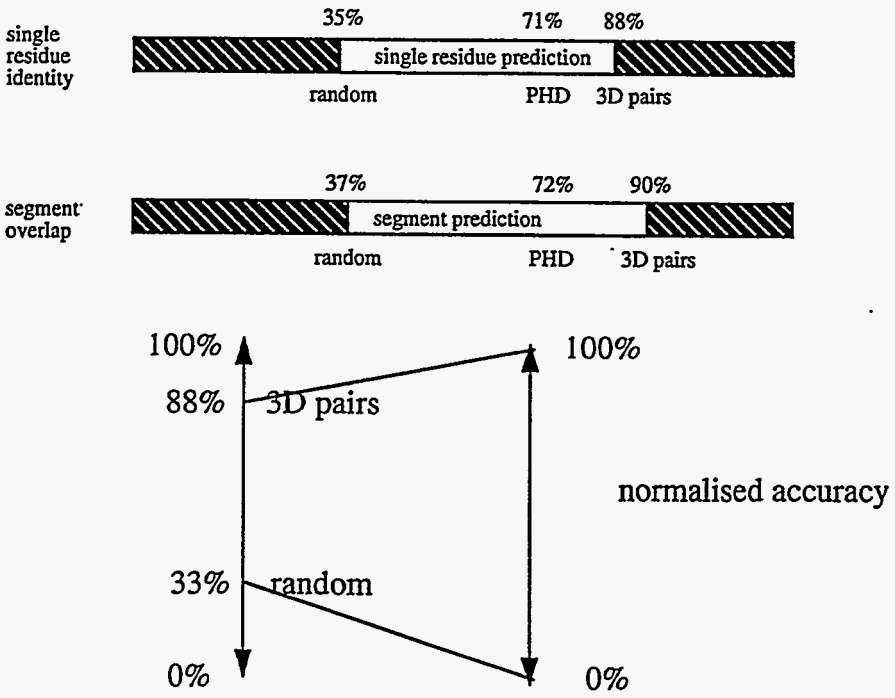
- How to choose the data set?
  - » in general, to be decided with respect to 'best alternative' secondary structure: pairwise sequence identity < 25%

    Fig. 3.4
  - » cross-validation

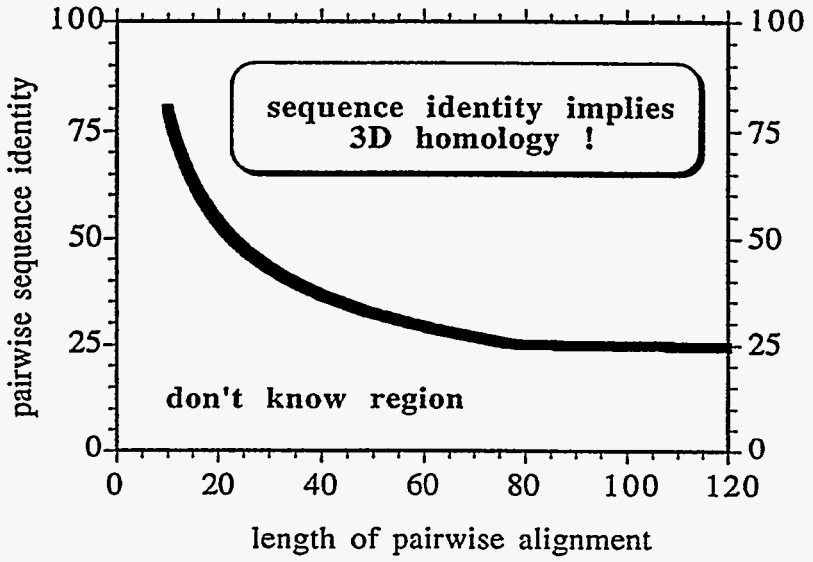
    Fig. 3.5




## Fig. 3.3: Best/worst prediction scale

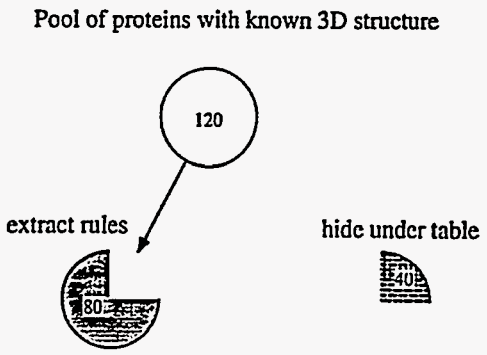


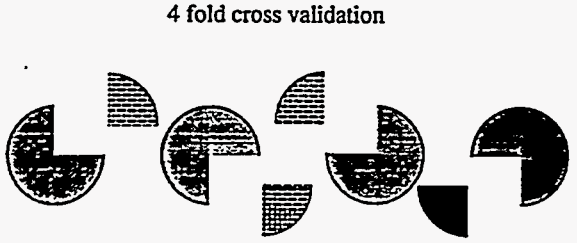
## Fig. 3.4: Significant sequence identity






## Fig. 3.5: Cross-validation



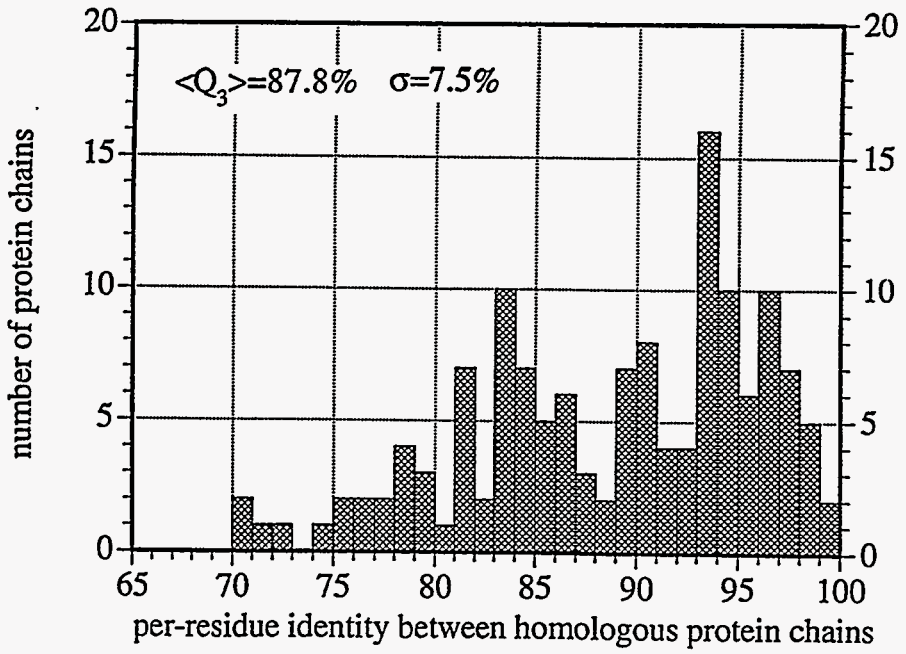
# Evaluation of prediction methods

- **How many proteins to use for the test set?**
  - » as many as possible, but...
  - » features of proteins are distributions, i.e., vary between different proteins

    Fig. 3.6
  - » => at least as many to mirror this variance
  - » rule of thumb; choose number of proteins N such that: $\sigma_N \approx \sigma_{2N}$ i.e. doubling test set => same result
- **Optimising free parameter with respect to test set?**
  - » optimise free parameters BEFORE cross-validation experiment is performed
- **How many cross-validation experiments have to be performed?**
  - » 2 × 150:150 ?
  - » 300 × 299: 1 ?
  - – **No difference with respect to generality of results**



# Fig. 3.6: Variance between different proteins



## Evaluation of prediction methods

- **Enough of testing?**
  - » 'pre-release test'
    ideally after manuscript has been written
- **How to measure performance accuracy?**
  - » what is the goal of the method?
    - e.g. prediction of secondary structure
  - » which measure best to describe goal?
    - per-residue: information; accuracy for helix, strand (%obs,%pred)
  - » which measure best to reflect biological reality of goal?
    - per-segment: optimise by structural comparisons
  - » which standard deviation is to be expected?
    - variance of accuracy with protein chain

– **NOTE: the expected variation may not necessarily follow from statistics based on the test set!**

prediction helix/non-helix, based on test set of 10 proteins may result in an estimate of ± 3% for the standard deviation, however from three-state predictions, it is known that the correct value is rather in the order of ± 10%



## Prediction of protein structure in 1D

- **Secondary structure prediction**
- **Solvent accessibility prediction**
- **Transmembrane helix prediction**

# Secondary structure prediction

- Goal and concept
- Methods
  - Statistics
  - Neural networks
  - Nearest neighbour algorithms
  - Break-through by using evolutionary information
- Results
  - » Measures for accuracy
  - » three-state accuracy > 72%
- Further methods
  - Prediction of secondary structure content
  - Prediction of secondary structure in two states
- Applications
  - » Post-processing prediction methods; chain tracing; mutational experiments; speculations about binding sites and function



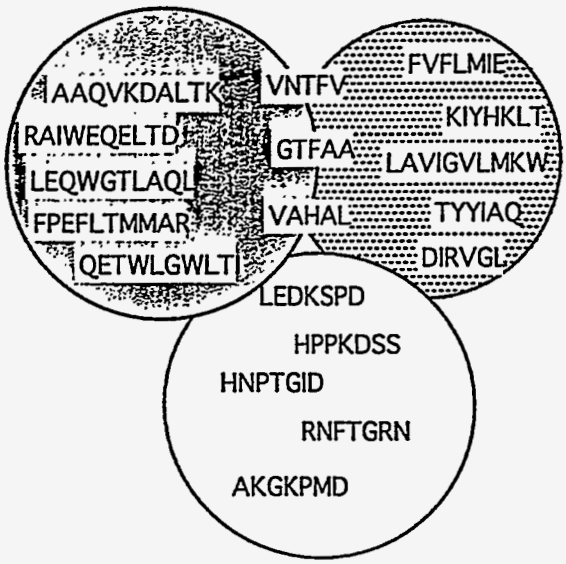
# The simple prediction problem: 1D secondary structure

- Basic idea:
  classification by similarity to known cases
  Fig. 3.7
  - » pentapeptides not unique, ...
  - » but, longer peptides are!
- screening secondary structure of central residue in a window of $w$ adjacent residues
  - typical values for w = 1-21
  Fig. 3.8

# Pattern classification



# Fig. 3.7: Classification by residue patterns




**Secondary structure prediction as a pattern recognition problem:** Certain oligopeptides have high preference to be in a particular secondary structure. Circles: upper left (dark shading): helix, upper right (light shading): strand, centre (no shading): loop. The 3 pentapeptides between the helix and strand circles are observed in both structures.

## Fig. 3.8: Central-residue screening




## Secondary structure prediction methods

- **Information theory**
  - » principles
  - » application to secondary structure prediction
- **Neural network**
  - » principles
- **Neural network**
  - » simple solution for secondary structure prediction
- **Neural network**
  - » problem specific adaptation
- **Nearest neighbour algorithm**
  - » principles
  - » application to secondary structure prediction
- **Break-through by using evolutionary information**
  - » information contained in evolutionary exchange patterns
  - » implementation of information into NN

# Secondary str. prediction by information theory

- **principle:**

(Robson & Paine, 1971; Garnier et al., 1978; Gibrat et al., 1987)

state S, one residue R:

$$I(S;R) = \log\left[\frac{p(S \mid R)}{p(R)p(S)}\right] \qquad p(S \mid R) = \frac{p(S,R)}{p(R)}$$

I(S,R), information of residue R in state S; p(S,R), probability of observing residue R in state S; p(R), probability of finding residue R; p(S), probability of finding state S

states S, S', one residue R:

$$I(S;R) - I(S';R) = I(S:S';R) = \log\left[\frac{p(S \mid R)}{p(S' \mid R)}\right] + \log\left[\frac{p(S')}{p(S)}\right]$$

I(S:S';R) information difference of residue R in states S and S'

states S, S', (2m+1) residues R:

$$I(S:S';R_{-m}...R_m) = \log\left[\frac{p(S \mid R_{-m}...R_m)}{p(S' \mid R_{-m}...R_m)}\right] + \log\left[\frac{p(S')}{p(S)}\right]$$

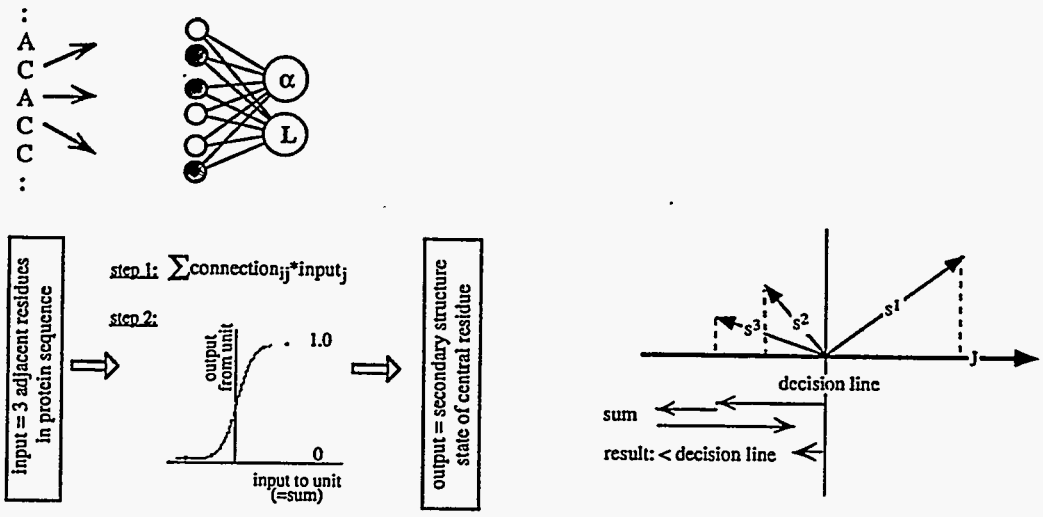
$$\approx I(S:S';R) + \sum_{j=-m, j\neq 0}^{j=+m} \log\left[\frac{p(S \mid R_j.)}{p(S' \mid R_j)}\right] + \log\left[\frac{p(S')}{p(S)}\right]$$

prediction = min (I(S:S';R)

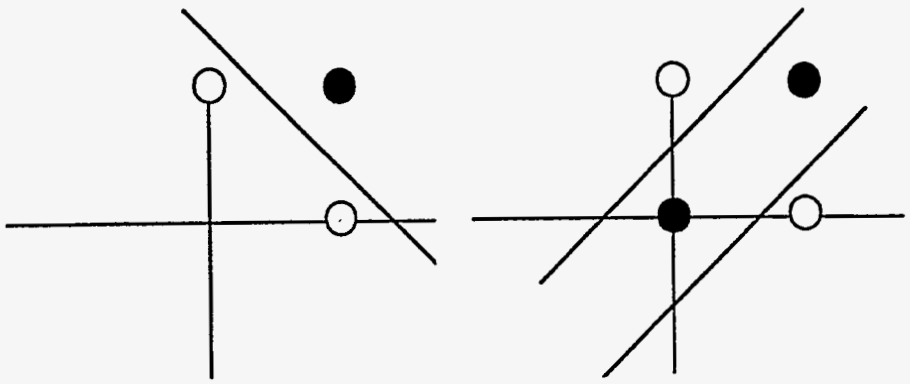


# Principles of neural networks: input -> output

» two steps:
1. linear: sum over all input × connection
2. non-linear: sigmoid trigger, i.e., project sum onto 0-1



## Fig. 3.9: Pattern classification by NN





## Principles of neural networks: error

– **output:**

$$out_i = \sum_{i=1}^{N^{in+1}} J_{ij}\, in_j$$

$in_j$ value of input unit $j$ ; $out_i$ value of output unit $i$ ;
$J_{ij}$ connection between input unit $j$ and output unit $i$

– **error:**

$$E = \sum_{i=1}^{N^{out}} (out_i - des_i)^2$$

$out_i$ value of output unit $i$ ; $des_i$ secondary structure state observed for central amino acid for output unit $i$ (e.g. for a helix: $des_1$=1, $des_2$=0, $des_3$=0)

– **free variables:** **connections { J }**

– **goal:**

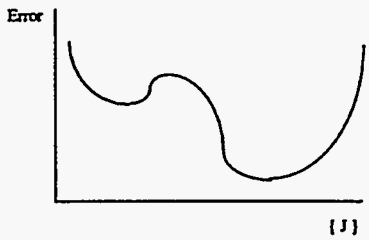
» representation of set of examples (training set) for which the mapping input->output is known, i.e., the secondary structure state of the central residue has been observed by the network

# Principles of neural networks: training

- training = change of connections $\{J\}$ such that $E$ decreases
- simplest procedure:
  - gradient descent

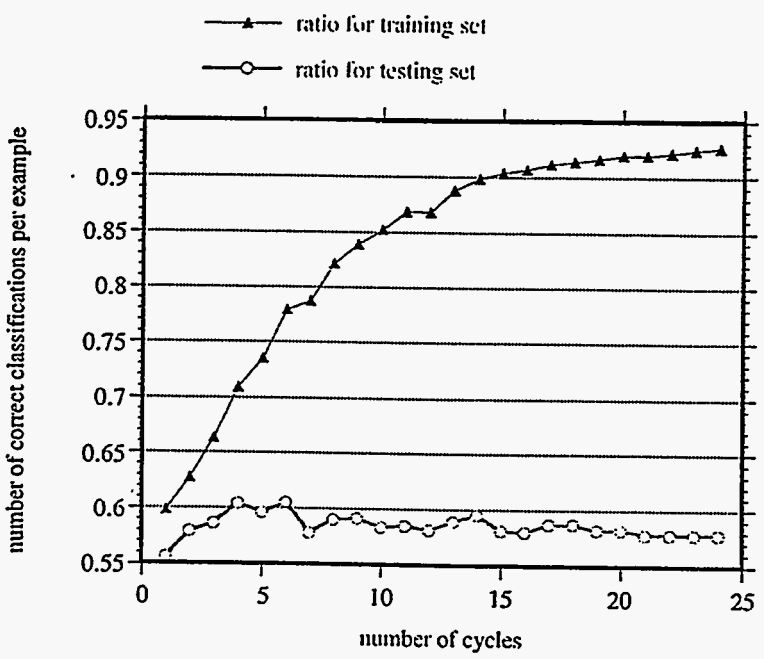
$$\Delta J_{ij}(t+1) = -\varepsilon \frac{\partial E(t)}{\partial J_{ij}(t)} + \alpha \Delta J_{ij}(t-1)$$

where $\partial E/\partial J$ is the derivative of the error with respect to the network connection; $t$ is the algorithmic time given by the presentation of one example; $\varepsilon$ determines the step width of the change (learning strength, typically some 0.01); $\alpha$ gives the contribution of the momentum term ($\Delta J(t-1)$, typically some 0.2), which permits uphill moves






# Fig. 3.10: Effect of overtraining



# Secondary str. prediction by neural networks

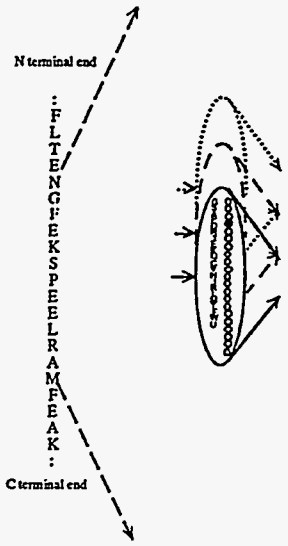
- input/output coding

  Fig. 3.9
- adapting the tool to the problem
  - balanced training
  - second level of networks
  - jury decision

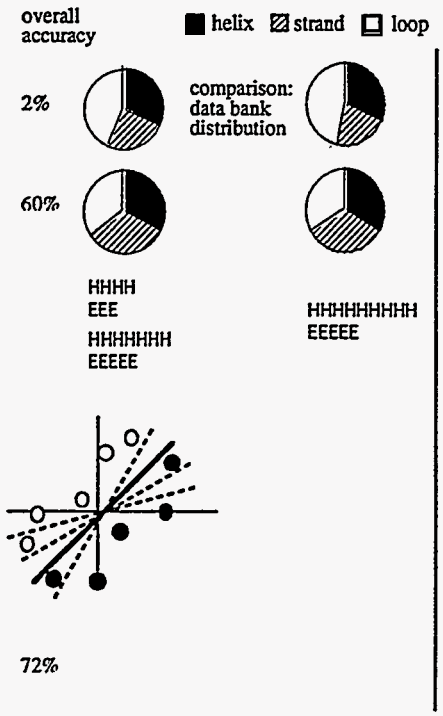
  Fig. 3.10



# Fig. 3.11: Simple NN for sec str pred



## Fig. 3.12: Adapting NN's to the problem






## Nearest neighbour algorithms: principle

- **principle idea: similarity to known structures**

(Kabsch & Sander, 1983b; Levin et al., 1986; Schneider, 1989; Zhang et al., 1992; Yi & Lander, 1993; Solovyev & Salamov, 1995)

| | |
|---|---|
| STNKDWW | unknown structure |
| KSNPDWW<br>EHQGEWW<br>RSTGDWW | known structures |

$$D\,(R_1^a \ldots R_w^a, R_1^b \ldots R_w^b) = \sum_{i=1}^{w} D\,(R_i^a, R_i^a)$$

$D\,(R_i^a, R_i^b)$, is the distance (or similarity) between the residues at position $i$ for the two strings $a$ and $b$

## Nearest neighbour algorithms: distances

- problem: what is similar?
- solutions
  - Hamming distance:
    - » equal residues: `D(R,R) = 1`
    - » different residues : `D(R,R') = 0`
  - Dayhoff matrix

    (Zhang et al., 1992)

$$D(R_i^a, R_i^b) = \frac{1}{w}\sum_{j=1}^{w} \left| p(S_j \mid R_i^a) - p(S_j \mid R_i^b) \right|$$

$$+ \frac{1}{w \cdot w \cdot 20}\sum_{j=1}^{w}\sum_{k=1}^{w}\sum_{h=1}^{20} \left| p(S_j \mid R_i^a, x_k^h) - p(S_j \mid R_i^b, x_k^h) \right|$$

$x_k^h$ denotes amino acid $x^h$ at window position $k$;



## Nearest neighbour algorithms: potentials

(Solovyev & Salamov, 1995)

- compute distances based on 'fitness-of-sequence-for-structure' potentials

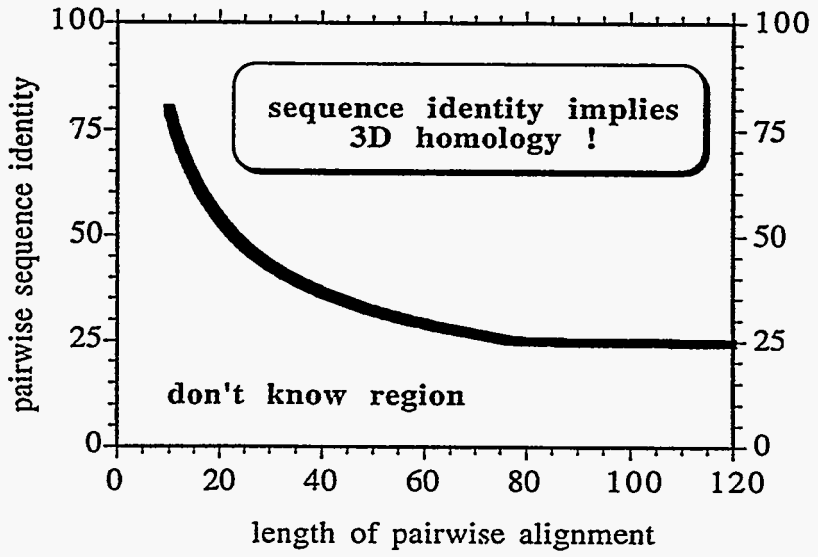
  (Bowie et al., 1990; Bowie et al., 1991; Ouzounis et al., 1993)
- distinguish between helix core, helix N- and C-term
- restrict list of possible similar segments by information theory
- balance statistics
- include evolutionary information

## Break-through by using evolutionary information

- Problem:
  different algorithms yield only marginal differences in prediction accuracy
- Reason:
  only local information processed, but secondary structure formation is strongly determined by non-local interactions
- Way out:
  » increase window size
  not possible, ultimately as not enough patterns in database
  » then, what?
- Evolution has it!



Fig. 3.13: Evolution has it!



## Fig. 3.14: Processing information from multiple alignments

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| secondary structure | DSSP | E | | | | | | E | E | E | E | E | E | | | E | E | E | E | E | E | H | H | H |
| | SB | N | S | T | N | K | D | W | W | K | V | E | V | N | D | R | Q | G | F | V | P | A | A | Y |
| alignment | s1 | N | K | S | N | P | D | W | W | E | G | E | L | N | G | Q | R | G | V | F | P | A | S | Y |
| | s2 | E | E | H | . | G | E | W | W | K | A | K | s | s | K | R | E | G | F | I | P | S | N | Y |
| | s3 | R | S | T | . | G | D | W | W | L | A | r | v | T | G | R | E | G | Y | V | P | S | N | F |
| | s4 | F | S | . | . | . | . | F | F | G | V | e | v | D | D | L | Q | V | F | V | P | P | A | Y |
| | V | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 | 60 | 0 | 0 | 0 | 0 | 20 | 20 | 60 | 0 | 0 | 0 | 0 |
| | L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 |
| | M | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | F | 20 | 0 | 0 | 0 | 0 | 0 | 20 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 60 | 20 | 0 | 0 | 0 | 20 |
| | W | 0 | 0 | 0 | 0 | 0 | 0 | 80 | 80 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 80 |
| profile | G | 0 | 0 | 0 | 0 | 50 | 0 | 0 | 0 | 20 | 20 | 0 | 0 | 0 | 40 | 0 | 0 | 80 | 0 | 0 | 0 | 0 | 0 | 0 |
| | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 40 | 0 |
| | P | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 20 | 0 | 0 |
| | S | 0 | 60 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 20 | 0 |
| | T | 0 | 0 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | H | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | R | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 60 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | K | 0 | 20 | 0 | 0 | 25 | 0 | 0 | 0 | 40 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Q | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| additional information | E | 20 | 20 | 0 | 0 | 0 | 25 | 0 | 0 | 20 | 0 | 60 | 0 | 0 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | N | 40 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 | 0 |
| | D | 0 | 0 | 0 | 0 | 0 | 75 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $N_{ins}$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $N_{del}$ | 0 | 0 | 1 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | GW | 1.0 | 0.8 | 0.7 | 0.8 | 0.6 | 1.1 | 1.5 | 1.5 | 0.8 | 0.9 | 1.0 | 0.7 | 0.7 | 0.9 | 0.9 | 0.7 | 1.5 | 1.0 | 1.2 | 1.5 | 0.9 | 0.7 | 1.5 |



## Per-residue measures for prediction accuracy

» full table **$A_{ij}$ = number of residues predicted to be in structure type j and observed to be in type i**

» The sums over the columns of *A* give the number of residues predicted to be in structure *i* :

$$a_i = \sum_{j=1}^{3} A_{ji}, \text{ for } i = \alpha, \beta, L$$

» The sums over the rows give the number of residues observed to be in structure *i* :

$$b_i = \sum_{j=1}^{3} A_{ij}, \text{ for } i = \alpha, \beta, L$$

» The sum over all elements of *A* is the number of residues in the data bank used, abbreviated by *b* :

$$b = \sum_{j=1}^{3} b_j = \sum_{j=1}^{3} a_j$$

» The percentages of residues correctly predicted to be in class *i* from all residues predicted to be in *i* are given by:

$$Q_i = Q_i^{\%obs} = \frac{A_{ii}}{b_i} * 100$$

» The percentages of residues correctly predicted to be in class *i* from all residues predicted to be in i are given by:

$$Q_i^{\%pred} = \frac{A_{ii}}{a_i} * 100$$

» Overall three-state accuracy
(correctly predicted residues/all residues):

$$Q_3 = \frac{\sum_{i=1}^{3} A_{ii}}{b} *100$$

» Matthews correlation:

$$C_i = \frac{p_i \cdot n_i - u_i \cdot o_i}{\sqrt{(p_i+u_i)\cdot(p_i+o_i)\cdot(n_i+u_i)\cdot(n_i+o_i)}}$$

with $p_i$ being the number of properly predicted residues in conformation i, $n_i$ the number of those correctly not assigned to structure i, $u_i$ the number of underestimated, and $o_i$ that of overestimated conformations.

» Information content:

$$I = 1 - \frac{\sum_{i=1}^{3} a_i * \ln a_i - \sum_{i,j=1}^{3} A_{ij} * \ln A_{ij}}{b * \ln b - \sum_{j=1}^{3} b_j * \ln b_j}$$

This information is related to the probability of deviation of table A from a random distribution:

$I = 0$, if: $A_{ij}=1/9$, for $i, j=1, 2, 3$
$I = 1$, if: $A_{ij}=0$ , for $i \neq j$ and $A_{ii}=b_i$, $i,j = 1, 2, 3$



**Fig. 3.15: Accuracy table**

| | net H | net E | net C | sum DS |
|---|---|---|---|---|
| DSSP H | 17407 | 1634 | 5421 | 24462 |
| DSSP E | 1242 | 10197 | 4793 | 16232 |
| DSSP C | 3623 | 4033 | 26551 | 34207 |
| sum Net | 22272 | 15864 | 36765 | 74901 |

| %obs | | | %pred | | |
|---|---|---|---|---|---|
| H | E | C | H | E | C |
| 71 | 6 | 22 | 78 | 10 | 14 |
| 7 | 62 | 29 | 5 | 64 | 13 |
| 10 | 11 | 77 | 16 | 25 | 72 |

## Per-segment measures for prediction accuracy

» average segment length:

$$\langle L_i \rangle = \frac{\text{sum of the lengths over all segments of structure i}}{\text{number of all segments of structure i}}$$

» distribution of segments

» loose overlap between segments

$$ov^{loose} = \frac{1}{N} \sum_{s} \Theta^{loose}(s_1, s_2) * len(s_1)$$

$\Theta = 1$ if helices or strands overlap by one half, and loops by at least 2 residues

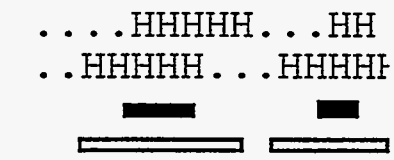
»

» optimised measure for segment overlap

$$Sov = \frac{1}{N} * \sum_{s} \frac{\min\{e(s_1); e(s_2)\} - \max\{b(s_1); b(s_2)\} + 1 + \delta}{\max\{e(s_1); e(s_2)\} - \min\{b(s_1); b(s_2)\} + 1} * len(s_1)$$

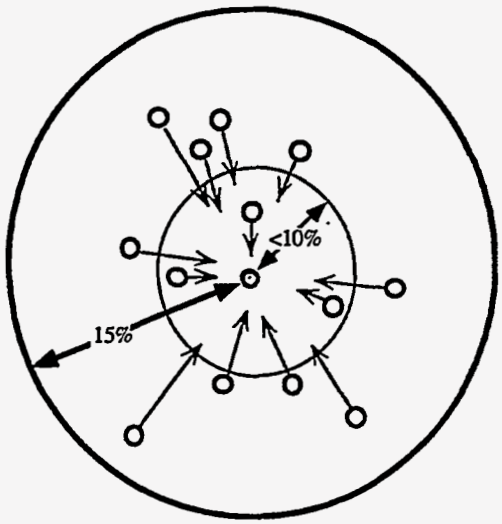


**Fig. 3.16: Per-segment measures**




$$Fov = \frac{3}{7} + \frac{2}{3}$$

## Fig. 3.17: Criterion for best segment measure






## Results of secondary str prediction

- Basic idea:

  classification by similarity to known samples
- Not as simple:

  accuracy in 3 states: helix, strand, rest ≈ 60%
- Improvement by:

  – new algorithms?
  – increase in number of known 3D structures?
  – more insight into protein folding?
- Projection from 3D onto 1D reduces information
  -> in search for more information

## Results of secondary str prediction

- Basic idea:

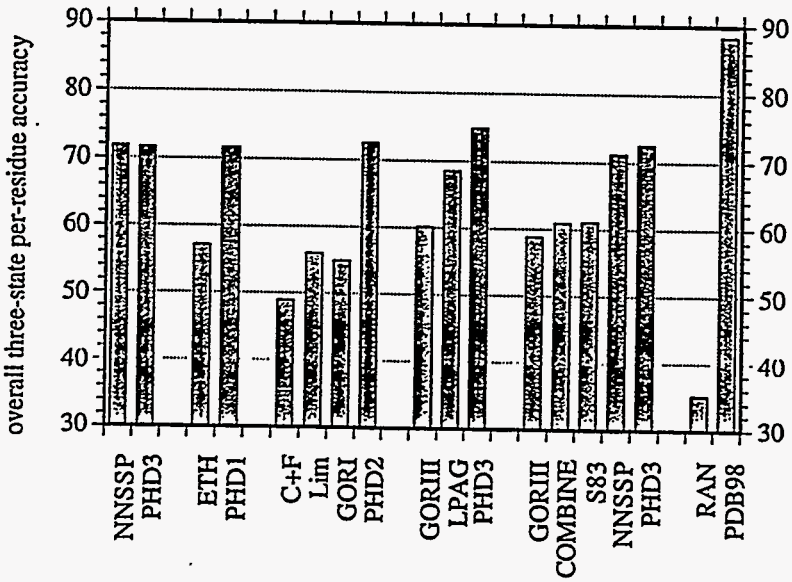
  classification by similarity to known samples
- Not as simple:

  accuracy in 3 states: helix, strand, rest ≈ *60%*
- Improvement by:

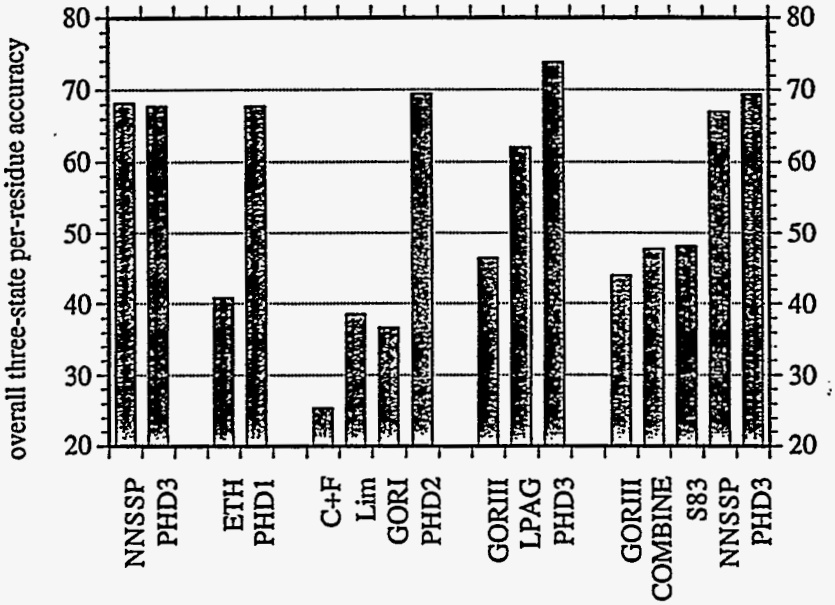
  – new algorithms?
  – increase in number of known 3D structures?
  – more insight into protein folding?
- Projection from 3D onto 1D reduces information
  -> in search for more information
- Evolutionary information pushes to > *70%*



## Fig. 3.18: Accuracy for various methods

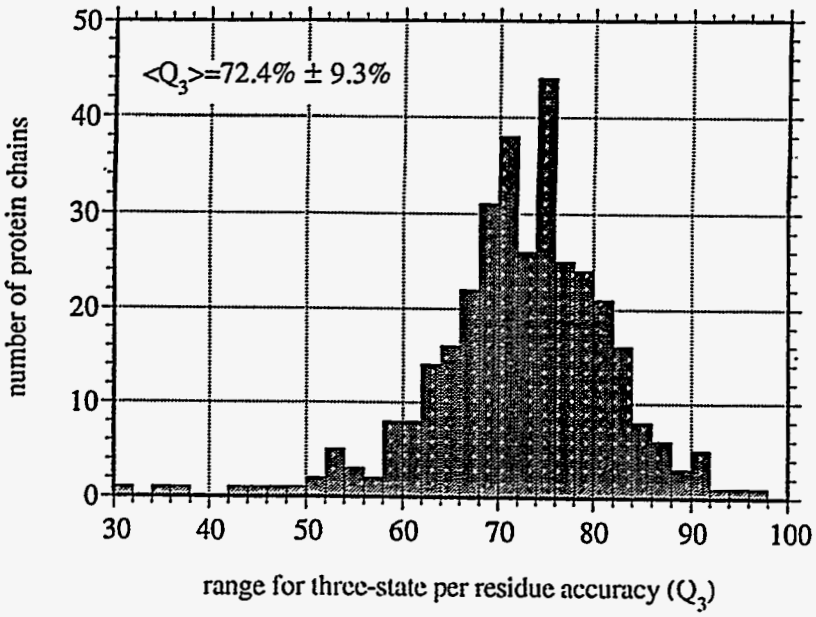


## Fig. 3.19: Normalised accuracy for various methods

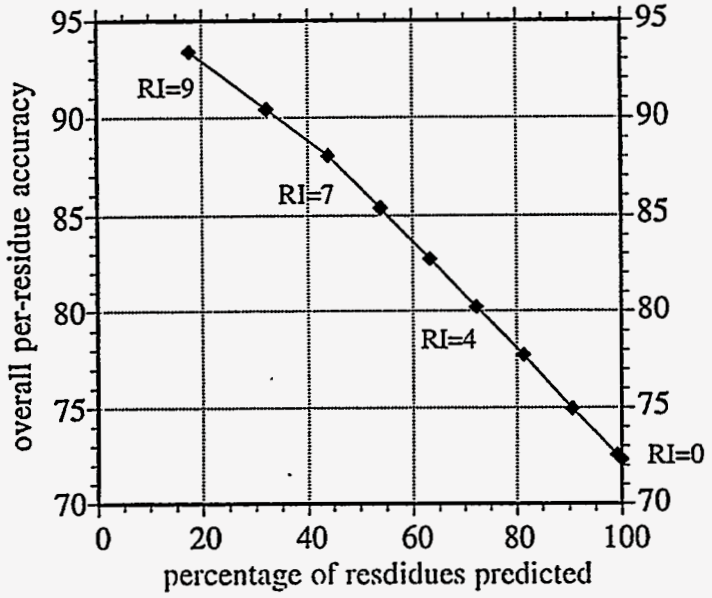





## Fig. 3.20: Distribution of prediction accuracy



## Fig. 3.21: Reliability of prediction

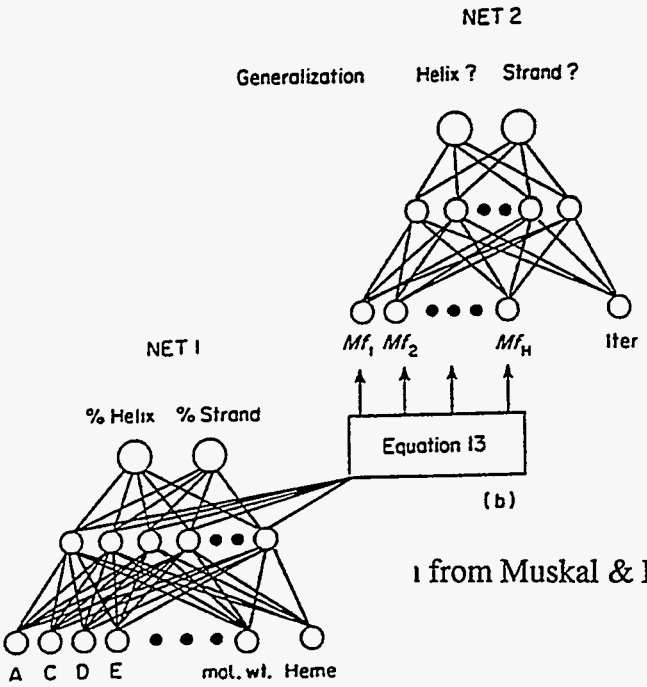





## Prediction of secondary structure content

- definition of structural content
- neural network specialists
- usual network (PHD)
- CD measurements

## Fig. 3.23: Tandem network to predict sec str content




ı from Muskal & Kim, 1992



## Fig. 3.22: Distinction of structural classes

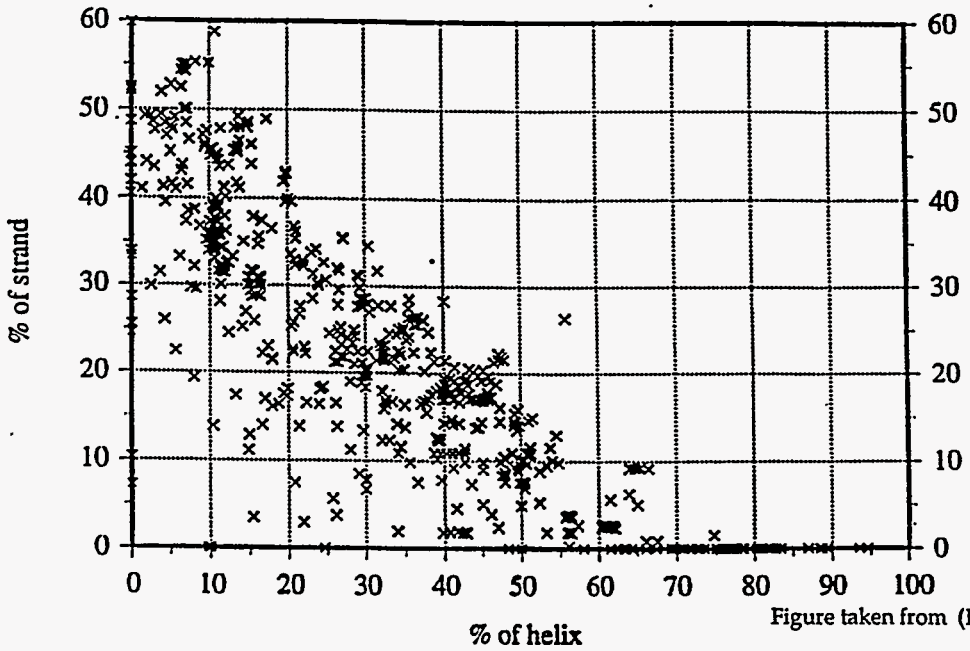



Figure taken from (Rost & Sander, 1994b)

## Fig. 3.24: Content prediction: experiment vs. theory

| Pearson correlation: | Nprot | helix | strand | quote from: |
|---|---|---|---|---|
| HM | 130 | 0.97 | 0.97 | Rost et al., 1994b |
| PHD | 124 | 0.91 | 0.73 | Rost & Sander, 1994b |
| COMBINE | 124 | 0.83 | 0.51 | Rost & Sander, 1994b |
| CD (Perczel et al., 1992) | 22 | 0.88 | <0.5 | Rost & Sander, 1993b |
| PHD | 22 | 0.86 | 0.88 | Rost & Sander, 1993b |



## Fig. 3.25: Accuracy in predicting sec str content

| method[a] | set[a] | Nprot[a] | Δhelix[b] | Δstrand[b] | All-α[c] | All-β[c] | α/β[c] | Rest[c] | $Q_{class}$[c] |
|---|---|---|---|---|---|---|---|---|---|
| HM:SeqAli | set 1 | 80 | 2.8±3.8 | 2.7±3.2 | 94.1 | 86.7 | 100.0 | 89.7 | 90.0 |
| RAN | set 1 | 80 | 32.1±20.8 | 21.3±14.5 | 0.0 | 0.0 | 0.0 | 71.2 | 44.7 |
| PHD | set 2 | 126 | 8.5±8.0 | 7.5±8.1 | 85.7 | 50.0 | 50.0 | 74.1 | 74.6 |
| PHD | set 3 | 124 | 7.8±6.8 | 7.3±7.9 | 94.1 | 0.0 | 55.6 | 74.5 | 75.8 |
| PHD | set 2-6 | 337 | 8.1±7.9 | 7.1±7.6 | 85.0 | 55.6 | 45.5 | 75.6 | 74.2 |

a See caption of Table I.

b Error in predicting the content of helix or strand averaged over all protein chains in the data set. The error is computed as the difference between the percentage of helix (Δhelix) or strand (Δstrand) between observed and predicted. ("±" values refer to one standard deviation).

c Percentage of protein chains correctly predicted in either of the four classes: all- α, all-β, α/β and all others. $Q_{class}$ gives the percentage of protein chains correctly predicted in any of the four classes.

## Prediction of helix/non-helix

- gain by specialising on one class

  (Maxfield & Scheraga, 1976; Taylor & Thornton, 1984; King & Sternberg, 1990; Kneller et al., 1992; Muggleton et al., 1992; Rost & Sander, 1993c)
- problem of most publications: too few examples
- results about 80% accuracy for helix/non-helix specialised prediction methods

  MKS (Muggleton et al., 1992): 80.5 %

  Helix network (Rost & Sander, 1993c): 82.7%
- marginally better than methods predicting 3 states

  PHD (Rost & Sander, 1993c): 81.2%
- BUT: inaccuracy in determining the class results in that specialists (two-state predictors) have on average lower prediction accuracy than, e.g., three-state predictors!
- MIND: two-state number not comparable to three-state numbers

  RAN (two states; Rost & Sander, 1993c): 54.5%

  RAN (three states; Rost et al., 1994b): 35.4%



## Resumé

- Evolution improves secondary structure prediction by 6-10 percentage points
- Neural networks are easy to be adapted to specific features of problems
- Prediction not perfect, but reasonably accurate
  - for 40% as good as homology modelling
  - well balanced
  - segments
- But:

  Goal is to predict 3D structure
- Evolution helpful to continue?

## Applications

- **Post-processing prediction methods**
  - » 3D modelling
  - » threading
  - » contact-predictions
- **Chain tracing**
- **Mutational experiments**
  - » change of secondary structure by exchange of residues, e.g., for finding (de) stabilising mutations
- **Speculations about binding sites and function**
  - » e.g. specific patterns, such as helix-turn-helix



## Solvent accessibility prediction

- **Goal and concept**
- **Methods**
  - – **Neural networks**
  - – **Statistics**
- **Results**
  - » Measures for accuracy
  - » Three-state accuracy < 60%
- **Evaluation**
  - » Accurate enough to seed predictions of secondary structure
  - » Not accurate enough to be as useful as secondary str. predictions
  - » Clear improvement by database growth (evolutionary information)
- **Applications**
  - » Post-processing prediction methods; Speculations about binding sites and function

## Accessibility of first hydration shell

- accessibility (DSSP) = 0-300 Å$^2$ *Acc*
- relative accessibility = 0-100 % *RelAcc*
- two-state modell :
  - *buried* <20%
  - *exposed* ≥20%
- three-state model:
  - *buried* < 5%
  - *intermediate* 5-20%
  - *exposed* ≥20%
- ten-state model :

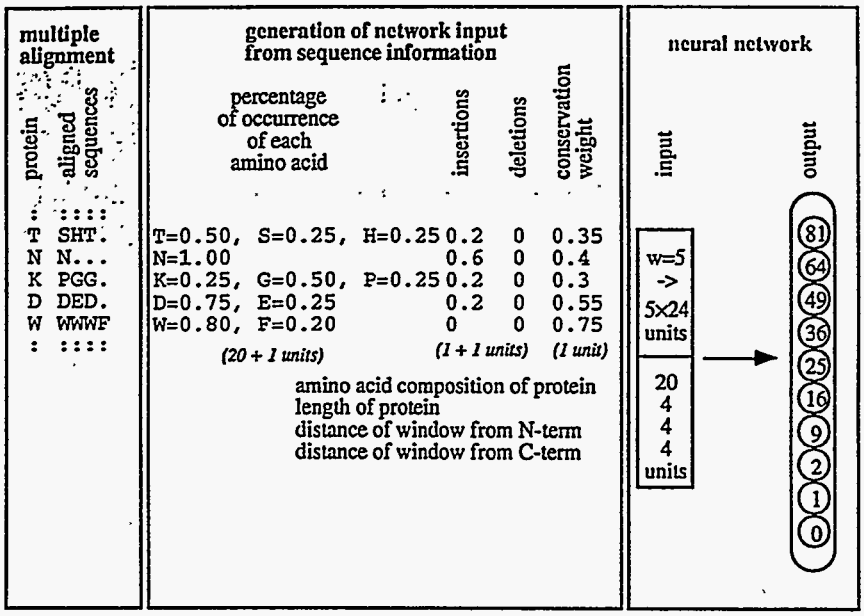
$$\mathrm{RelAcc}_{10} = \mathrm{INTEGER}\sqrt{100 \times \mathrm{RelAcc}}$$



## Lessons from 3D families

- 10 state description sufficiently detailed
  - binary and ternary descriptions lead to a frustrating ambivalence in choosing the thresholds for state distinctions
- Solvent accessibility is less conserved than is secondary structure
- Accuracy of homology prediction sharply decreases with sequence identity
- Small residues are conserved best

## Fig. 3.26: Neural network for accessibility prediction






## Accessibility prediction by statistics

(Wako & Blundell, 1994a)

- two states: buried (<20%), exposed (>20%)
- information theory on multiple alignments

# Measuring prediction accuracy

$$CorAcc = \frac{<x \times y> - <x> \times <y>}{\sqrt{<x^2> - <x>^2} \times \sqrt{<y^2> - <y>^2}}$$

Correlation of accessibility, with x and y being the relative accessibility a pair of homologue proteins (for the analysis of accessibility conservation in 3D families), or for a prediction and the observation (for the analysis of prediction accuracy).

$Q_2$ = percentage of conserved (or correctly predicted) residues in two states (B, E) defined by thresholds given above.

$Q_3$ = percentage of conserved (or correctly predicted) residues in three states (B, I, E) defined by thresholds given above.

$Q_{nX}$ = for $n$ states: percentage of conserved (or correctly predicted) residues in state $X$.

$Q_{3X}^{\%obs}$ = same as before, for the prediction of accessibility the percentages are normalised by the number of residues observed to be in state $X$.

$Q_{3X}^{\%pred}$ = probability for a correct prediction, i.e. the number of residues predicted correctly in state $X$ (× 100) divided by the number of all residues predicted to be in state $X$.



# Accuracy of predicting solvent accessibility

| | 2 states | 3 states | | | 10 states | |
|---|---|---|---|---|---|---|
| | Q2 | Q3 | Q3B | Q3I | Q3E | CorAcc |
| ***PREVIOUS METHODS*** | | | | | | |
| ‡ Wako & Blundell (13 families) | 76.5 | | | | | |
| ‡ Holbrook et al. (5 proteins) | 72.0 | 52.0 | 51 | 44 | 62 | |
| ***PHDacc for different testing sets*** | | | | | | |
| * *PHDacc* 126 = cross-validation set | 75.0 | 57.9 | 76 | 12 | 81 | 0.54 |
| * *PHDacc* 112 = pre-release set | 74.7 | 57.9 | 77 | 12 | 75 | 0.54 |
| * *PHDacc* 99 monomers (of 238) | 77.7 | 60.5 | 77 | 13 | 81 | 0.59 |
| * *PHDacc* 13 from Wako & Blundell | 79.2 | 60.8 | 77 | 12 | 86 | 0.61 |
| * *PHDacc* 5 from Holbrook et al. | 75.7 | 58.4 | 76 | 10 | 79 | 0.55 |

- **most accurately predicted:**
  **residues in helices and in buried strands**

## Accessibility prediction: conclusion

- Evaluation
  - Accurate enough to seed predictions of secondary structure (Wako & Blundell, 1994b; Benner et al., 1994)
  - Not accurate enough to be as useful as secondary str. predictions
  - Clear improvement by database growth (evolutionary information)
- Applications
  - Post-processing prediction methods
    - » prediction of contact maps: upper and lower limits
    - » threading
  - Speculations about binding sites and function



## Transmembrane helix prediction

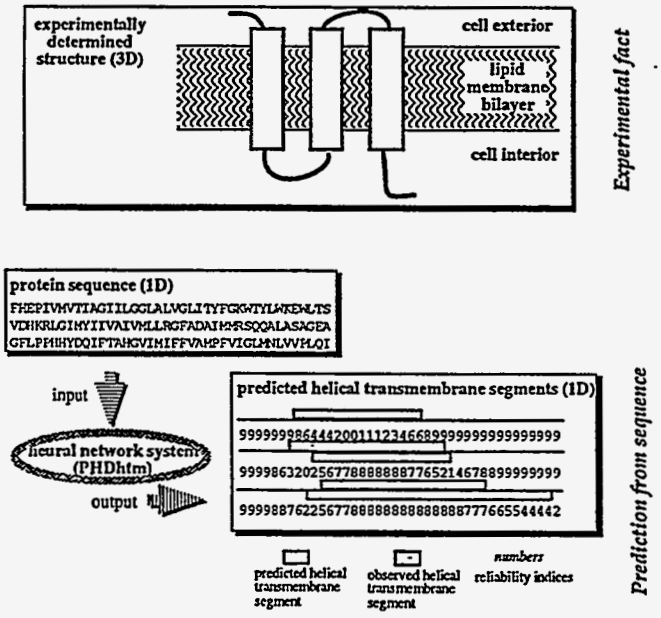
- Goal and concept
- Methods
  - Expert rules based on physico-chemical properties
  - Statistics
  - Neural networks
- Results
  - » Measures for accuracy
  - » Two-state accuracy about 95%
- Evaluation
  - » Often accurate enough to seed 3D or topology predictions
  - » Improvement by database growth (evolutionary information)
- Further method (β-strand segments)
- Applications
  - » Design mutation experiments; Speculations about binding sites and function; Fast mapping of all proteins from entire chromosomes

# Transmembrane helices: prediction goal

- **residues bound to lipid bilayer**
  Fig. 3.27
- **problem: lack of reliable data**
  - X-ray
    - photosynthetic reation, 1prc (Deisenhofer et al., 1985)
    - bacteriorhodopsin, 1brd (Henderson et al., 1990)
    - light harvesting complex II, (Wang et al., 1993; Kühlbrandt et al., 1994)
    - porin (16-stranded-beta-barrel)
      (Weiss & Schulz, 1992; Cowan & Rosenbusch, 1994; Kreusch & Schulz, 1994)
  - experimental assignment
    - Swissprot (Bairoch & Boeckmann, 1994)
    - lists compiled by:
      (Manoil & Beckwith, 1986; von Heijne & Gavel, 1988; von Heijne, 1992; Sipos & von Heijne, 1993; Jones et al., 1994)



# Fig. 3.27: Locations of transmembrane helices



# Methods for predicting TM-helix-locations

- **hydrophobicity scales**

  (Argos et al., 1982; Kyte & Doolittle, 1982; Eisenberg et al., 1984a; Eisenberg et al., 1984b; Engelman et al., 1986; Cornette et al., 1987; Degli Exposti et al., 1990; Claverie & Daulmiere, 1991)

- **expert rules**
  - **positive-inside rule:**

    positively charged amino acids (R, K) are more abundant in cytoplasmic than in periplasmic segments

    (von Heijne, 1981, 1986, 1991, 1992; von Heijne & Gavel, 1988; von Heijne & Manoil, 1990;Boyd & Beckwith, 1990; Dalbey, 1990; Sipos & von Heijne, 1993)

- **information theory**

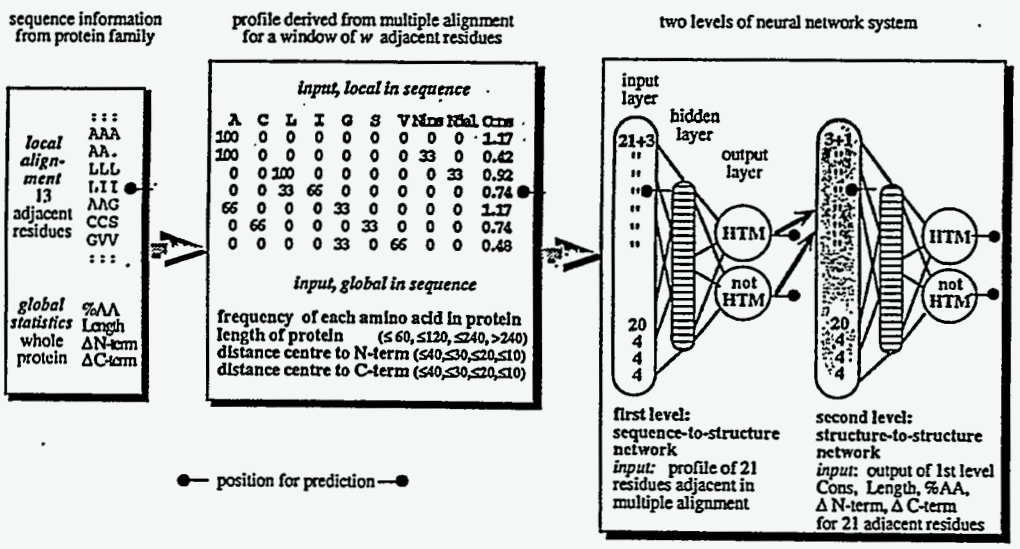
  (Engelman, 1993; Jones et al., 1994; Persson & Argos, 1994)

- **neural networks**

  (Fariselli et al., 1993; Casadio et al., 1995; Rost et al., 1995)



# Fig. 3.28: HTM prediction by neural network



# Fig. 3.29: Filter for HTM prediction

**too short helices**

if { L < 17 ∩ RI>7 (at either end of helix) }--> elongate helix by one residue until L ≥ 17

if { only one helix predicted }
  if { L < 17 } --> cut helix

if { at least 2 helices predicted }
  if { L < 11 } --> cut helix

**too long helices**

if { L > 35 } --> split helix at position L/2 into two helices of length L/2

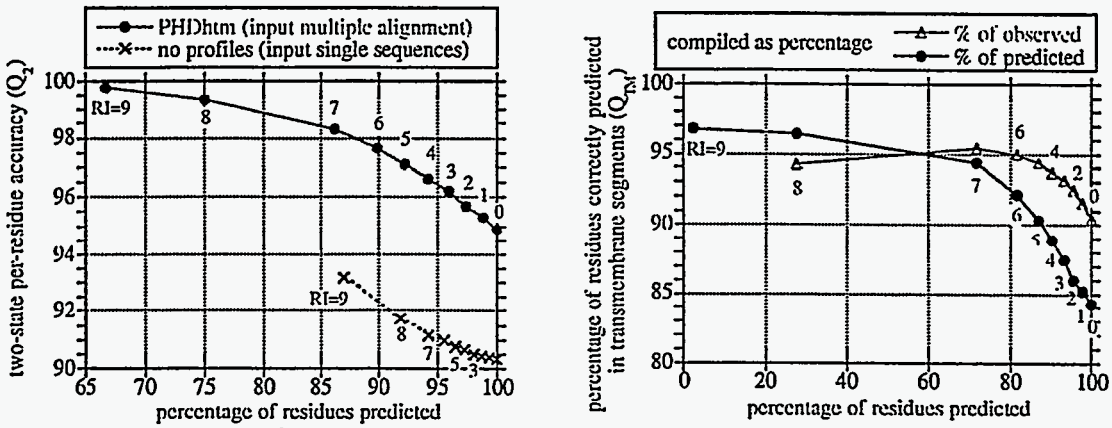
if { L > n × 22, n=3,4,... } --> split helix into n of length L/n



# Accuracy of HTM prediction

| | | | Overall | | Helical transmembrane segments only | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Per-residue score | | | | Segment-based scores | | | |
| Set[b] | Method[c] | $N$ | $Q_2$ | Info | %Obs $Q_{TM}$ | %Prd $Q_{TM}$ | Corr | $\langle L \rangle$ | %Obs Sov | %Prd Sov | Nseg[d] over | Nseg under |
| Set 1 | No profiles | 69 | 90 | 0.45 | 84 | 70 | 0.71 | 23 | 90 | 81 | 15<br>6.3% | 47<br>17% |
| | PHDhtm | 69 | 95 | 0.64 | 91 | 84 | 0.84 | 23 | 96 | 96 | 5<br>1.9% | 10<br>3.8% |
| Set 2 | PHDhtm | 37 | 95 | | 91 | | 0.85 | 23 | | | | |
| | Edelman (1993) | 37 | 88 | | 90 | | 0.70 | 26 | | | | |
| Set 3 | Jones et al. (1994) | 67 | | | | | | | | | 15<br>4.5% | 6<br>1.9% |
| Set 4 | PHDhtm | 28 | | | | | | | | | 3-2[e]<br>1.6% | 3<br>2.3% |
| | Persson and Argos (1994)<br>Not cross-validated[f] | 28 | | | | | | | | | 2-3[e]<br>1.6% | 3<br>2.3% |

## Fig. 3.30: Reliability of HTM prediction






## Transmembrane helix prediction: conclusion

- **Evaluation**
  - » Often accurate enough to seed 3D or topology predictions
  - » Improvement by database growth (evolutionary information)
- **Further method (β-strand segments)**
  - » prediction methods for globular proteins (secondary structure prediction) often accurate enough, but...
  - » no general method available.
- **Applications**
  - » Design mutation experiments
  - » Speculations about binding sites and function
  - » Fast mapping of all proteins from entire chromosomes

Fig. 3.31

# Fig. 3.31: HTM regions for entire chromosome: yeast VIII

Table 5. *Prediction of transmembrane helices for yeast chromosome VIII*[a]

| Identifier | Nres[b] | Nali[b] | Locations of predicted segments | | | | Nhtm[b] |
|---|---|---|---|---|---|---|---|
| YHL040c | 627 | 5 | 75–88 | 116–127 | 141–157 | 173–190 | |
| | | | 205–216 | 231–252 | 285–308 | 326–342 | |
| | | | 363–387 | 404–418 | 429–441 | 458–477 | |
| | | | 568–581 | | | | 13 |
| YHL047c | 637 | 5 | 70–83 | 111–122 | 136–152 | 168–185 | |
| | | | 200–211 | 226–247 | 280–303 | 321–337 | |
| | | | 358–382 | 400–413 | 425–436 | 453–473 | |
| | | | 563–576 | | | | 13 |
| YHR092c | 560 | 21 | 70–87 | 124–139 | 152–171 | 179–196 | |
| | | | 215–226 | 247–261 | 369–385 | 400–413 | |
| | | | 435–459 | 474–492 | 500–518 | | 11 |
| YHR096c | 592 | 18 | 85–101 | 138–154 | 167–186 | 194–212 | |
| | | | 230–241 | 262–276 | 385–400 | 415–428 | |
| | | | 450–475 | 489–507 | 515–533 | | 11 |
| YHR094c | 570 | 17 | 64–80 | 118–133 | 146–165 | 173–191 | |
| | | | 209–220 | 241–255 | 363–379 | 394–407 | |
| | | | 429–453 | 468–486 | 494–512 | | |
| YHR026w | 213 | 18 | 20–37 | 56–80 | 94–122 | 145–168 | |
| | | | 180–205 | | | | 5 |
| YHR002w | 357 | 8 | 37–53 | 102–115 | 141–153 | 201–227 | |
| | | | 271–281 | | | | 5 |
| YHL048w | 381 | 4 | 39–62 | 70–93 | 233–252 | 260–277 | 4 |
| YHR190w | 444 | 4 | 272–283 | 295–310 | 425–440 | | 3 |
| YHR129c | 384 | 258 | 137–153 | 349–360 | | | 2 |
| YHR005c | 472 | 153 | 337–347 | 377–387 | | | 2 |
| YHR183w | 489 | 39 | 360–371 | 418–429 | | | 2 |
| YHR046c | 295 | 7 | 103–117 | 201–216 | | | 2 |
| YHR176w | 373 | 6 | 262–272 | 338–351 | | | 2 |
| YHR039c | 644 | 5 | 49–66 | 247–264 | | | 2 |
| YHL011c | 320 | 22 | 73–92 | | | | 1 |
| YHR028c | 818 | 8 | 26–44 | | | | 1 |
| YHR007c | 530 | 7 | 25–47 | | | | 1 |
| YHR037w | 575 | 4 | 209–227 | | | | 1 |



# Prediction of protein structure in 2D

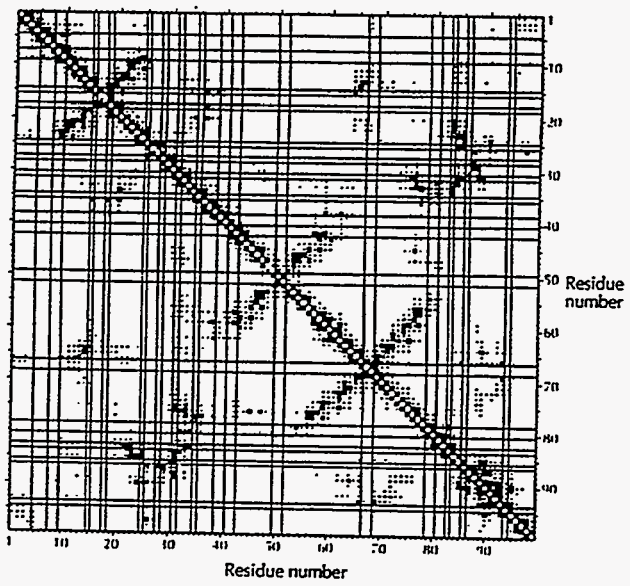
- Prediction of (long-range) inter-residue contacts
- Prediction of contacts between beta-strands
- Prediction of disulphide bonds

# Prediction of inter-residue contacts

- Goal and concept
- Methods
  - Statistics (correlated mutations)
  - Neural networks
- Results
  - » Predictions based on correlated mutations: between 1.4 and 5.1 times better than random predictions
  - » For others, results difficult to assess
- Evaluation
  - » Distinction between alternative models for 3D structure?
  - » No prediction of conformations ab initio
- Applications
  - » Possibly many, none so far



# Fig. 3.32: Contact map



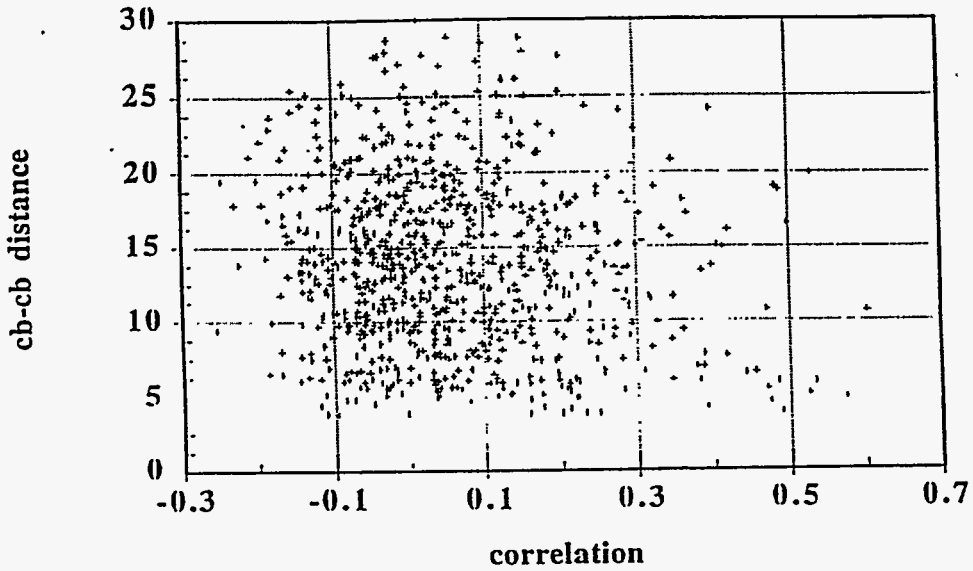
## Prediction of contacts by correlated mutations

- evolutionary constraints on protein sequences
  - selective pressure from need to maintain protein function
  - consequently, conservation and mutation patterns evidence of functional or structural constraints plus mutational drift
    - » functional constraints: surface residues
    - » mutational drift: loop regions
    - » structural constraints: core
  - simplifying assumption:
    residues in contact show correlated mutational behaviour, i.e., if one residue mutates, its contact partners also tend to mutatate
- Do correlated mutations imply spatial proximity?
  - sometimes

(Altshuh et al., 1987; Altshuh et al., 1988; Neher, 1994; Taylor & Hatrick, 1994; Shindyalov et al., 1994; jGoebel et al., 1994)



## Fig. 3.33: Mutations correlated to distance




(Figure 2 from Goebel et al., 1994)

# Correlation between mutations

- **starting point: multiple alignment derived mutation matrix**

  Fig. 3.34

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{w_{kl}\,(s_{ikl} - \langle s_i \rangle)\,(s_{jkl} - \langle s_j \rangle)}{\sigma_i\,\sigma_j}$$

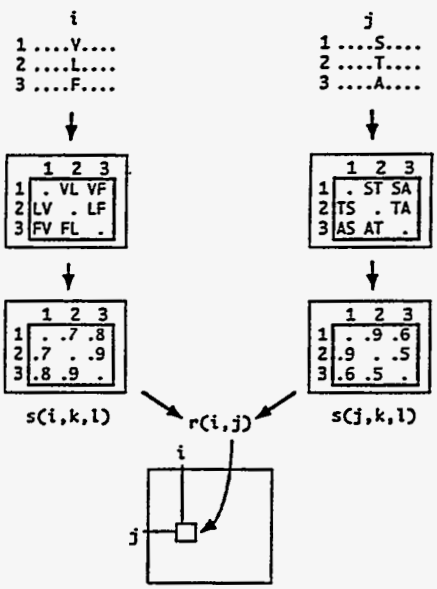
*rij* distance between residues at position *i* and *j* ; *sikl* mutation matrix for residue at position *i* , *k*, *l* = *1*, ..., *Nali* , where *Nali* is the number of sequences in the alignment; <si> is the average over all *k* and *l* , and $\sigma_i$ the respective standard deviation

- **contact predicted, if** r(i,j) > threshold
  - exclude positions with > 10% gaps
  - exclude completely conserved positions
  - define clusters of correlated residues:

    cluster of rank n:
    residue part of cluster n, if it is correlated with at least n other residues in the cluster

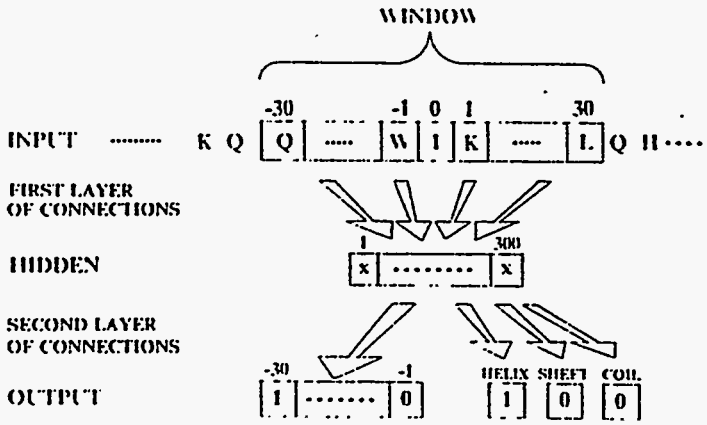


# Fig. 3.34: Correlated mutations




(Figure 1 from Goebel et al., 1994)

## Fig. 3.35: Distance matrix prediction by neural network




(Figure 1 from Bohr et al., 1990)



## Accuracy of inter-residue contact prediction

- **Accuracy:**

$$Acc_{pred} = \frac{C_{correct}}{C_{predicted}}$$

How many of the *predicted* contacts are observed?

- **Coverage:**

$$Cov^{pred} = \frac{C_{correctly\ predicted}}{C_{observed}}$$

How many of the *observed* contacts are predicted?

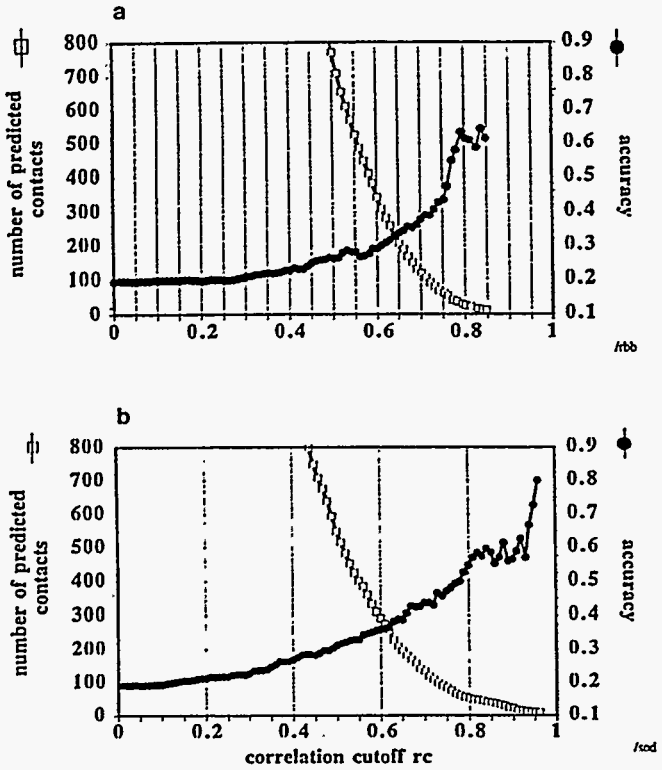
- **Improvement over random:**

$$R_{improve} = \frac{Acc^{pred}}{Acc^{random}}$$

random prediction: contact density
-> dependent on size, e.g.:

trypsin inhibitor (56) => random = 0.39
trypsin (223) random = 0.13

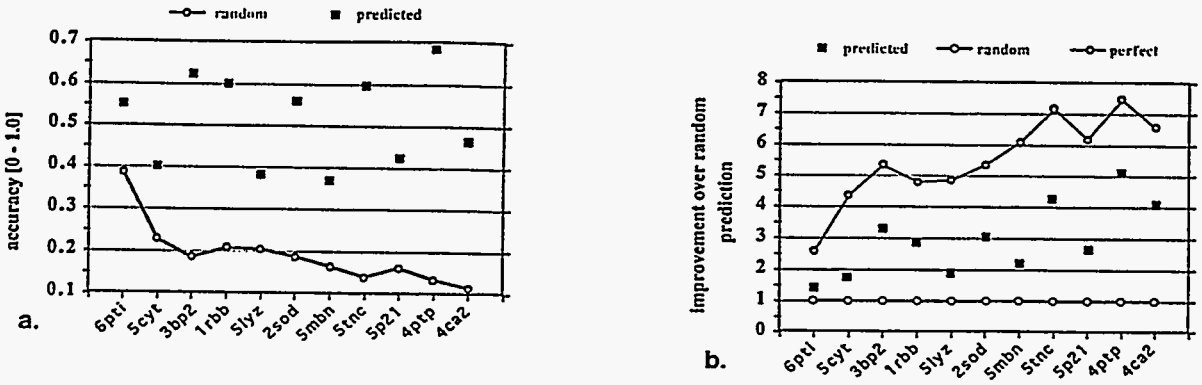
## Fig. 3.36: Pay-off between accuracy and coverage




(Figure 3 from Goebel et al., 1994)

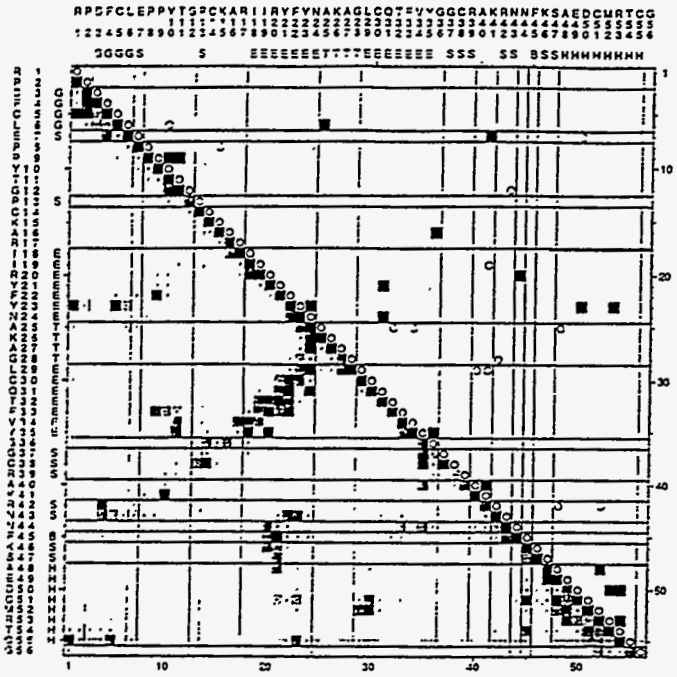


## Fig. 3.37: Accuracy of contact prediction (CorrMut)




(Figure 5 from Goebel et al., 1994)

## Fig. 3.38: Predicted contact map (CorrMut)



(Figure 4 from Goebel et al., 1994)



## Fig. 3.39: Predicted contact map (Neural Network)

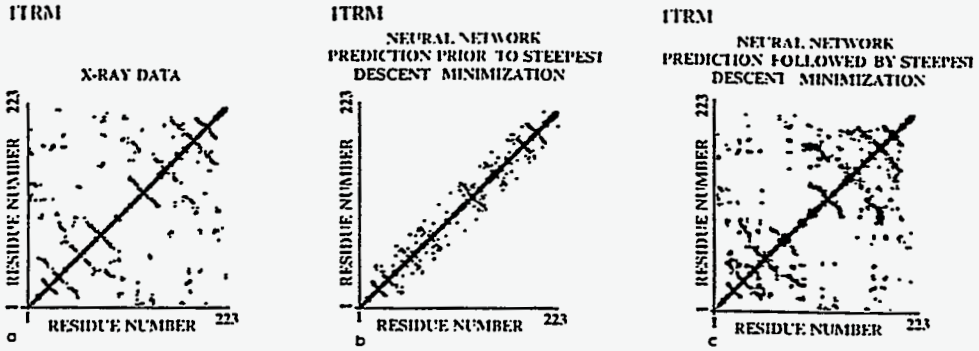



Fig.2. Binary distance matrices for 1TRM. The matrices (223 × 223) show which $C_\alpha$ atoms are within an 8 Å distance to each other $C_\alpha$ atom in the folded protein. a) The matrix corresponding to the structure determined from the X-ray data (8 Å threshold). b) Neural network prediction of an 8 Å distance matrix. A 61-residue band centered along the diagonal is generated. The network predicts this band with an accuracy of 96.6%. c) The matrix corresponding to the structure produced by steepest descent minimization, using the neural network prediction as a starting point

(Figure 2 fromBohr et al, 1990)

## Inter-residue prediction: conclusion

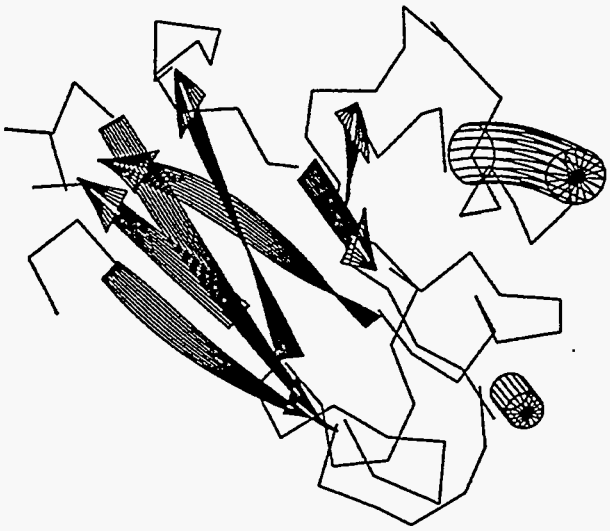
- **Results**
  - » problem is a hard one (at least for non-local contacts)
  - » improvement of 1.4 - 5.1 times over random predictions
- **Evaluation**
  - **– ab initio prediction of conformations not possible, ...**
  - **– ... but, distinction between alternative models may be possible**
  - **– open: combine information from correlated mutations with:**
    - » conservation of residues (Taylor & Hatrick, 1994)
    - » statistical predictions (Galaktionov & Rodionov, 1980; Galaktionov & Marshall, 1994)
    - » other...
- **Applications**
  - **– post-processing prediction methods**
  - **– speculations about function**
  - **– HOWEVER, none so far ...**

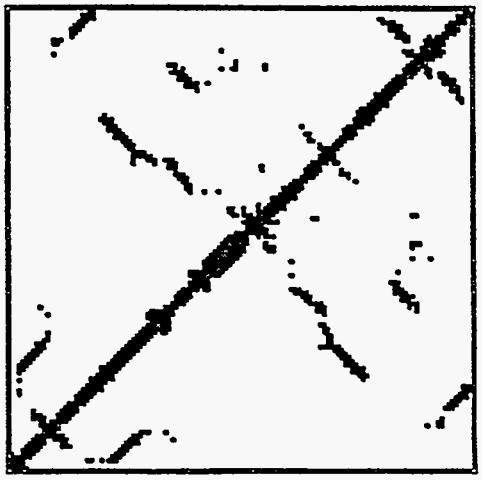


## Prediction of contacts between β-strands

- **Goal and concept**
- **Methods**
  - **– Statistics (potentials of mean force)**
- **Results**
- **Evaluation**
  - » Less accurate for predicted strands,
  - » But used successfully for predicting higher aspects of 3D structure
- **Applications**
  - » Post-processing prediction methods
  - » Speculations about binding sites and function

## Fig. 3.40: Contacts between strands





[WHAT IF]



## Fig. 3.41: Generation of propensity tables

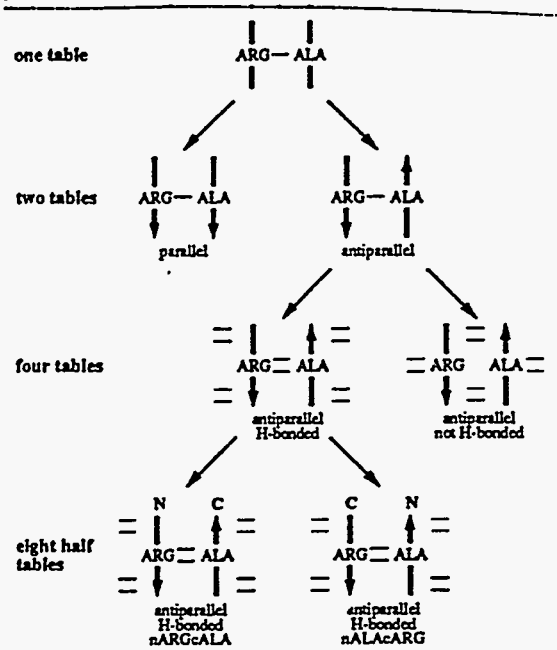



Figure 1 Subdivisions of β–strand residue pairs by parallel/antiparallel; hydrogen-bonding pattern and with respect to N and C termini. Some subdivisions have been omitted from the figure for clarity.

(Figure 1 from Hubbard, 1994)

# Contact propensities

- **residue-contact propensities:**

$$p(a,b) = -\log\left[\frac{\dfrac{n(a,b)}{n(a)\,n(b)}}{\dfrac{\sum_{x}^{\text{all res}}\sum_{y}^{\text{all res}} n(x,y)}{\sum_{x}^{\text{all res}}\{n(x)\}^2}}\right]$$



# Fig. 3.42: Distinguishing 5 classes

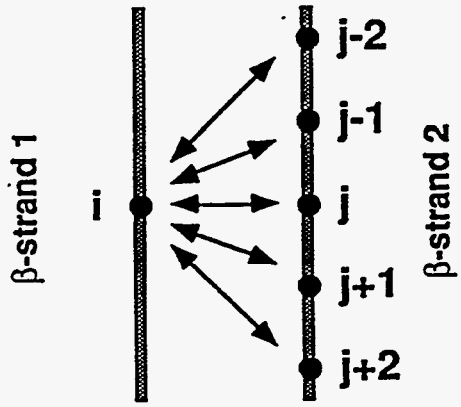



Figure 2 For each β–strand residue pair ij, the occurrences of pairs ij-2, ij-1, ij, ij+1, ij+2 are counted in separate tables.

(Figure 2 from Hubbard, 1994)

# Definition of pseudo-potentials

- definition of pseudo-potential:
  sum of propensities from relevant tables for all pair interactions (Fig. 3.42),
  divided by total number of interactions summed
  (four different for tables in Fig. 3.41)
- selective for:
  - parallel / antiparallel
  - correct / incorrect hydrogen-bonding
  - correct / incorrect strand order
- ability to identify correct strand alignments
  (without knowing length of strand-strand interaction)
  Fig. 3.43
- accuracy about 35-45%



# Fig. 3.43: Identifying the correct strand alignment

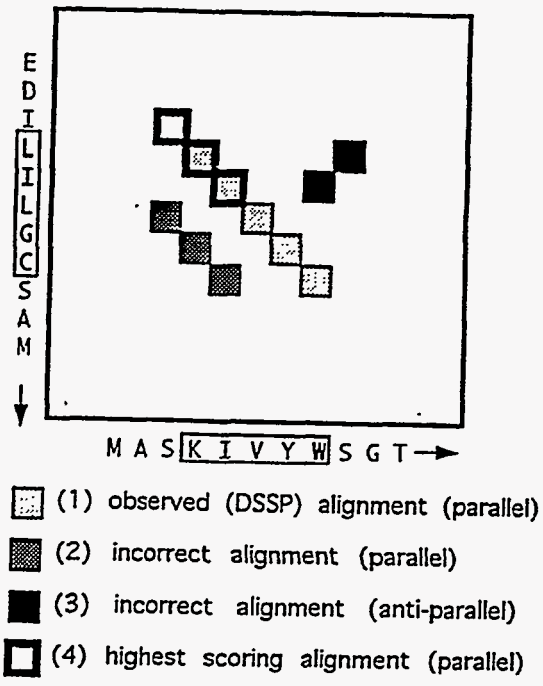



Figure 5 Method for searching local alignment space around a β–strand pair (KIVYW and LILGC) which interact to form a parallel β–sheet. Arrows following the sequences point towards the C-terminus of the protein. Each box indicates an IJ alignment between corresponding residues on each axis. (1) indicates the correct alignment. (2) indicates an misalignment of the strands by 3 residues. (3) indicates an alignment of the wrong sheet type. (4) indicates an alignment which aligns the correct residues, but which is a different length and overlap.

(Figure 5 from Hubbard, 1994)

# Fig. 3.44: SH3: contacts observed

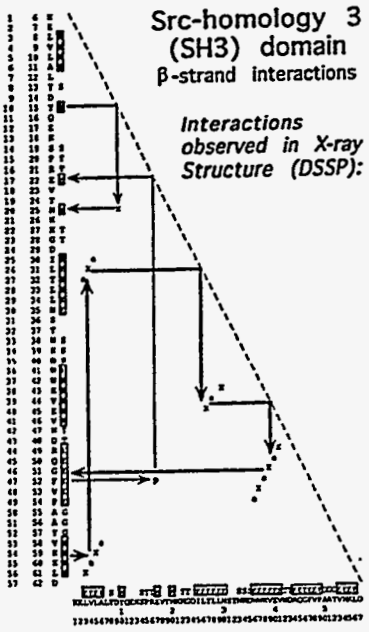



Figure 8 The β–strand contact map for Src-homology 3 (SH3) domain protein derived from the known structure also showing the connectivity of the sheet. Contacts containing 'a' are antiparallel and 'p' parallel. The 4 columns on the left are sequence index (1-n), residue number (from PDB file), aminoacid, DSSP summary information [B=isolated β–bridge, E=strand, G=$3_{10}$ turn, H=helix, S = bend, T=turn].

(Figure 8 from Hubbard, 1994)



# Fig. 3.45: SH3: all contacts predicted

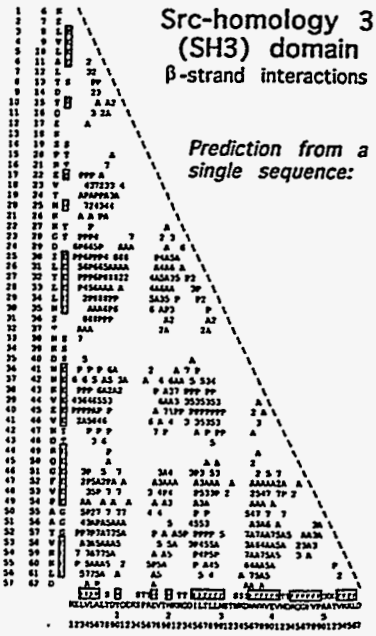



Figure 9 Predicted β–strand contact map for SH3 from a single sequence. The predicted contacts are identified as A5A5A and P4P4 for a length 5 antiparallel interaction and a length 4 parallel interaction respectively. No information about the score of the predicted contact is shown.

(Figure 9 from Hubbard, 1994)

# Fig. 3.46: SH3: contacts predicted from alignment

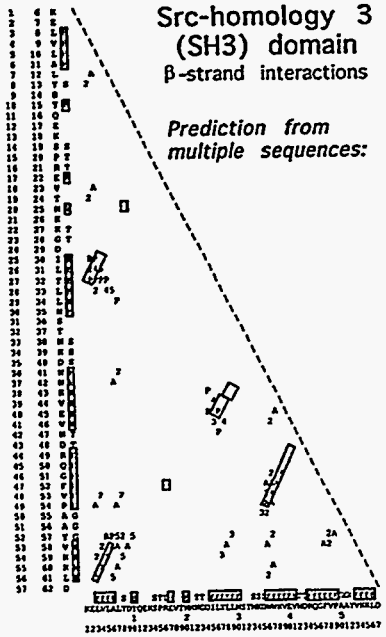



Figure 10 Predicted β-strand contact map for SH3 smoothed by considering multiple sequences. The Diagonal boxes show the location of the true β-interactions (from fig 7). Predicted contacts are identified as in figure 9.

(Figure 10 from Hubbard, 1994)



# Inter-strand contact prediction: conclusion

- Results
  - accurate enough to be useful
- Evaluation
  - less accurate for predicted strands (PHD), ...
  - ... but successfully applied

    (Hubbard& Park, 1995; Defay & Cohen, 1995)
- Applications
  - post-processing prediction methods
    - » fold recognition
    - » 3D modelling
  - speculations about binding sites

# Prediction of disulphide bonds

- Goal and concept
- Methods
  - Neural network
- Results
  - » About 80% of contacts correctly predicted
- Evaluation
  - » Small test set (some hundred examples)
- Application
  - » ?



# Predicting disulphide bonds: goal

- special rôle of cysteine residues
  - » Cysteine (C) is of particular importance since it forms covalent disulphide bridges often crucial for the stability of a protein: (Creighton, 1984)

$$CH_2\text{-}SH + SH\text{-}CH_2 \rightarrow CH_2\text{-}S{=}S\text{-}CH_2$$

  - » An SS bridge form is through a mixed disulphide intermediate. All Cysteine/SS reactions are through the $S^-$ anion (Ewbank, 1992)
- prediction of:

  S=S
  S=X

-

# Fig. 3.47: Neural network for disulphide-bond prediction

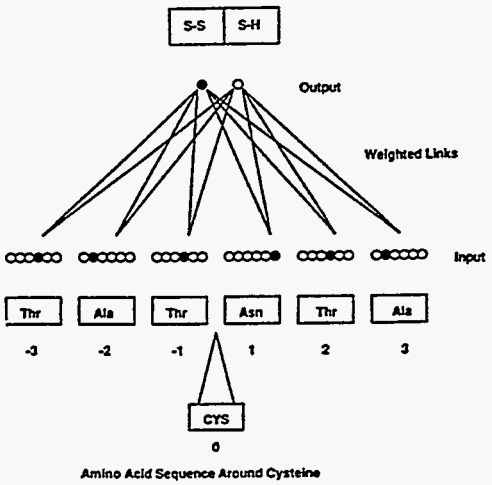



Fig. 1. A diagram of network architecture. For clarity, only six window positions (three amino acids to the N-terminal and three amino acids to the C-terminal side of an assumed centered cysteine) and six nodes per window position are illustrated. Within a window position, an amino acid is represented by giving a value of 1.0 to its node while setting all other nodes in that window position to 0.0. Input values are propagated through weighted links to produce activities at the two output nodes, S−S and S−H. The output node with the highest activity is the network's decision.

(Figure 1 from Muskal et al., 1990)



# Fig. 3.48: Pay-off between accuracy and coverage

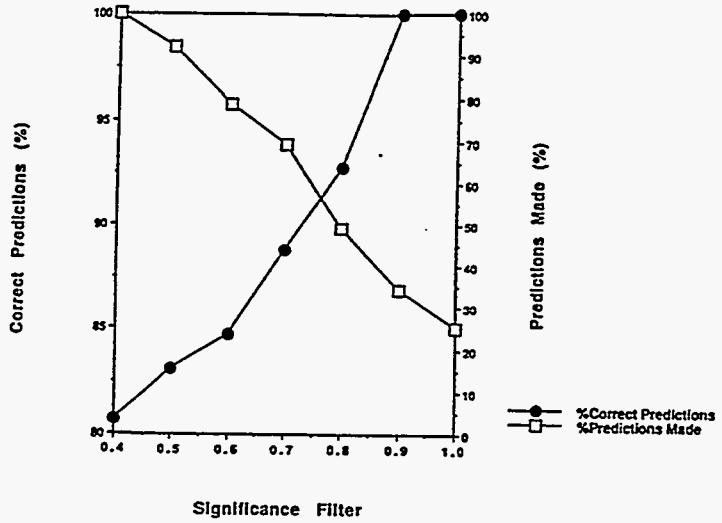



Fig. 2. Dependence of predictive accuracy on the strength of output node activities. The significance filter is placed over the output nodes so that only activities greater than the filter can pass through for prediction. Data were average from the seven testing sets in Table III.

(Figure 2 from Muskal et al., 1990)

## Predicting disulphide bonds: conclusion

- Results
  - SS 81%
  - SX 80%
- Evaluation
  - extremely small testing set $7 \times 20$
- Applications
  - » filtering contact predictions
  - » post-processing prediction methods
  - » BUT, non so far published



## Prediction of protein structure in 3D

- Sequence alignment THE prediction tool
- Homology modelling
- Potentials of mean force
- Remote homology modelling (threading)

## Sequence alignment THE prediction tool

- Goal and concept
- Methods
  - Hashing
  - Dynamic programming
- Results
  - » Straightforward for high levels of pairwise sequence identity
  - » Tricky below about 30% pairwise sequence identity
- Evaluation
  - » Power of dynamic programming grows with databases
  - » Sensitive and fast enough as first step for sequence analysis
  - » Drawback: few methods provide cut-off criteria
- Applications
  - » Post-processing prediction methods
  - » Prediction of function or binding sites



## Alignment methods: fast, hashing

- FASTA:
  - 1. search identical 'words' (e.g. pairs)
  - 2. widen range of identity (profile based)

    (Dumas & Ninio, 1982; Wilbur & Lipman, 1983; Lipman & Pearson, 1985; Pearson & Lipman, 1988)
- BLAST
  - 1. list of high scoring words, typically words of length four with high information
  - 2. search database for identical words
  - 3. expand words to segments

    (Altschul et al., 1990; Karlin & Altschul, 1990; Karlin et al., 1990; Altschul, 1991, 1993)

# Alignment methods: slow, dynamic programming

- exchange matrices
  - PAM: accepted point mutations (Dayhoff, 1978)
    (percent accepted mutations; point accepted mutations per 100 residues)
    merely counts of occurrences
  - mutation matrix: probability of amino acid exchanges
  - log-odds matrices: logarithm of exchange probabilities
  - comparison of various matrices: (Henikoff & Henikoff, 1993)
- dynamic programming (optimal alignment)
  - gaps originally length independent (Needleman & Wunsch, 1970)
  - length dependent: (Sellers, 1974)

    $$g(k) = g_0 + g_e k$$

    $g_0$ gap open penalty; $g_e$ gap elongation penalty; k length of gap
    typically $g_e / g_o = 1/10$
  - problem: global alignment, i.e., full length of aligned sequences optimised
  - alternative: align similarities (Smith & Waterman, 1981)



# Fig. 3.49: Dynamic programming

| | A | B | C | N | J | R | O | C | L | C | R | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 8 | 7 | 6 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| J | 7 | 7 | 6 | 6 | 6 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 6 | 6 | 7 | 6 | 5 | 4 | 4 | 4 | 3 | 3 | 1 | 0 | 0 |
| J | 6 | 6 | 6 | 5 | 6 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| N | 5 | 5 | 5 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | 0 | 0 |
| R | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 3 | 3 | 2 | 2 | 0 | 0 |
| C | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 4 | 3 | 3 | 1 | 0 | 0 |
| K | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 |
| C | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 1 | 0 | 0 |
| R | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 0 | 0 |
| B | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

FIG. 2. Contributors to the maximum match in the completed array.
The alternative pathways that could form the maximum match are illustrated. The maximum match terminates at the largest number in the first row or first column, 8 in this case.

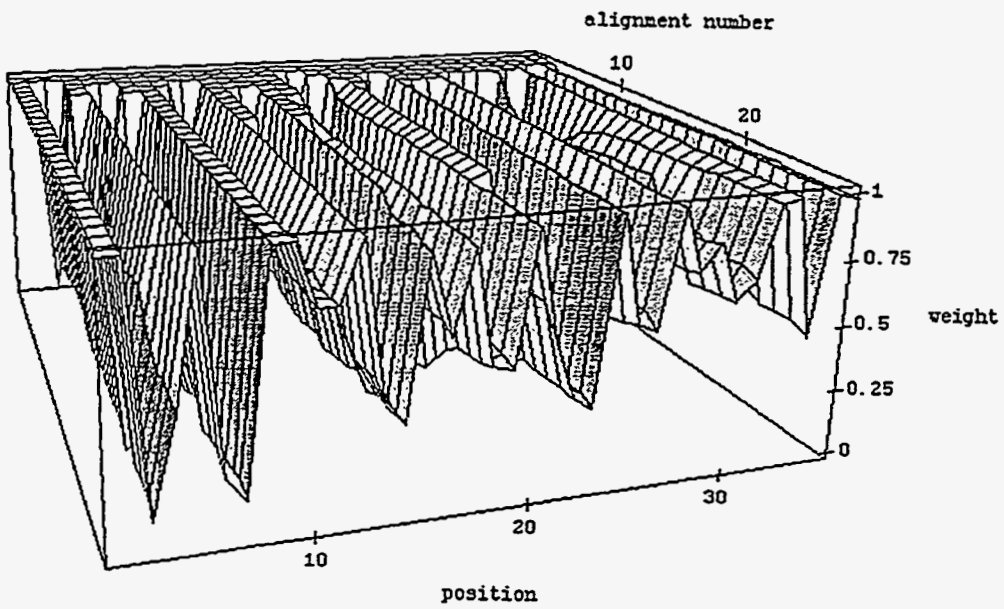
(Figure 2 from Needleman & Wunsch, 1970)

# Alignment methods: multiple alignment

- optimal alignment practicable for Nali <=3
  (Murata et al., 1985; Murata, 1990)
- pairwise alignment -> multiple alignment
  (Barton & Sternberg, 1987; Feng & Doolittle, 1987; Taylor, 1987; Corpet, 1988; Higgins & Sharp, 1988; Vingron & Argos, 1987; Sander & Schneider, 1991; Higgins et al., 1992; Schneider, 1994)
- profile-based alignment
  e.g. MaxHom (Sander & Schneider, 1991; Schneider, 1994)
  - position dependent conservation weight
  - 1. pairwise alignment of homologous sequences based on conservation weight of previously aligned sequences
  - 2. fix conservation weights
  - 3. repeat pairwise alignments with fixed conservation weights

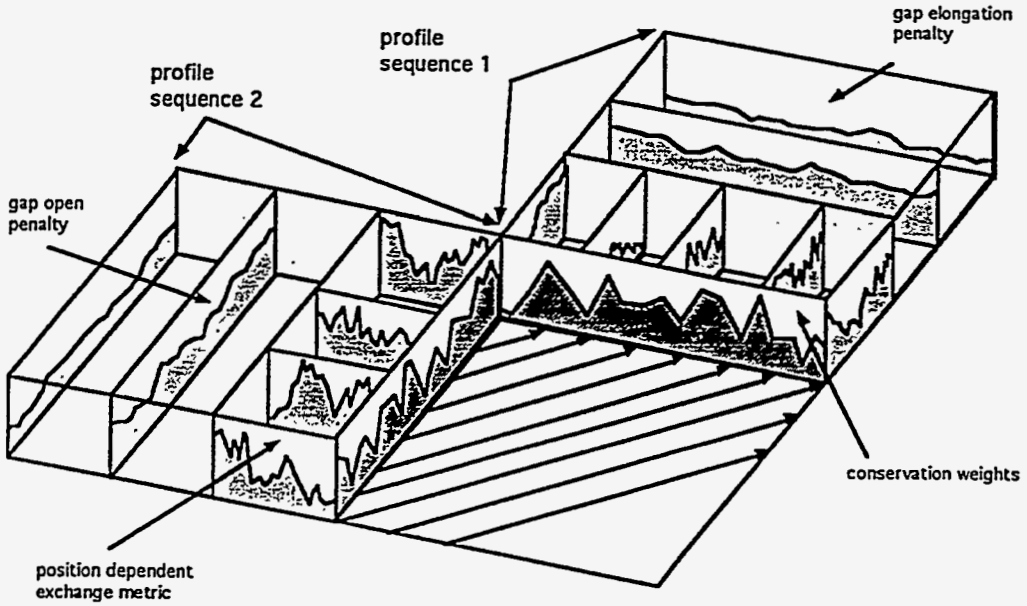


# Fig. 3.50: Evolution of conservation weights




(Figure 12 from Schneider, 1994)

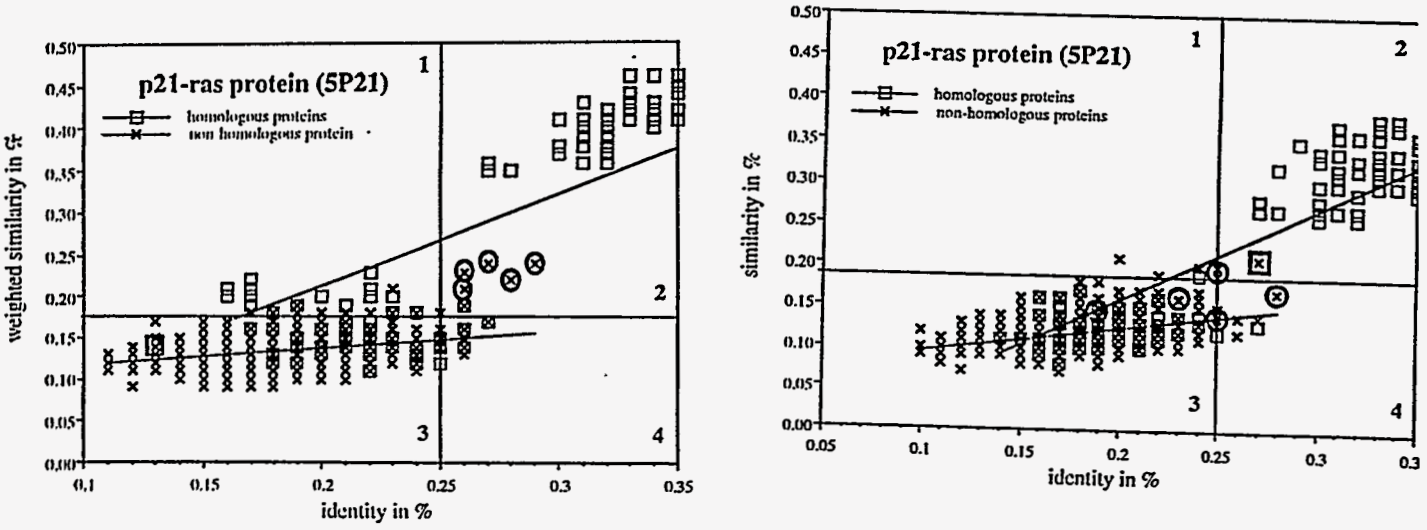
## Fig. 3.51: Profile-based alignment algorithm: MaxHom




(Figure 13 from Schneider, 1994)



## Fig. 3.52: Profile-based alignment algorithm: p21-ras




(Figure 14 from Schneider, 1994)

## Position dependent conservation weight

$$cw(i) = \frac{\sum_{kl=1}^{Nali} w_{kl}\, sim_{kl}(i)}{\sum_{kl=1}^{Nali} w_{kl}}, \quad \text{with } w_{kl} = \left(1 - \frac{1}{100}\,\%identity_{kl}\right)$$

*cw(i)* conservation weight at position *i* ; *Nali* number of sequence in alignment; *k, l* indices for sequences in multiple alignment; $w_{kl}$ weighting factor to balance uneven distribution in sequence space; $sim_{kl}(i)$ similarity between amino acids at position *i* of sequences *k* and *l* ; $\%identity_{kl}$ percent identity between sequences *k* and *l*

- normalised such that <cw> = 1
- include only sequences above threshold for homology



## Alignments: conclusion

- **Results**
  - » Straightforward for high pairwise sequence identity
  - » Tricky below 30% pairwise sequence identity
- **Evaluation**
  - » Power of dynamic programming grows with databases
  - » Sensitive and fast enough as first step for any sequence analysis
  - » Drawback 1: few methods provide cut-off criteria
  - » Drawback 2: lack of thourough tests on performance accuracy
- **Applications**
  - » Post-processing prediction methods
    - prediction in 1D, 2D, 3D
  - » Prediction of function or binding sites

# Homology modelling

- Goal and concept
- Methods
  - Rotamer libraries
- Results
  - » Accuracy depends on level of pairwise sequence identity
- Evaluation
  - » Sufficiently accurate to predict 3D structure
- Applications
  - » Site-directed mutations
  - » Prediction of function and binding sites



# Homology modelling: goal

- protein structure is more conserved than is sequence

  (Chothia & Lesk, 1986; Pastore & Lesk, 1990; Lesk, 1991; Lesk & Boswell, 1992; Holm et al., 1993; Holm & Sander, 1993; Holm & Sander, 1994a)
- single point mutations can be fatal to protein structure and function, but ...

  (Dao-pin et al., 1990; 1991a-c; Grenzin et al., 1992)
- most often, proteins within a sequence family have homologous 3D structure

  (Chothia & Lesk, 1986; Sander & Schneider, 1991)
- given a protein of unknown structure (SOUS), try to model its 3D structure by using the $C^{\alpha}$-backbone of a known structure as template

  early work: (Dickerson, 1976; Greer, 1980, 1981, 1990, 1991)
- limiting steps: function of pairwise sequence identity

  Fig.

## Fig. 3.53: Limiting steps of homology modelling

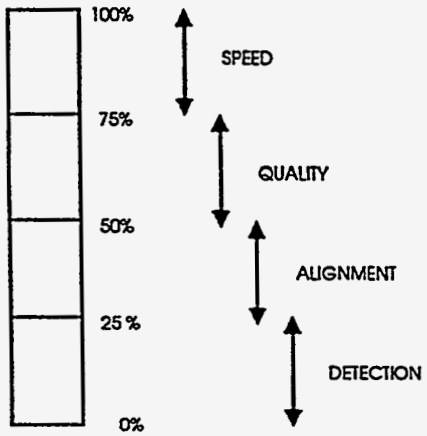



**Figure 1.** The main limiting steps for model building by homology as function of the percentage sequenc identity between the structure and the model.

(Figure 1 from Holm et al., 1994)



## Homology modelling: limitations

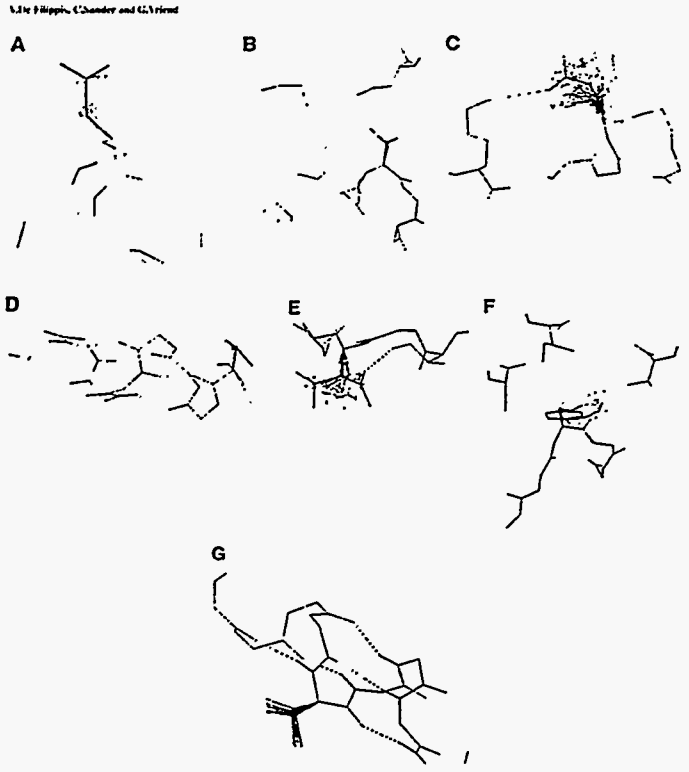
- **High homology: placing new side chains in the structure**
  - **side chains can be 'grown' during molecular dynamcis**
    (Karplus & Petsko, 1990; Cornell et al., 1991; Berendsen, 1991)
    - » problem: time (useful for difference of one residue)
  - **similar environment in database of known structures**
    (Ponder et al., 1987; Summers & Karplus, 1989; Summers & Karplus, 1990; Holm & Sander, 1992; Levit, 1992; Eisenmenger et al., 1993; Vriend & Sander, 1993; Vriend & Eijsink, 1993; De Fillippis et al., 1994; Vriend et al., 1994)
    - » problem 1: what is similar?
    - » problem 2: quick scan, i.e., database systems that allow for fast, easy and flexible retrieval of specific information
      (Bryant, 1989; Islam & Sternberg, 1989; Vriend, 1990)
- **Intermediate homology:**
  - **building loops if there is an insertion in the model**
  - **verification of quality of models**
- **Low homology: improving the alignment**

# Homology modelling: rotamer libraries

- **position-specific rotamer analysis**
  (Jones & Thirup, 1986; Vriend & Eijsink, 1993; Vriend et al., 1994)
  - **start: database of non-redundant sequences**
    (Hobohm et al., 1992; Hobohm & Sander, 1994)
  - **extract rotamer distribution**
  - **fragment lengths:**
    - » helix and strand: seven residues
    - » loop: five residues
  - **accepted fragments:**
    - » identical amino acid in centre
    - » local backbone similar to that around evaluated position (<0.5 Å r.m.s.d.)



# Fig. 3.54: Rotamer distributions




(Figure 2 from De Filippis et al, 1994)

## Homology modelling: conclusion

- **Results**
  - » Accuracy depends on level of pairwise sequence identity
  - » for high homology > 60% correct (De Filippis et al., 1994)
  - » for intermediate homology: sometimes loops correct (Abagyan & Totrov, 1993; Abagyan et al., 1994; Totrov & Abagyan, 1994)
  - » for low homology: rough estimate, not sufficient in general to design experiments
- **Evaluation**
  - » Sufficiently accurate to predict 3D structure
- **Applications**
  - » Site-directed mutations
  - » Drug design
  - » Prediction of function and binding sites



## Mean-force potentials

- **Goal and concept**
- **Methods**
  - **– Sippl potentials**
- **Results**
  - » Accurate enough to spot incorrect structures
- **Applications**
  - » Post-processing prediction methods (e.g. threading)
  - » Site-directed mutations
  - » Selection of the best among an ensemble of possible structures
  - » Spot stresses in structures

# Mean-force potentials goal

- **inductive approach: quantum-mechanics**
  - » semi-empirical force fields
    (Momany et al., 1975; Brooks et al., 1988; van Gunsteren, 1988, 1993; Brünger et al., 1986; Karplus & Petsko, 1990)
- **deductive: knowledge-based potentials of mean force**
  - » Boltzmann's principle
    (Sippl, 1990; Sippl et al., 1992; Hendlich et al., 1992; Sippl, 1993a)
-



# Fig. 3.55: Mean-force approach

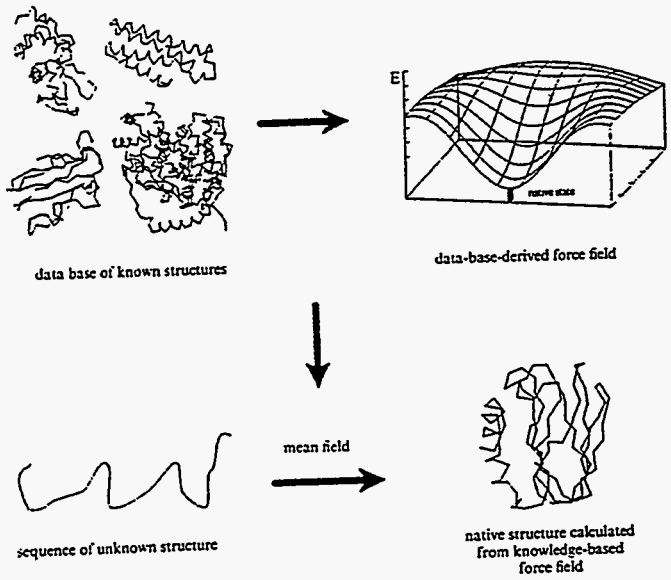



Fig. 1. Outline of the mean field approach to protein folding. The set of available 3D structures of proteins is used to extract a data-base-derived force field. If this attempt is successful the force field can be employed in the computational determination of protein structures.

(Figure 1 from Sippl, 1993a)

## Boltzmann's principle

- Boltzmann law:

$$p_{ijk} = \frac{1}{Z} \exp \frac{E_{ijk}}{kT}$$

i, j, k variables of system; k Boltzmann constant;
T temperature; Z partition function:

$$Z = \sum_{ijk} \exp \frac{E_{ijk}}{kT}$$

- general goal in statistical mechanics:
  given energy E ->
  compute partition function Z and probabilities p
  - problem 1: accurate energy function
  - problem 2: analytical or numerical computation of Z



## Boltzmann: inverse law

$$E_{ijk} = -k\,T \ln [f_{ijk}] + k\,T \ln Z$$

E: potential of mean force; f: relative frequencies obtained from measurements

note: $\lim_{n \to \infty} f_{ijk} = p_{ijk}$ , i.e.,
relative frequencies equal probability densities

- Z is constant, thus, no effect on energy differences
- consequently, here the following choice is made:

$$Z = 1$$

which is consistent with definition of partition function

## Boltzmann: reference system

$$\Delta E_{kl}^{ij} = E_{kl}^{ij} - \langle E_{kl} \rangle = - k\,T \ln \left[ \frac{p_{kl}^{ij}}{\sum_{ij} p_{kl}^{ij}} \right]$$

system described by four variables: i, j, k, l;
subset of variables: k, l; $\Delta E$: net potential of mean force;
note: net mean force energy contains only those components which are particular to the subsystem labelled i and j

Forces = partial derivatives of energies:

$$F_m^{ij} = \frac{\partial \Delta E_{kl}^{ij}}{\partial m}, \qquad \text{with } m = l, k$$



## Mean-force potentials for pair interactions

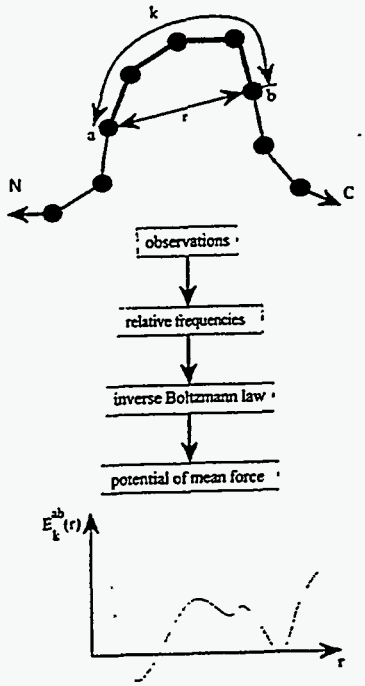
- Variables
  - amino acids: a, b
  - atom types: c, d
  - sequence separation: k
  - spatial distance: r
- thus, compilation of $f_{abcdkr}$ straightforward
- next: choice of subsystems and reference frame

$$\Delta E_r^{abcdk} = E_r^{abcdk} - \langle E_r^{cdk} \rangle = - k\,T \ln \left[ \frac{p_r^{abcdk}}{\sum_{ab} p_r^{abcdk}} \right]$$

$$F_r^{abcdk} = \frac{\partial \Delta E_r^{abcdk}}{\partial r}$$

– problem: sparse data

# Fig. 3.56: Mean-force: pair interactions




(Figure 2 from Sippl, 1993a)



# Fig. 3.57: Mean-force: potentials

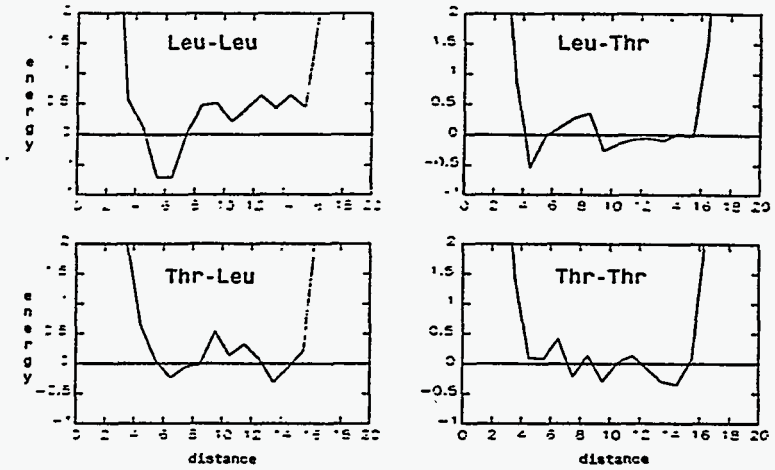



Fig. 3. Examples of $C^\beta$-$C^\beta$ mean force potentials for separation k = 4 along the amino acid sequence. Energies are scaled in the form E/kT. For small values of k particular values of r correlate strongly with local structures. The deep minimum of Leu-Leu at r ≈ 6 Å reflects the strong preference for α-helical structures. In contrast, α-helical conformations are energetically unfavourable for Thr-Thr. The mixed pairs are intermediate. Thr-Leu, for example, has two minima of comparable depth at α-helical and extended conformations.

(Figure 3 from Sippl, 1993a)

# Mean-force potentials: total energy

$$\Delta E_s^{ac} = -kT \ln\left[\frac{f_s^{ac}}{\sum_a p_s^{ac}}\right]$$

s number of atoms in a sphere of radius R around atom a;

$$\Delta S(S,C) = \sum_i \Delta E_s^{a(i)c}$$

$\Delta S(S,C)$ total surface energy of sequence S in conformation C;

$$\Delta E(S,C) = w_P \Delta P(S,C) + w_S \Delta S(S,C)$$

$$F(S,C) = \frac{\partial \Delta E}{\partial r \partial s}$$

total energy and total molecular force field



# Fig. 3.58: Mean-force: total energy

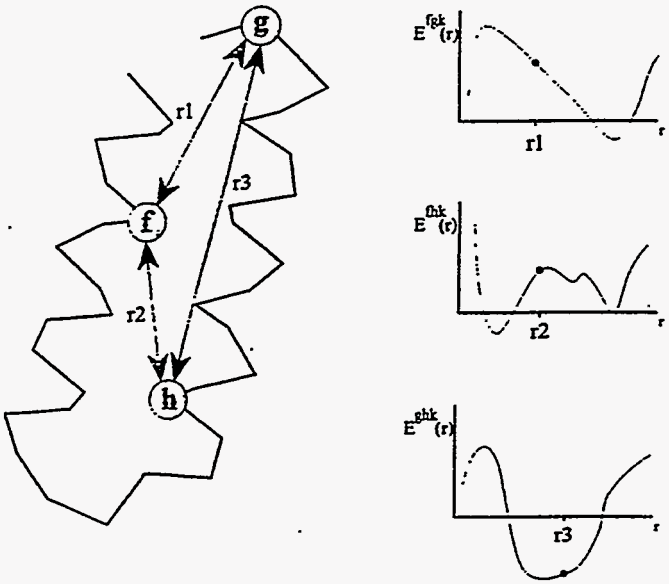



Fig. 5. Outline of the computation of the total pair interaction energy of proteins. The distances between atoms are calculated. The residue types a and b, atom types c and d, the separation k along the sequence determine the type of potential used to evaluate the energy at distance r. The total pair interaction energy is obtained by summing over all atom pairs in the molecule.

(Figure 5 from Sippl, 1993a)

# Predictive power of mean-force potentials

- reference system:
  - polyprotein
- goal:
  - evaluate likelihood of background, i.e., find system with lowest energy
- compute z-scores:

$$z_q = \frac{\Delta E(S,C_q) - \sum_q \Delta E(S,C_q)}{\sigma}$$

$C_q$ conformation q along the polyprotein; $\sigma$ standard deviation of the average energy over all conformations q;



# Fig. 3.59: Potentials for known structures

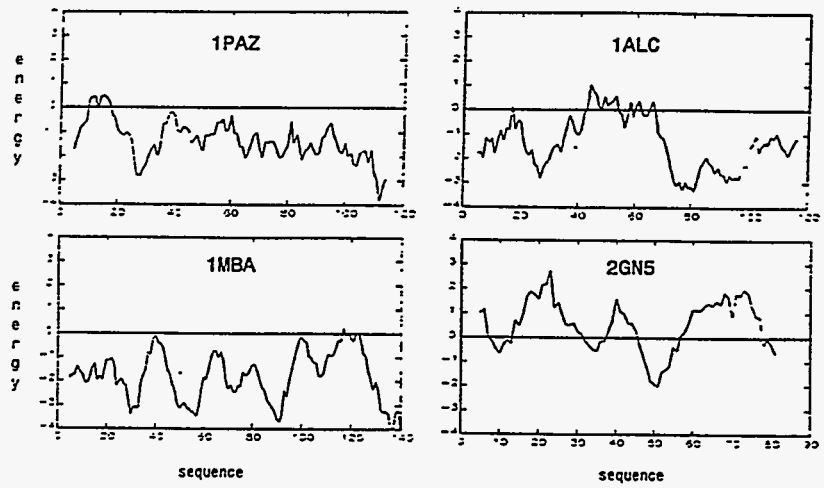
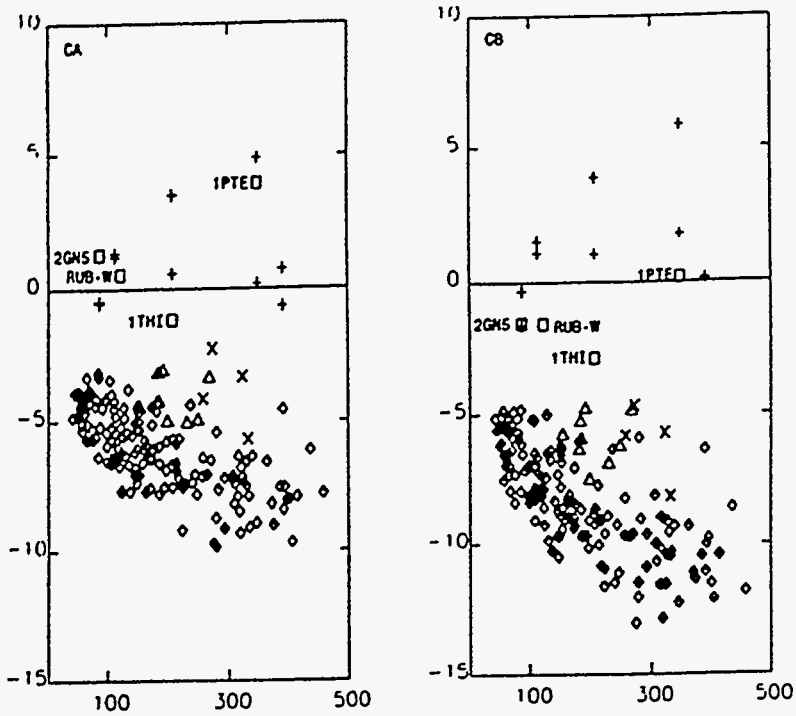



Fig. 11. Residue profiles for several protein structures determined by X-ray analysis. The energies were calculated from $C^\beta$ interactions only. In the plastocyanin (1PAZ), myoglobin (1MBA), and α-lactalbumin (1ALC) profiles the energy remains mostly below zero. Only occasionally we encounter small positive peaks. In contrast, the residue profile of 2GN5 contains large positive peaks. The conformation appears to be extremely strained. It is noteworthy that this strain is not a consequence of steric overlap. The energies for all distances r less than 5 Å were excluded from the calculations. The window used for gliding averages amounts to 10 residues.

(Figure 11 from Sippl, 1993a)

## Fig. 3.60: Mean-force energy z-scores for known structures




(Figure 1 from Sippl, 1993)



## Fig. 3.61: Potentials for 2GN5 and 1BGH

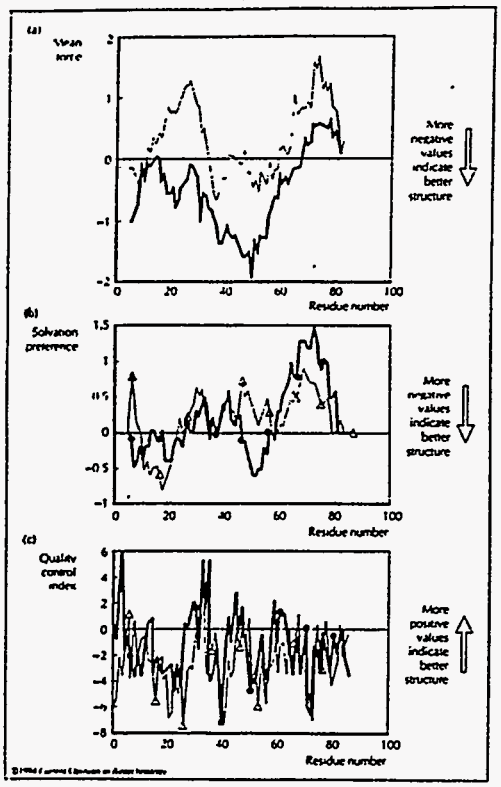



Fig. 4. Discrimination between native-like three-dimensional structures and incorrect model structures. Three methods plotting sequence against potential/pseudopotential are compared for the two structures of the DNA-binding gene V protein (Protein Data Bank datasets 2GN5 and 1BGH). (a) Distance-based potential of mean force (where more negative values indicate a better structure). For 2GN5 and 1BGH, the average over all residues was −2.53 and −4.95, respectively [87,94**]. (b) Atomic solvation preferences (where smaller values indicate a better structure). For 2GN5 and 1BGH, the average over all residues was 22.2 and 16.3, respectively [91]. (c) Contact-based quality control index (where more positive values indicate a better structure). For 2GN5 and 1BGH, the average over all residues was −2.75 and −1.43, respectively [93*]. All three methods identify 2GN5 as containing many errors. 1BGH may be inaccurate towards the carboxyl terminus. A similar qualitative result is found using a method [92] that relies on various structural features (backbone dihedral angles, bond lengths, planarity of rings and hydrogen-bonding patterns). On the basis of such differences, structures deposited in the data bank were predicted to contain errors [92,93*,94**].

(Figure 4 from Rost & Sander, 1994)

## Mean-force potentials: conclusion

- **Results**
  - » Accurate enough to spot incorrect structures
- **Applications**
  - » Post-processing prediction methods (e.g. threading)
  - » Site-directed mutations
  - » Selection of the best among an ensemble of possible structures
  - » Spot stresses in structures



## Remote homology modelling (threading)

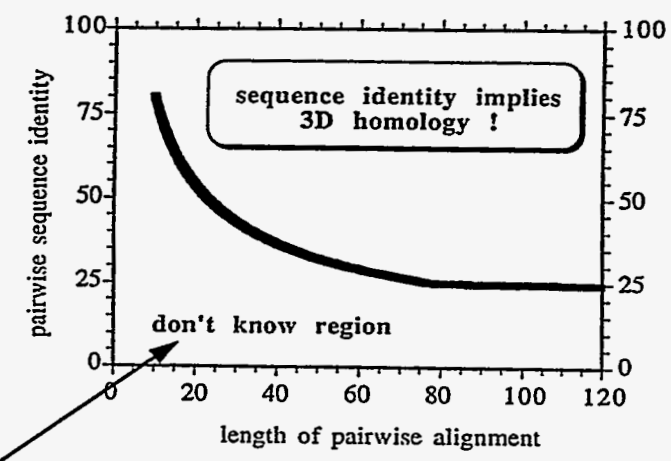
- **Goal and concept**
- **Methods**
  - **– Sippl potentials**
  - **– Fosfos potentials**
- **Improvement by evolutionary information**
- **Results**
  - » Potentials can retrieve the original structure
  - » Correct remote homologue often found
  - » Prediction of 3D structure seems to work sometimes
- **Evaluation**
  - » Evaluation of tools a shame!
  - » Prediction accuracy overemphasised in the past, ...
  - » but, methods will probably become increasingly important
- **Applications**
  - » If successful, same as for homology modelling

# Threading: goal

- 'simple' program:
  - » given sequence of unknown structure SOUS
  - » generate all possible conformations
  - » select best
- not so simple:
  - » semi-empirical force-fields cannot even distinguish the correct from a grossly misfolded structure, in general
    (Novotny et al., 1984; Novotny et al., 1988)
- alternative simplify potentials
  - » base distinction on inter-residue contacts or averages over contacts
- goal:
  fitness of sequence for structure (fosfos)



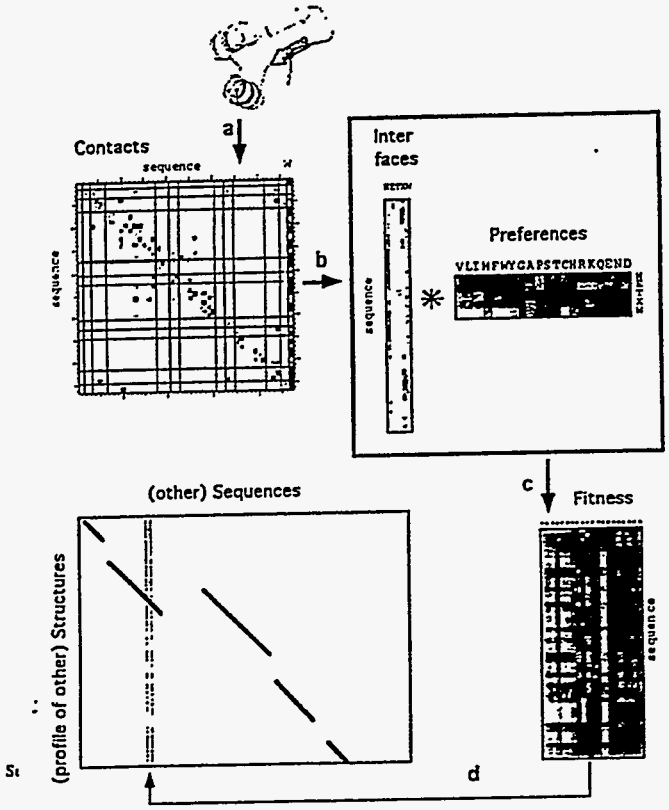
# Fig. 3.62: Remote homology




current PDB (3.000 structures):
some 5.000 pairs in
"don't know" = "remote homology" region

(Figure from Sander & Schneider, 1991)

## Fig. 3.63: Fosfos potentials - principle idea




(Figure 1 from Ouzouniz et al., 1993)



## Threading: fosfos potentials

- 3D - 1D potentials
  - simplest:
    hydrophobicity matching accessibility
    (Bowie et al., 1990)
  - more elaborated description:
    18 classes (accessibility, polarity, secondary str.)
    (Bowie et al., 1991; Lüthy et al., 1991)
  - contact interface potentials:
    29 classes
    » helix, strand, turn, rest
    » buried, intermediate, exposed
    » residue, solvent

    » + core weights: conserved and not exposed
    (Ouzounis et al., 1993)

# Fig. 3.64: Aligning accessibility potentials

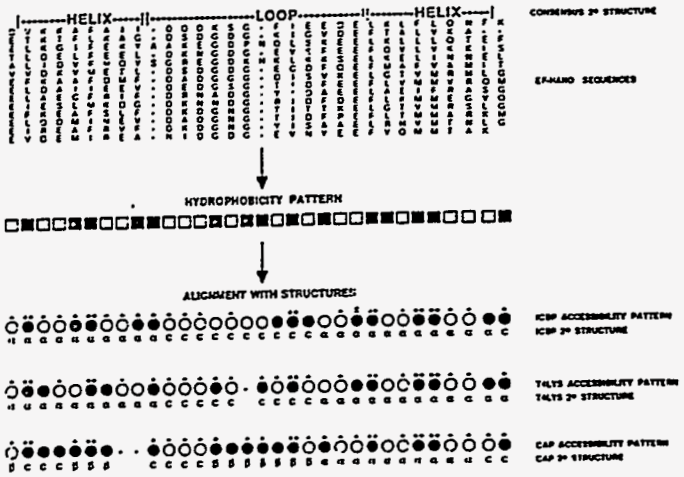



Fig. 6. Generation of the hydrophobicity pattern of EF-hand sequences and the alignments with the solvent accessibility patterns for the three proteins that gave the top scores. The amino acid sequences and consensus secondary structure are from Szebenyi et al.[14] The hydrophobicity pattern is shown below the sequences. Filled-in boxes correspond to members of the $H_1$ hydrophobicity group, shaded boxes to members of the $H_2$ group and open boxes to members of the $H_3$ group. The lower part of the figure shows the aligned solvent accessibility patterns of the proteins that gave the top three scores: vitamin D dependent calcium binding protein from bovine intestine (ICBP, residues 5 through 36), T4 lysozyme [T4-Lys; residues 41 through 71] and catabolite gene activating protein [CAP; residues 291 through 320]. Filled-in circles correspond to members of the $B_1$ group, shaded circles to members of the $B_2$ group, and open circles to members of the $B_3$ group. Gaps in the alignment are indicated by dashes. Symbols above the accessibility patterns indicate the quality of the matches: - - indicates a score of greater than 0.5; +, a score between 0 and 0.5, and an X, a score less than 0.5. The secondary structure of the three proteins is shown below each accessibility pattern: α denotes an α-helix; β, a β-strand; and C, a coil region.

(Figure 6 from Bowie et al., 1990)



# Fig. 3.65: Separating positives and false positives

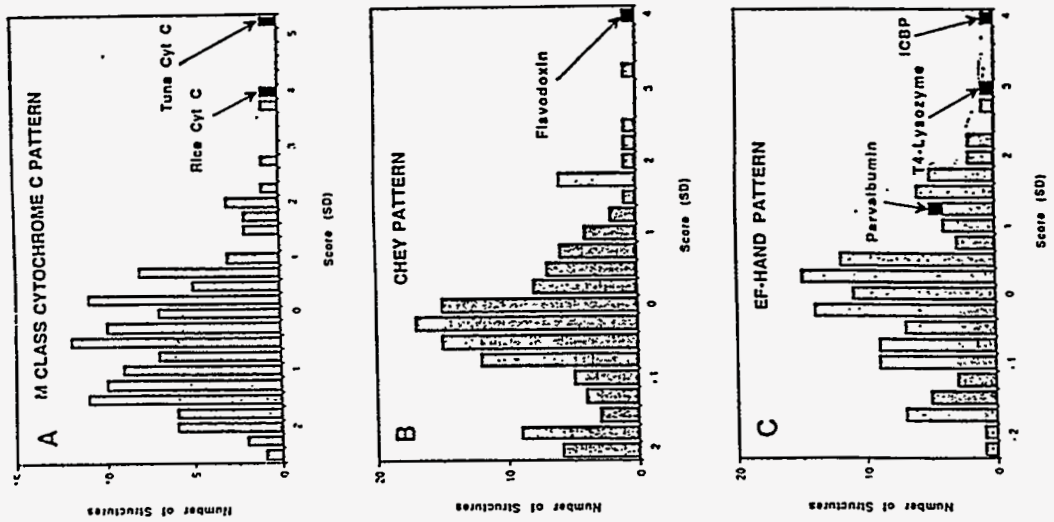
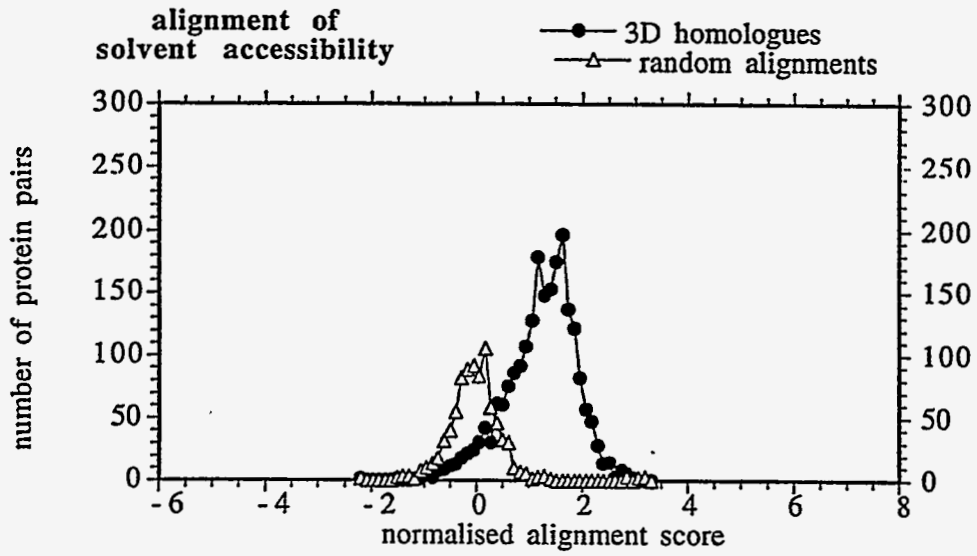



Fig. 4. Alignment scores (see the legend for Fig. 3) for the M class cytochromes c, the CheY protein family, and the EF-hand. The names, in order, of the top five scoring proteins for each histogram are: A) tuna cytochrome c, rice cytochrome c, cytochrome B562, superoxide dismutase, flavodoxin; B) flavodoxin, T4 lysozyme, cytochrome B562, yeast phosphoglycerate kinase, chymotrypsinogen A; C) the EF-hand: vitamin D dependent calcium binding protein, T4 lysozyme, catabolite gene activating protein, 434 repressor, arabinose binding protein.

(Figure 4 from Bowie et al., 1990)

## Fig. 3.66: Separating positives and false positives - more cases

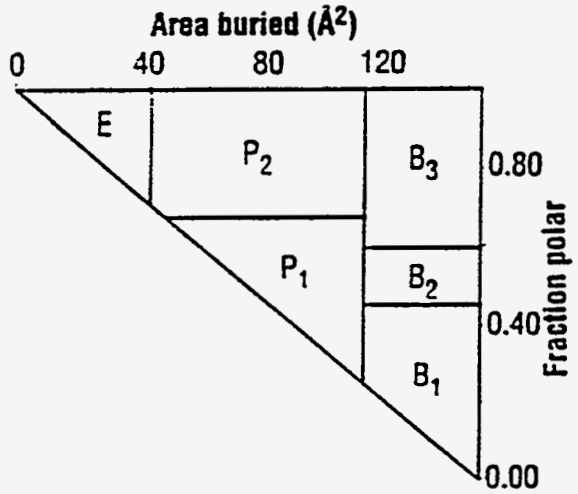



(Figure 6 from Rost, 1995a)



## Fig. 3.67: Bowie & Eisenberg potentials: classes

- 18 classes:
- 6 : accessibility and polarity
- 3: helix strand rest




(Figure 4 from Bowie et al., 1991)

# Fig. 3.68: Bowie & Eisenberg potentials

| Environment class | W | F | Y | L | I | V | M | A | G | P | C | T | S | Q | N | E | D | H | K | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $B_1\alpha$ | 1.00 | 1.22 | 0.18 | 1.27 | 1.17 | 0.66 | 1.26 | -0.66 | -2.53 | -1.16 | -0.73 | -1.29 | -2.73 | -1.08 | -1.93 | -1.74 | -1.97 | -0.34 | -1.82 | -1.67 |
| $B_1\beta$ | 1.17 | 0.85 | 0.07 | 1.13 | 1.47 | 1.09 | 0.55 | -0.79 | -2.02 | -0.94 | -0.22 | -1.12 | -2.91 | -1.57 | -1.42 | -1.93 | -2.56 | -1.91 | -2.69 | -1.16 |
| $B_1$ | 1.05 | 1.45 | 0.17 | 1.10 | 1.11 | 1.02 | 0.98 | -0.91 | -1.92 | 0.26 | -1.22 | -1.53 | -2.81 | -1.17 | -2.42 | -2.52 | -1.76 | -1.12 | -2.59 | -2.16 |
| $B_2\alpha$ | 0.50 | 0.90 | 0.85 | 1.01 | 0.63 | 0.68 | 1.12 | -0.69 | -1.49 | -2.21 | -0.10 | -1.50 | -1.47 | -0.23 | -0.81 | -0.71 | -1.62 | 0.23 | -0.78 | 0.06 |
| $B_2\beta$ | 0.01 | 1.18 | 1.06 | 0.78 | 1.31 | 1.06 | 0.64 | -1.55 | -2.26 | -0.49 | -0.87 | -2.27 | -1.77 | -1.22 | -2.07 | -1.07 | -1.41 | -0.77 | -1.14 | -0.20 |
| $B_2$ | 1.02 | 1.05 | 1.12 | 0.84 | 0.81 | 0.60 | 0.90 | -0.66 | -1.65 | 0.19 | -0.05 | -0.76 | -1.17 | -0.76 | -0.66 | -1.35 | -1.23 | 0.46 | -2.34 | -0.80 |
| $B_3\alpha$ | 0.92 | -0.03 | 0.58 | 0.15 | 0.04 | -0.02 | 0.89 | -0.57 | -1.88 | -0.68 | -1.56 | -0.57 | -0.96 | 0.22 | -0.08 | 0.08 | -0.50 | 0.73 | 0.43 | 0.96 |
| $B_3\beta$ | 0.75 | 0.81 | 1.30 | 0.18 | 0.54 | 0.56 | -0.57 | -0.93 | -1.93 | -0.34 | -0.54 | -0.44 | -0.74 | 0.21 | -0.24 | -0.14 | -0.88 | 0.82 | -0.53 | 0.13 |
| $B_3$ | 1.07 | 0.70 | 1.13 | 0.35 | -0.17 | -0.03 | 0.23 | -0.96 | -0.98 | -0.13 | -1.20 | -0.53 | -0.54 | 0.05 | 0.04 | -0.36 | -1.05 | 1.01 | 0.10 | 0.66 |
| $P_1\alpha$ | -1.35 | -0.82 | -0.59 | -0.52 | -0.24 | 0.10 | -0.03 | 0.73 | -0.49 | -0.25 | 0.95 | 0.31 | 0.34 | -0.14 | -0.54 | -0.17 | -0.25 | -0.52 | -0.21 | -0.28 |
| $P_1\beta$ | 0.38 | -0.49 | 0.17 | -1.03 | 0.20 | 0.48 | -0.27 | 0.64 | -0.82 | -0.55 | 1.49 | 0.93 | 0.33 | -2.27 | -1.32 | -0.73 | -1.07 | -0.42 | -1.21 | -0.77 |
| $P_1$ | -1.28 | -1.20 | -1.31 | -0.82 | -0.23 | -0.01 | -1.19 | 0.46 | -0.24 | 0.68 | 1.35 | 0.56 | 0.49 | -0.63 | -0.13 | -0.81 | 0.38 | -1.12 | -0.74 | -1.29 |
| $P_2\alpha$ | -1.14 | -1.43 | -0.79 | -0.35 | -0.54 | -0.48 | -0.45 | 0.06 | -0.50 | -0.26 | -0.93 | -0.05 | -0.18 | 0.55 | -0.05 | 0.56 | 0.28 | 0.06 | 0.81 | 0.50 |
| $P_2\beta$ | -0.79 | -0.54 | -0.84 | -1.30 | -0.33 | 0.13 | -0.72 | -0.55 | -0.98 | -1.29 | -0.57 | 0.84 | 0.59 | -0.08 | -0.16 | 0.32 | 0.19 | -0.87 | 0.59 | 0.10 |
| $P_2$ | -0.82 | -0.86 | -0.51 | -0.70 | -1.09 | -0.88 | -0.89 | -0.15 | -0.40 | 0.44 | -0.60 | 0.06 | 0.26 | 0.27 | 0.50 | 0.27 | 0.49 | 0.13 | 0.44 | 0.30 |
| $E\alpha$ | -1.35 | -2.20 | -2.10 | -1.58 | -2.76 | -1.10 | -0.72 | 0.46 | 0.68 | 0.04 | -0.44 | -0.17 | 0.15 | 0.38 | 0.28 | 0.59 | 0.44 | -0.19 | 0.13 | -0.34 |
| $E\beta$ | 0.64 | -0.90 | 0.30 | -1.66 | -1.47 | -1.74 | -0.68 | 0.06 | 1.46 | -0.96 | -0.24 | 0.14 | 0.65 | -0.19 | -0.06 | -0.16 | -0.78 | -0.83 | -0.52 | -0.49 |
| $E$ | -2.14 | -1.90 | -0.94 | -1.19 | -1.61 | -0.91 | -1.67 | 0.12 | 1.13 | 0.20 | -0.46 | 0.12 | 0.32 | -0.03 | 0.41 | 0.03 | 0.22 | -0.25 | -0.14 | -0.32 |

Fig. 5. The 3D-1D scoring table. The scores for pairing a residue $i$ with an environment $j$ is given by the information value (61).

$$\text{3D-1D score } ij = \ln\left(\frac{P(i;j)}{P_i}\right)$$

where $P(i;j)$ is the probability of finding residue $i$ in environment $j$ and $P_i$ is the overall probability of finding residue $i$ in any environment. These probabilities were determined from a database of 16 known protein structures and sets of homologous sequences aligned to the sequence of known structure as described in Luthy *et al.* (23). For each position in the aligned set of sequences, we determined the environment category of the position from the known structure and counted the number of each residue type found at the position within the set of aligned sequences. A residue type was counted only once per position. For example, if there were ten aspartates and one glycine found at a position in a set of aligned sequences, then both the Asp and Gly counters were both incremented by only one. The total number of residue replacements in our database was 8273. If the number of residues $i$ in an environment $j$ was found to be zero, the number was increased to one so that $P(i;j)$ was never zero. Boundaries for the environment categories (shown in Fig. 3) were adjusted iteratively to maximize the total 3D-1D score summed over all residues in our database:

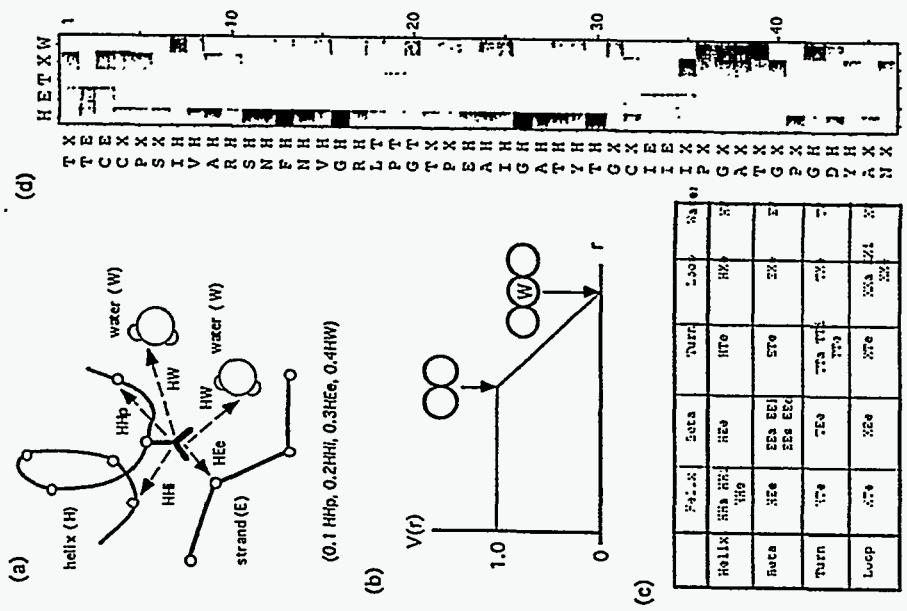
$$\text{Total 3D-1D score} = \sum_{ij} N_{ij} \ln\left(\frac{P(i;j)}{P_i}\right)$$

where $N_{ij}$ is the number of residues $i$ in environment $j$. In this case, if $N_{ij}$ was zero, the number was not increased to one. Instead, that term in the sum was treated as zero.

(Figure 5 from Bowie et al., 1991)

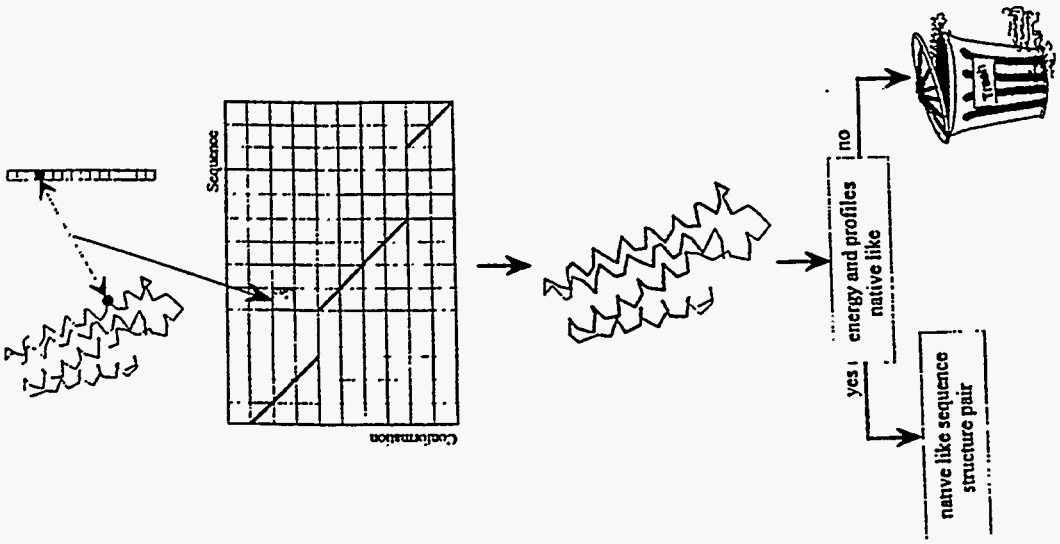


# Fig. 3.69: Fosfos potentials




(Figure 2 from Ouzouniz et al., 1993)

# Fig. 3.70: Sippl potentials




(Figure 14 fromSippl, 1993a)



# Fig. 3.71: Threading: a non-trivial problem

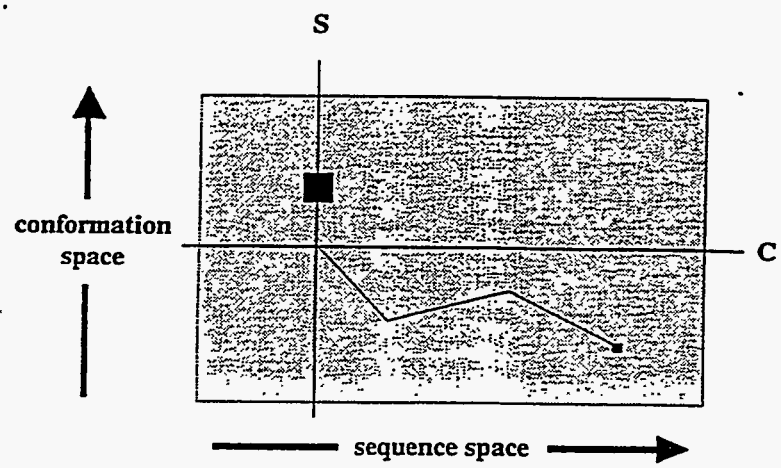
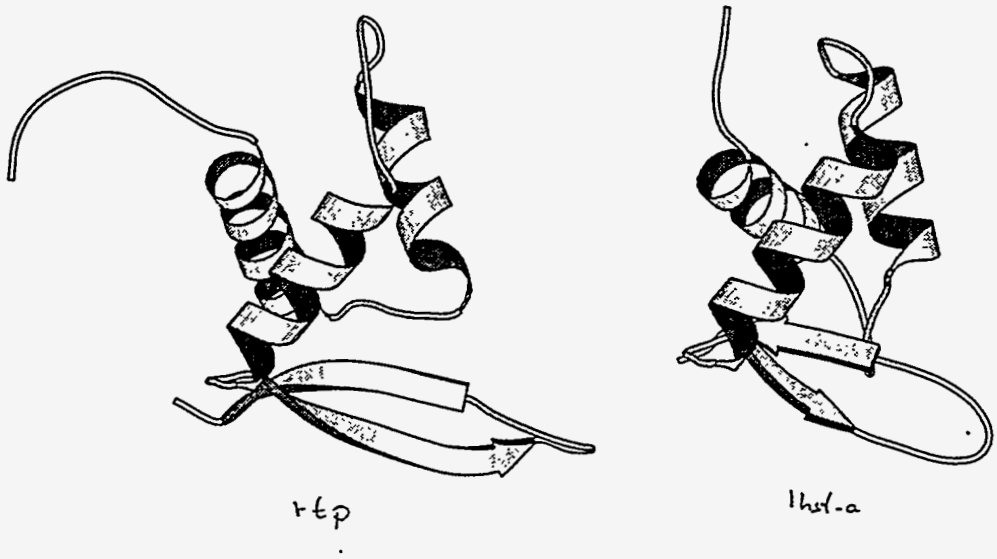



Fig. 16. Sequence structure alignment in terms of sequence structure space. In the alignment of sequence S with conformation C there is no guarantee that we find a native-like sequence structure pair. The introduction of gaps changes the sequence S, hence the folding postulate is not applicable, and the chances are that the alignment goes astray. Nonnative-like alignments are unmasked by their nonnative energies and profiles but it may be impossible to find the native-like pair (filled square) even if we start at a point (S,C) in its immediate neighbourhood.

(Figure 14 fromSippl, 1993a)

## Fig. 3.72: One successful 3D prediction




(Figure 1 from Flöckner et al., 1995)



## Threading: conclusion

- **Results**
  - » Potentials can retrieve the original structure
  - » Correct remote homologue often found
  - » Prediction of 3D structure seems to work sometimes
- **Evaluation**
  - » Evaluation of tools a shame!
  - » Prediction accuracy overemphasised in the past, ...
  - » but, methods will probably become increasingly important
- **Applications**
  - » If successful, same as for homology modelling

# Acknowledgement


- Chris Sander, EMBL
- Reinhard Schneider, EMBL

- Gerritt Vriend, EMBL
- Antoine de Daruvar, Christos Ouzounis, EMBL
- THE GROUP (Protein Design Group, EMBL)

- Manfred Sippl, Salzburg
- Michael Braxenthaler, CARB, Washington D.C.
- SØren Brunak, Jacob Engelbrecht, Copenhagen
- Tim Hubbard, MRC, Cambridge, U.K.

## Abbreviations used

| | |
|---|---|
| 1D | one-dimensional |
| 2D | two-dimensional |
| 3D | three-dimensional |
| 4D | four-dimensional |
| ADP/ATP | Adenosine Di-Phospate/Adenosine Tri-Phospate, the reaction: ATP->ADP releases ≈ 7kcal/mol. |
| CW | Continuous wave |
| DG | Distance geometry |
| DSA | Dynamical simulated annealing |
| DSSP | data base containing the secondary structure and solvent accessibility for proteins of known 3D structure (Kabsch & Sander 1983a) |
| FSSP | data base of remote homologues of known 3D structure (Holm, et al. 1993, Holm & Sander 1993, Holm & Sander 1994a) |
| FT | Fourier transform |
| GOR | prediction of secondary structure based on statistics (Garnier, et al. 1978, Gibrat, et al. 1987, Biou, et al. 1988) |
| HM | Homology Modelling: modelling the 3D structure of a protein based on a significant level of pairwise sequence identity to a protein of known 3D structure |
| HSSP | data base containing for each PDB protein of known 3D structure the alignments of all SWISSPROT sequences homologue to the known structure (Sander & Schneider 1991, Sander & Schneider 1994). |
| HTM | Trans-Membrane-helix, helix crossing the lipid bilayer of integral transmembrane proteins |
| MD | Molecular dynamics |
| NMR | Nuclear Magnetic Resonance |
| NOE | Nuclear Overhauser effect |
| PDB | Protein Data Bank of experimentally determined 3D structures of proteins (Bernstein, et al. 1977, Abola, et al. 1988). |
| PHD | Profile based neural network prediction of |
| PHDacc | solvent accessibility (PHDacc; (Rost & Sander 1994c, Rost 1995b)), and |
| PHDhtm | transmembrane helices (PHDhtm; (Rost 1995b, Rost, et al. 1995)). |
| PHDsec | secondary structure (PHDsec; (Rost & Sander 1994b, Rost 1995b)), |
| RMSD | Root-mean-square deviation |
| SOUS | Sequence Of Unknown Structure. |
| SWISSPROT | data base of protein sequences (Bairoch & Boeckmann 1994). |
| TM | Trans-Membrane, region bound to lipid bilayer of integral trans-membrane proteins. |
| XRC | X-ray crystallography |

## Sources of Figures

### Introduction

| | |
|---|---|
| Fig. 1.1 | Basic tetrahedron of all amino acids (Rost 1993) |
| Fig. 1.2 | The 20 amino acids (Rost 1993) |
| Fig. 1.3 | Biosynthesis of amino acids to poly-peptides (Rost 1993) |
| Fig. 1.4 | The dihedral angles (Rost 1993) |
| Fig. 1.5 | Simplified view of protein folding (Sander, et al. 1992) |
| Fig. 1.6 | Chaperone mediated protein folding (Martin & Hartl 1993) |
| Fig. 1.7 | Hydrogen bond pattern of helix (Schulz & Schirmer 1979) |
| Fig. 1.8 | Hydrogen bond patterns of strand (Schulz & Schirmer 1979) |
| Fig. 1.9 | Calcium binding motif: helix-loop-helix (Brändén & Tooze 1991) |
| Fig. 1.10 | Greek-key motif: four strands (Brändén & Tooze 1991) |
| Fig. 1.11 | Relationship between structural homolo-gy and sequence identity (Rost & Sander 1994b) |
| Fig. 1.12 | Protein jigsaw puzzle (Taylor 1992) |
| Fig. 1.13 | Relation between structural homology and sequence identity (Sander & Schneider 1991) |

### Determination methods

| | |
|---|---|
| Fig. 2.1 | The growth of the protein data bank |
| Fig. 2.2 | The PDB format - showing its age |
| Fig. 2.3 | Chemical shifts for different hyrdogen atoms in peptides (Creighton, 1993) |
| Fig. 2.4 | Continuous wave vs Fourier transform spectroscopy (Ernst, 1994) |
| Fig. 2.5 | Pulse sequences - a black art (Ernst, 1994) |
| Fig. 2.6 | Schematic representations of 2D spectra (Ernst, 1994) |

## Prediction methods

## References

Abagyan, R.; Frishman, D. and Argos, P. 1994. Recognition of distantly related proteins through energy calculations. *Proteins* 19:132-140.

Abagyan, R. & Totrov, M. 1994. Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins. *J. Mol. Biol.* 235:983-1002.

Abagyan, R.; Totrov, M. and Kuznetsov, D. 1994. ICM - A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* 15:488-506.

Abola, E. E.; Bernstein, F. C. and Koetzle, T. F. 1988. The Protein Data Bank. In Lesk A. M., E. eds. *Computational molecular biology. Sources and methods for sequence analysis.* Oxford: Oxford University Press.

Alexandrov, N. N. 1992. Local multiple alignment by consensus matrix. *CABIOS* 8:339-345.

Altschuh, D.; Lesk, A. M.; Bloomer, A. C. and Klug, A. 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* 193:693-707.

Altschuh, D.; Vernet, T.; Moras, D. and Nagai, K. 1988. Coordinated amino acid changes in homologous protein families. *Prot. Engin.* 2:193-199.

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555-565.

Altschul, S. F. 1993. A protein Alignment Scoring System Sensitive at All Evolutionary Distances. *J. Mol. Evol.* 36:290-300.

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Andrade, M. A.; Chacón, P.; Merelo, J. J. and Morán, F. 1993. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Prot. Engin.* 6:383-390.

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

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science* 181:223-230.

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

Argos, P.; Rao, J. K. M. and Hargrave, P. A. 1982. Structural prediction of Membrane-bound proteins. *Eur. J. Biochem.* 128:565-575.

Asai, K.; Hayamizu, S. and Handa, K. 1993. Prediction of protein secondary structure by the hidden Markov model. *CABIOS* 9:141-146.

Bacon, D. J. & Anderson, W. F. 1990. Multiple sequence comparison. *Meth. Enzymol.* 183:438-447.

Bairoch, A. & Boeckmann, B. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucl. Acids Res.* 22:3578-3580.

Barton, G. J. & Russell, R. B. 1993. Protein structure prediction. *Nature* 361:505-506.

Bassolino, D. A.; Hirata, F.; Kitchen, D.; Kominos, D.; Pardi, A. & Levy, R. M. 1988. Determination of protein structures in solution using NMR data and IMPACT. *Int. J. Supercomput. Appl.* 2:41-61.

Bauer, A. & Beyer, A. 1994. An Improved Pair Potential to Recognize Native Protein Folds. *Proteins* 18:254-261.

Baumann, G.; Frömmel, C. and Sander, C. 1989. Polarity as a criterion in protein design. *Prot. Engin.* 2:329-334.

Benner, S. A. & Gerloff, D. 1990. Patterns of Divergence in Homologous Proteins as Indicators of Secondary and Tertiary Structure of the Catalytic Domain of Protein Kinases. *Adv. Enz. Reg.* 31:121-181.

Benner, S. A. 1992. Predicting de novo the folded structure of proteins. *Curr. Opin. Str. Biol.* 2:402-412.

Benner, S. A.; Badcoe, I.; Cohen, M. A. and Gerloff, D. L. 1994. Bona Fide Prediction of Aspects of Protein Conformation. *J. Mol. Biol.* 235:926-958.

Benner, S. A.; Cohen, M. A. and Gerloff, D. 1993. Predicted Secondary Structure for the Src Homology 3 Domain. *J. Mol. Biol.* 229:295-305.

Berendsen, H. J. C. 1991. Molecular dynamics studies of proteins and nucleic acids. *Curr. Opin. Str. Biol.* 1:191-195.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.

Biou, V.; Gibrat, J. F.; Levin, J. M.; Robson, B. and Garnier, J. 1988. Secondary structure prediction: combination of three different methods. *Prot. Engin.* 2:185-91.

Blout, E. R. 1962. The dependence of the conformation of polypetides and proteins upon amino acid composition. In Stahman, M. eds. *Polyamino Acids, Polypeptides, and Proteins.* Madison: Univ. of Wisconsin Press.

Blundell, T. L. & Johnson, M. S. 1993. Catching a common fold. *Prot. Sci.* 2:877-883.

Blundell, T. L., et al. 1988. Knowledge-based protein modelling and design. *Eur. J. Biochem.* 172:513-520.

Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M. J.; Lautrup, B.; Nørskov, L.; Olsen, O. H. and Petersen, S. B. 1988. Protein secondary structure and homology by neural networks. *FEBS Lett.* 241:223-228.

Bohr, H.; Bohr, J.; Brunak, S.; Fredholm, H.; Lautrup, B. and Petersen, S. B. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.* 261:43-46.

Bohr, J.; Bohr, H.; Brunak, S.; Cotterill, R. M. J.; Fredholm, H.; Lautrup, B. and Petersen, S. B. 1993. Protein Structures from Distance Inequaliteies. *J. Mol. Biol.* 231:861-869.

Bonvin, A. M. J. J.; Boelens, R. & Kaptein, R. 1994. Time- and ensemble-averaged direct NOE restraints. *J. Biomol. NMR* 4:143-149.

Bordo, D. 1993. ENVIRON: a software package to compare protein three-dimensional structures with homologous sequences using local structural motifs. *CABIOS* 9:639-645.

Bordo, D.; Djinovic, K. and Bolognesi, M. 1994. Conserved Patterns in the Cu, Zn Superoxide Dismutase Family. *J. Mol. Biol.* 238:366-386.

Bork, P. & Grundwald, C. 1990. Recognition of different nucleotide-binding sites in primary structures using a

property-pattern approach. *Eur. J. Biochem.* 191:347-358.

Bork, P. 1992. Mobile modules and motifs. *Curr. Biol.* 413-421.

Bork, P.; Ouzounis, C. and Sander, C. 1994. From genome sequences to protein function. *Curr. Opin. Str. Biol.* 4:393-403.

Bork, P.; Ouzounis, C.; Sander, C.; Scharf, M.; Schneider, R. and Sonnhammer, E. 1992a. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Prot. Sci.* 1:1677-1690.

Bork, P.; Ouzounis, C.; Sander, C.; Scharf, M.; Schneider, R. and Sonnhammer, E. 1992b. What's in a genome? *Nature* 358:287.

Bork, P.; Sander, C.; Valencia, A. and Bukau, B. 1992c. A module of the DnaJ heat shock proteins found in malaria parasites. *TIBS* 129.

Bossa, F. & Pascarella, S. 1990. PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. *CABIOS* 5:319-320.

Bowie, J. U.; Clarke, N. D.; Pabo, C. O. and Sauer, R. T. 1990. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins* 7:257-264.

Bowie, J. U.; Lüthy, R. and Eisenberg, D. 1991. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* 253:164-169.

Boyd, D. & Beckwith, J. 1990. The role of charged amino acids in the localization of secreted and membrane proteins. *Cell* 62:1031-1033.

Brändén, C. & Tooze, J. 1991. *Introduction to Protein Structure*. New York, London: Garland Publ.

Braun, W. & Go, N. 1985. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* 186:611-626.

Braun, W. & Gö, N. 1985. Calculation of Protein Conformations by Proton-Proton Distance Constraints. *J. Mol. Biol.* 186:611-626.

Bricogne, G. 1984. Maximum entropy and the foundation of direct methods. *Acta Crystallogr. A* 40:410-445.

Bricogne, G. 1991. Maximum entropy as a common statistical basis for all phase determination methods. In *Crystallographic computing 5, From chemistry to biology* D. Moras; A. D. Podjarny & J. C. Thierry eds. New York: Oxford University Press.

Brooks, C. L.; Karplus, M. and Pettitt, B. M. 1988. *Proteins: A theoretical perspective of dynamics, structure, and thermodynamics*. New York: John Wiley & Sons.

Brünger, A. T. & Nilges, M. 1993. Computational challenges for macromolecular structure determination by X-ray crystallography and solutioin NMR-spectroscopy. *Quart. Rev. Biophys.* 26:49-125.

Brünger, A. T. 1992. The free *R* factor: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355:472-474.

Brünger, A. T. 1993. Assessment of phase accuracy and cross-validation: the free *R* value. Methods and applications. *Acta Crystallogr. D* 49:24-36.

Brünger, A. T.; Clore, G. M.; Gronenborn, A. M. & Karplus, M. 1986. Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: Applications to crambin. *Proc. Natl. Acad. Sci. USA* 83:3801-3805.

Brünger, A. T.; Clore, G. M.; Gronenborn, A. M. & Karplus, M. 1987. Solution conformations of human growth hormone releasing factor: comparison of the restrained molecular dynamics and distance geometry methods for a system without long-range distance data. *Protein Engng* 1:399-406.

Brünger, A. T.; Clore, M. G.; Gronenborn, A. M. and Karplus, M. 1986. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc. Natl. Acad. Sc. U.S.A.* 83:3801-3805.

Brünger, A. T.; Clore, M. G.; Gronenborn, A. M.; Saffrich, R. & Nilges, M. 1993. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science* 261:328-331.

Brünger, A. T.; Kuriyan, J. & Karplus, M. 1987. Crystallographic *R* factor refinement by molecular dynamics. *Science* 235:458-460.

Brüschweiler, R. & Case, D. A. 19894 Charaterization of biomolecular structure dynamics by NMR cross relaxation. *Prog NMR Spectr.* 26:27-58.

Bryant, S. H. & Lawrence, C. E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16:92-112.

Bryant, S. H. 1989. PKB: a program system and database for analysis of protein structure. *Proteins* 5:233-247.

Byrne, D.; Li, J.; Platt, E.; Robson, B. & Weiner, P. 1994. Novel algorithms for searching conformational space. *J. Comput.-Aided Mol. Design* 8:67-82.

Casadio, R.; Fariselli, P.; Taroni, C. and Compiani, M. 1994. A predictor of transmembrane α-helix domains of proteins based on neural networks. *European Journal of Biophysics* submitted, 8/94.

Casari, G. & Sippl, M. J. 1992. Structure-derived Hydrophobic Potential. *J. Mol. Biol.* 224:725-732.

Chandonia, J.-M. & Karplus, M. 1995. Neural Networks for secondary structure and structural class predictions. *Prot. Sci.* 4:275-285.

Chiche, L.; Gregoret, L. M.; Cohen, F. E. and Kollman, P. A. 1990. Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sc. U.S.A.* 87:3240-3243.

Chothia, C. & Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823-826.

Chothia, C. 1976. The Nature of the Accessible and Buried Surfaces in Proteins. *J. Mol. Biol.* 105:1-12.

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

Chou, P. Y. & Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47:45-148.

Chou, P. Y. & Fasman, U. D. 1974. Prediction of protein conformation. *Biochem.* 13:211-215.

Chou, P. Y. 1989. Prediction of protein structural classes from amino acid composition. In D., F. G. eds. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press.

Claverie, J.-M. & Daulmiere, C. 1991. Smoothing profiles with sliding windows: better to wear a hat! *CABIOS* 7:113-115.

Cohen, F. E. & Kuntz, I. D. 1989. Tertiary Structure Prediction. In Fasman, G. D. eds. *Prediction of Protein Structure and the Principles of Protein Conformation*. New York, London: Plenum Press.

Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D. and Fletterick, R. J. 1983. Secondary Structure Assignment for a/b Proteins by a Combinatorial Approach. *Biochem.* 22:4894-4904.

Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D. and Fletterick, R. J. 1986. Turn Prediction in Proteins Using a Pattern-Matching Approach. *Biochem.* 25:266-275.

Cohen, F. E.; Sternberg, M. J. E. and Taylor, W. R. 1980. Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature* 285:378-382.

Colloc'h, N.; Etchebest, C.; Thoreau, E.; Henrissat, B. and Mornon, J.-P. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Prot. Engin.* 6:377-382.

Cornell, W. D.; Howard, A. E. and Kollman, P. 1991. Molecular mechanical potential functions and their application to study molecular systems. *Curr. Opin. Str. Biol.* 1:201-212.

Cornette, J. L.; Cease, K. B.; Margalit, H.; Spouge, J. L.; Berzofsky, J. A. and DeLisi, C. 1987. Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins. *J. Mol. Biol.* 195:659-685.

Cowan, S. W. & Rosenbusch, J. P. 1994. Folding pattern diversity of integral membrane proteins. *Science* 264:914-916.

Creighton, T. 1984. *Proteins: Structures and molecular properties*. New York: W. H. Freeman.

Creighton, T. E. 1991. Unfolding protein folding. *Nature* 352:17-18.

Crippen, G. M. & Maiorov, V. N. 1994. A Potential Function that Identifies Correct Protein Folds. In Bohr, H. and Brunak, S. eds. *Protein Structure by Distance Analysis*. Amsterdam, Oxford, Washington: IOS Press.

Crippen, G. M. 1977. Correlation of sequence and tertiary structure in globular proteins. *Biopolymers* 16:2189-2201.

Crippen, G. M. 1989. Linearized embedding: a new metric matrix method for calculating molecular conformations subject to geometric constraints. *J. Comput. Chem.* 10:896-902.

Crippen, G., M. 1991. Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces. *Biochem.* 30:4232-4237.

Dalbey, R. E. 1990. Positively charged residues are important determinants of membrane protein topology. *TIBS* 15:253-257.

Dao-pin, S.; Baase, W. A. and Matthews, B. W. 1990. A mutant T4 lysozyme (Val131->Ala) designed to increase thermostability by the reduction of strain within an alpha-helix. *Proteins* 7:198-204.

Dao-pin, S.; Sauer, U.; Nicholson, H. and Matthews, B. W. 1991a. Contributions of surface salt bridges to the stability of bacteriophage T4 lysozyme determined by directed mutagenesis. *Biochem.* 30:7142-7153.

Dao-pin, S.; Söderlind, E.; Baase, W. A.; Wozniak, J. A.; Sauer, U. and Matthews, B. W. 1991b. Cumulative site-directed charge-change replacements in bacteriophage T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability. *J. Mol. Biol.* 221:873-887.

Darwin, C. 1859. *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Davies, D. R. 1964. A correlation between amino acid composition and protein structure. *J. Mol. Biol.* 9:605-609.

De Filippis, V.; Sander, C. and Vriend, G. 1994. Predicting local structural changes that result from point mutations. *Prot. Engin.* 7:1203-1208.

Defay, T. & Cohen, F. E. 1995. Evaluation of current techniques for ab-initio protein structure prediction. *Proteins* in press.

Degli Esposti, M.; Crimi, M. and Venturoli, G. 1990. A critical evaluation of the hydropathy profile of membrane proteins. *Eur. J. Biochem.* 190:207-219.

Deisenhofer, J.; Epp, O.; Mii, K.; Huber, R. and Michel, H. 1985. Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution. *Nature* 318:618-624.

Deperieux, E. & Feytmans, E. 1992. MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *CABIOS* 8:501-509.

Dickerson, R. E.; Timkovich, R. and Almassy, R. J. 1976. The Cytochrome Fold and the Evolution of Bacterial Energy Metabolism. *J. Mol. Biol.* 100:473-491.

Dill, K. A. 1993. Folding proteins: finding a needle in a haystack. *Curr. Opin. Str. Biol.* 3:99-103.

Donnelly, D.; Overington, J. P. and Blundell, T. L. 1994. The prediction and orientation of a-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules. *Prot. Engin.* 7:645-653.

Doolittle, R. F. & Bork, P. 1993. Evolutionarily Mobile Modules in Proteins. *Scient. Am.* October:50-56.

Doolittle, R. F. 1986. *Of URFs and ORFs: a primer on how to analyze derived amino acid sequences*. Mill Valley California: University Science Books.

Doolittle, R. F. 1994. Convergent evolution: the need to be explicit. *TIBS* 19:15-18.

Dumas, J.-P. & Ninio, J. 1982. Efficient algorithms for folding and comparing nucleic acid sequences. *Nucl. Acids Res.* 10:197-206.

Dunbrack, R. L. & Karplus, M. 1993. Backbone-dependant rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* 230:543-574.

Edelman, J. 1993. Quadratic minimization of predictors for protein secondary structure: application to transmembrane α-helices. *J. Mol. Biol.* 232:165-191.

Eisenberg, D. & McLachlan, A. D. 1986. Solvation energy in protein folding and binding. *Nature* 319:199-203.

Eisenberg, D.; Lüthy, R. and McLachlan, A. D. 1991. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* 10:229-239.

Eisenberg, D.; Schwartz, E.; Komaromy, M. and Wall, R. 1984a. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 179:125-142.

Eisenberg, D.; Weiss, R. M. and Terwilliger, T. C. 1984b. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sc. U.S.A.* 81:140-144.

Eisenmenger, F.; Argos, P. and Abagyan, R. 1993. A Method to Configure Protein Side-chains from the Main-chain Trace in Homology Modelling. *J. Mol. Biol.* 231:849-860.

Engelman, D. M.; Steitz, T. A. and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual Review of Biophysics and Biophysical Chemistry* 15:321-353.

Epstein, C. J.; Goldberger, R. F. and Anfinsen, C. B. 1963. The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harbour Symp. Quant. Biol.* 28:439-449.

Ernst, R. R. 1994. Nuclear magnetic resonance fourier transform spectroscopy (Nobel lecture). *Bull. Magn. Reson.* 16:5-32.

Ernst, R. R.; Bodenhausen, G. & Wokaun, A. 1990. *Principles of nuclear magnetic resonance in one and two dimensions*. Oxford: Oxford University Press .

Esposito, G.; Lesk, A. M.; Molinari, H.; Motta, A.; Niccolai, N. and Pastore, A. 1994. Probing Protein Structure by Solvent Perturbation of NMR Spectra: III. Combination of Experiment and Theory. In Bohr, H. and Brunak, S. eds. *Protein Structure by Distance Analysis.* Amsterdam, Oxford, Washington DC: IOS Press.

Ewbank, J. J. & Creighton, T. E. 1991. The molten globule protein conformation probed by disulphide bonds. *Nature* 350:518-520.

Ewbank, J. J. & Creighton, T. E. 1992. Protein folding by stages. *Curr. Opin. Str. Biol.* 2:347-349.

Ewbank, J. J. 1992. The conformational dependence of disulphide bond formation and reduction in α-lactalbumin, Ph. D. Thesis, Trinity College, Cambridge.

Ewbank, J. J.; Creighton, T.; Hayer-Hartl, M. K. and Hartl, F. U. 1995. What is the molten globule? *Nature Struct. Biol.* 2:10.

Farber, G. K. & Petsko, G. A. 1990. The evolution of α/β barrel enzymes. *TIBS* 15:228-234.

Fariselli, P.; Compiani, M. and Casadio, R. 1993. Predicting secondary structures of membrane proteins with neural networks. *Eur. Biophys. J.* 22:41-51.

Fasman, G. D. 1989a. The development of the prediction of protein structure. In Fasman, G. D. eds. *Prediction of protein structure and the principles of protein conformation.* New York, London: Plenum Press.

Fasman, G. D. 1989b. *Prediction of Protein Structure and the Principles of Protein Conformation* . New York, London: Plenum.

Fasman, G. D. 1990. The prediction of the secondary structure of proteins: fact or fiction. *Curr. Science* 59:839-845.

Finkelstein, A. V. & Nakamura, H. 1993. Weak points of antiparallel β-sheets. How are they filled up in globular proteins? *Prot. Engin.* 6:367-372.

Finkelstein, A. V. & Reva, B. A. 1991. A search for the most stable folds of protein chains. *Nature* 351:497-499.

Finkelstein, A. V. & Reva, B. A. 1992. Search for the stable state of a short chain in a molecular field. *Prot. Engin.* 5:617-624.

Finkelstein, A. V.; Gutun, A. M. and Badretdinov, A. Y. 1993. Why are the same protein folds used to perform different functions? *FEBS Lett.* 325:23-28.

Flöckner, H.; Braxenthaler, M.; Lackner, P.; Jaritz, M.; Ortner, M. and Sippl, M. J. 1995. Progress in fold recognition, preprint, Center for Applied Molecular Engineering; Inst. Chemistry and Biochem.; Univ. Salzburg; Jakob Haringer Str. 1; A-5020 Salzburg; Austria.

Flores, T. P.; Orengo, C. A.; Moss, D. S. and Thornton, J. M. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Prot. Sci.* 2:1811-1826.

Forster, M. J. & Mulloy, B. 1994. Rationalizing nuclear Overhauser effect data for compounds adopting multiple-solution conformations. *J. Comput. Chem.* 15:155-161.

Fortier, S.; Castleiden, J.; Glasgow, J.; Conklin, D.; Walmsley, C.; Leherte, L. & Allen, F. H. 1993. Molecular scene analysis: the integration of direct methods and artifical intelligence stategies for solving protein crystal structures. *Acta Crystallogr. D* 49:168-178.

Galaktionov, S. G. & Marshall, G. R. 1994. Properties of Intraglobular Contacts in Proteins: An Approach to Prediction of Tertiary Structure. In 27th Hawaii International Conference on System Sciences, 326-335. Wailea, HI, U.S.A.: IEEE Computer Society Press.

Galaktionov, S. G. & Rodionov, M. A. 1980. Calculation of the tertiary structure of proteins on the basis of analysis of the matrices of contacts between amino acid residues. *Biophysics* 25:395-403 (translation of Biofizika, 1980, 25:385-392).

Galaktionov, S. G. & Rodionov, M. A. 1981. Calculation of the tertiary structure of proteins on the basis of analysis of the matrices of contacts between amino acid residues. *Biophysics* 25:395-403.

Garnier, J. & Levin, J. M. 1991. The protein structure code: what is its present status? *CABIOS* 7:133-142.

Garnier, J.; Osguthorpe, D. J. and Robson, B. 1978. Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *J. Mol. Biol.* 120:97-120.

Gascuel, O. & Golmard, J. L. 1988. A simple method for predicting the secondary structure of globular proteins: implications and accuracy. *CABIOS* 4:357-365.

Gerstein, M.; Sonnhammer, E. L. L. and Chothia, C. 1994. Volume Changes in Protein Evolution. *J. Mol. Biol.* 236:1067-1078.

Gibrat, J.-F.; Garnier, J. and Robson, B. 1987. Further Developments of Protein Secondary Structure Prediction Using Information Theory. New Parameters and Consideration of Residue Pairs. *J. Mol. Biol.* 198:425-443.

Gibson, T. J.; Thompson, J. D. and Abagyan, R. A. 1993. Proposed structure for the DNA-binding domain of the Helix-Loop-Helix family of eukaryotic gene regulatory proteins. *Prot. Engin.* 6:41-50.

Godzik, A. & Skolnick, J. 1992. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sc. U.S.A.* 89:12098-12102.

Godzik, A.; Kolinski, A. and Skolnick, J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* 227:227-238.

Goebel, U.; Sander, C.; Schneider, R. and Valencia, A. 1994. Correlated mutations and residue contacts in proteins. *Proteins* 18:309-317.

Goldstein, R. A.; Luthey-Schulten, Z. A. and Wolynes, P. 1994. A Bayesian Approach to Sequence Alignment Algorithms for Protein Structure Recognition. In 27th Hawaii International Conference on System Sciences, 306-315. Wailea, HI, U.S.A.: IEEE Computer Society Press.

Goldstein, R. A.; Luthey-Schulten, Z. A. and Wolynes, P. G. 1992. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sc. U.S.A.* 89:9029-9033.

Green, P.; Lipman, D.; Hillier, L.; Waterston, R.; States, D. and Claverie, J.-M. 1993. Ancient Conserved Regions in New Gene Sequences and the Protein Databases. *Science* 259:1711-1715.

Greer, J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sc. U.S.A.* 77:3393-3397.

Greer, J. 1981. Comparative Model-building of the Mammalian Serine Proteases. *J. Mol. Biol.* 153:1027-1042.

Greer, J. 1990. Comparative Modeling Methods: Application to the Family of the Mammalian Serine Proteases. *Proteins* 7:317-334.

Greer, J. 1991. Comparative modeling of homologous proteins. *Meth. Enzymol.* 202:239-252.

Gregoret, L. M. & Cohen, F. E. 1990. Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* 211:959-974.

Gregoret, L. M. & Cohen, F. E. 1991. Protein folding. Effect of packing density on chain conformation. *J. Mol. Biol.* 219:109-122.

Gribskov, M.; McLachlan, M. and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sc. U.S.A.* 84:4355-5358.

Griewank, A. O. 1981. Generalised descent for global optimisation. *J. Optimization Theory Appl.* 24:11-39.

Güntert, P.; Braun, W. & Wüthrich, K. 1991. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS, and GLOMSA. *J. Mol. Biol.* 217:517-530.

Guzzo, A. V. 1965. The influence of amino acid sequence on protein structure. *Biophys. J.* 5:809-822.

Hartl, F.-U.; Hlodan, R. and Langer, T. 1994. Molecular chaperones in protein folding: the art of avoiding sticky situations. *TIBS* 19:20-25.

Haussler, D.; Krogh, A.; Mian, I. S. and Sjölander, K. 1993. Protein Modeling using Hidden Markov Models: Analysis of Globins. In Proceedings for the 26th Hawaii International Conference on Systems Sciences, 792-802. Wailea, HI, U.S.A.: IEEE Computer Society Press.

Havel, T. & Wüthrich, K. 1984. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear megnetic resonance measurements of intramolecular $^1$H-$^1$H proximitieds in solution. *Bull. Math Biol.* 46:673-698.

Havel, T. F. & Wüthrich, K. 1984. A distance geometry program for determing the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular $^1$H$^1$H proximities in solution. *Bull. Math. Biol.* 46:673-698.

Havel, T. F. 1991. An Evaluation of Computational Strategies for Use in the Determination of Protein Structure from Distance Constraints Obtained by Nuclear Magnetic Resonance. *Prog. Biophys. Mol. Biol.* 56:43-78.

Havel, T. F.; Kuntz, I. D. & Crippen, G. M. 1983. The theory and practice of distance geometry. *Bull. Math. Biol.* 45:665-720.

Havel, T.; Kuntz, I. D. and Crippen, G. M. 1983. The Combinatorial Distance Geometry Method for the Calculation of Molecular Conformation I. A New Approach to an Old Problem. *J. Theor. Biol.* 104:359-381.

Hayward, S. & Collins, J. F. 1992. Limits on α-Helix Prediction With Neural Network Models. *Proteins* 14:372-381.

Henderson, R.; Baldwin, J. M.; Ceska, T. A.; Zemlin, F.; Beckmann, E. and Downing, K. H. 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* 213:899-929.

Hendlich, M.; Lackner, P.; Weitckus, S.; Flöckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G. and Sippl, M. J. 1990. Identification of Native Protein Folds Amongst a Large Number of Incorrect Models. The Calculation of Low Energy Conformations from Potentials of Mean Force. *J. Mol. Biol.* 216:167-180.

Henikoff, S. & Henikoff, J. G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* 17:49-61.

Henikoff, S. & Henikoff, J. G. 1994. Position-based sequence weights. *J. Mol. Biol.* 243:574-578.

Henikoff, S. 1991. Playing with blocks: some pitfalls of forcing multiple alignments. *The New Biologist* 3:1148-1154.

Heringa, J. & Argos, P. 1993. A method to recognise distant repeats in protein sequences. *Proteins* 17:391-411.

Higgins, D. G. & Sharp, P. M. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73:237-244.

Higgins, D. G. & Sharp, P. M. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* 5:151-153.

Higgins, D. G.; Bleasby, A. J. and Fuchs, R. 1992. CLUSTAL V: improved sofware for multiple sequence alignment. *CABIOS* 8:189-191.

Hirst, J. D. & Sternberg, M. J. E. 1991. Prediction of ATP-binding motifs a comparison of a perceptron-type neural network and a consensus sequence method. *Prot. Engin.* 4:615-623.

Hobohm, U. & Sander, C. 1994. Enlarged representative set of protein structures. *Prot. Sci.* 3:522-524.

Hobohm, U.; Scharf, M.; Schneider, R. and Sander, C. 1992. Selection of representative protein data sets. *Prot. Sci.* 1:409-17.

Hol, W. G. J.; Halie, L. M. and Sander, C. 1981. Dipoles of the α-helix and β-sheet: their role in protein folding. *Nature* 294:532-536.

Holbrook, S. R.; Muskal, S. M. and Kim, S.-H. 1990. Predicting surface exposure of amino acids from protein sequence. *Prot. Engin.* 3:659-665.

Holley, H. L. & Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sc. U.S.A.* 86:152-156.

Holm, L. & Sander, C. 1992a. Evaluation of Protein Models by Atomic Solvation Preference. *J. Mol. Biol.* 225:93-105.

Holm, L. & Sander, C. 1992b. Fast and Simple Monte Carlo Algorithm for Side Chain Optimization in Proteins: Application to Model Building by Homology. *Proteins* 14:213-223.

Holm, L. & Sander, C. 1993. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* 233:123-138.

Holm, L. & Sander, C. 1994a. The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* 22:3600-3609.

Holm, L. & Sander, C. 1994b. Parser for protein folding units. *Proteins* 19:256-268.

Holm, L. & Sander, C. 1994c. Searching Protein Structure Databases Has Come of Age. *Proteins* 19:165-173.

Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G. and Vriend, G. 1993. A database of protein structure families with common folding motifs. *Prot. Sci.* 1:1691-1698.

Holm, L.; Rost, B.; Sander, C.; Schneider, R. and Vriend, G. 1994. Data based modeling of proteins. In Statistical Mechanics, Protein Structure, and Protein Substrate Interactions, 277-296. New York: Plenum Press.

Hoppe, W. 1957. Die Faltmolekülmethode - eine neue Methode zur Bestimmung der Kristallstruktur bei ganz oder teilweise bekannter Molekülstruktur. *Acta Crystallogr.* 10:750-751.

Hubbard, T. J. P. & Blundell, T. L. 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Prot. Engin.* 1:159-171.

Hubbard, T. J. P. & Park, J. 1995. Fold recognition and ab initio sturcture predictions using Hidden Markov models and β-strand pair potentials. *Proteins* in press.

Hubbard, T. J. P. & Sander, C. 1991. The role of heat-shock and chaperone proteins in protein folding: possible molecular mechanisms. *Prot. Engin.* 4:711-717.

Hubbard, T. J. P. 1994. Use of β-strand Interaction Pseudo-Potential in Protein Structure Prediction and Modelling. In 27th Hawaii International Conference on System Sciences, 336-344. Maui, Hawaii, USA: IEEE Society Press.

Hutchinson, E. G. & Thornton, J. M. 1993. The Greek key motif: extraction, classification and analysis. *Prot. Engin.* 6:233-245.

Islam, S. A. & Sternberg, M. J. E. 1989. A relational database of protein structures designed for flexible enquiries about conformation. *Protein Eng.* 2:431-442.

IUPAC-IUB 1970. Commission on Beochemical Nomenclature 1969. *Biochem.* 9:3471-3479.

Jackson, R. M. & Sternberg, M. J. E. 1993. Protein surface area defined. *Nature* 366:638.

Jaennicke, R. 1987. Folding and Association of Proteins\. *Prog. Biophys. molec. Biol.*, 49:117-237.

Janin, J. 1976. Surface Area of Globular Proteins. *J. Mol. Biol.* 105:13-14.

Jaynes, E. T. 1978. Where do we stand on maximum entropy? In *Papers on probability, statistics, and statistical physics* R. D. Rosenkrantz eds. Reidel: Dordrecht, Holland.

Johnson, M. S. 1991. Comparisons of protein structures. *Curr. Opin. Str. Biol.* 1:334-344.

Johnson, M. S.; Overington, J. P. and Blundell, T. L. 1993. Alignment and Searching for Common Protein Folds Using a Data Bank of Structural Templates. *J. Mol. Biol.* 231:735-752.

Johnston, M., et al. 1994. Complete nucleotide sequence of saccaromyces cerevisiae chromosome VIII. *Science* 265:2077-2082.

Jones, D. T.; Taylor, W. R. and Thornton, J. M. 1992a. A new approach to protein fold recognition. *Nature* 358:86-89.

Jones, D. T.; Taylor, W. R. and Thornton, J. M. 1992b. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8:275-282.

Jones, D. T.; Taylor, W. R. and Thornton, J. M. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochem.* 33:3038-3049.

Jones, T. A. & Thirup, S. 1986. Using known substructures in protein model building and crystallography. *EMBO J.* 5:819-822.

Kabsch, W. & Sander, C. 1983a. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 22:2577-2637.

Kabsch, W. & Sander, C. 1983b. How good are predictions of protein secondary structure? *FEBS Lett.* 155:179-182.

Kabsch, W. & Sander, C. 1983c. Segment83. *unpublished* .

Kabsch, W. & Sander, C. 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sc. U.S.A.* 81:1075-1078.

Karlin, S. & Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sc. U.S.A.* 87:2264-2268.

Karlin, S.; Dembo, A. and Kawabata, T. 1990. Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* 18:571-581.

Karplus, M. & Petsko, G. A. 1990. Molecular dynamics simulations in biology. *Nature* 347:631-639.

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

Keepers, J. W. & James, T. L. 1984. A theoretical study of distance determinations from NMR. Two-dimensional nuclear Overhauser effect spectra. *J. Magn. Reson.* 24:343-366.

Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. J.; Davies, D. R. and Phillips, D. C. 1960. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 185:422-427.

King, R. D. & Sternberg, M. J. 1990. Machine Learning Approach for the Prediction of Protein Secondary Structure. *J. Mol. Biol.* 216:441-457.

Kneller, D. G.; Cohen, F. E. and Langridge, R. 1990. Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.* 214:171-182.

Koehl, P. & Delarue, M. 1995. A self consistent mean field approach to simulataneous gap closure and side-chain positioning in homology modelling. *Nature Struct. Biol.* 2:163-170.

Kreusch, A. & Schulz, G. E. 1994. Refined structure of the porin from rhodpseudomonas blastica. *J. Mol. Biol.* 243:891-905.

Krogh, A.; Brown, M.; Mian, I. S.; Sjölander, K. and Haussler, D. 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *J. Mol. Biol.* 235:1501-1531.

Kühlbrandt, W.; Wang, D. N. and Fujiyoshi, Y. 1994. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* 367:614-621.

Kuntz, I. D. 1972. Protein folding. *J. Am. Chem. Soc.* 49:4009-4012.

Kuntz, I. D.; Thomason, J. F. & Oshiro, C. M. 1989. Distance geometry. In *Methods in Enzymology* 177, N. J. Oppenheimer & T. J. James eds. San Diego, USA: Academic Press.

Kyte, J. & Doolittle, R. F. 1982. A Simple Method for Displaying the Hydrophathic Character of a Protein. *J. Mol. Biol.* 157:105-132.

Lamzin, V. S. & Wilson, K. S. 1993. Automated refinement of protein models. *Acta Crystallogr. D* 49:129-147.

Landis, C. & Allured, V. 1991. *J. Am. Chem. Soc.* 113:9493.

Laskowski, R. A.; Moss, D. S. and Thornton, J. M. 1993. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* 231:1049-1067.

Lathrop, R. H. & Smith, T. F. 1994. A Branch-and-Bound Algorithm for Optimal Protein Threading with Pairwise (Contact Potential) Amino Acid Interactoins. In 27th Hawaii International Conference on System Sciences, 365-374. Wailea, HI, U.S.A.: IEEE Computer Society Press.

Lattman, E. E. & Rose, G. D. 1993. Protein folding-what's the question? *Proc. Natl. Acad. Sc. U.S.A.* 90:439-441.

Lattman, E. E. 1994. Protein crystallography for all. *Proteins* 18:103-106.

Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F. and Wootton, J. C. 1993. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 262:208-214.

Lawrence, C.; Auger, I. and Mannella, C. 1987. Distribution of Accessible Surfaces of Amino Acids in Globular Proteins. *Proteins* 2:153-161.

Lee, B. K. & Richards, F. M. 1971. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* 55:379-400.

Lesk, A. M. & Boswell, R. D. 1992. Homology modelling: inferences from tables of aligned sequences. *Curr. Opin. Str. Biol.* 2:242-247.

Lesk, A. M. 1991. *Protein Architecture - A Practical Approach* . Oxford, New York, Tokyo: Oxford University Press.

Levin, J. M.; Pascarella, S.; Argos, P. and Garnier, J. 1993. Quantification of Secondary Structure Prediction Improvement Using Multiple Alignments. *Prot. Engin.* 6:849-854.

Levin, J. M.; Robson, B. and Garnier, J. 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* 205:303-308.

Levitt, M. 1992. Accurate Modeling of Protein Conformation by Automatic Segment Matching. *J. Mol. Biol.* 226:507-533.

Li, J.; Platt, E.; Waskowycz, B.; Cotterill, R. M. J. & Robson, B. 1992. *Biophys. Chem.* 43:221.

Lifson, S. & Sander, C. 1979. Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature* 282:109-111.

Lim, V. I. 1974. Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure. *J. Mol. Biol.* 88:857-872.

Lipman, D. J. & Pearson, W. R. 1985. Rapid and sensitive protein similarity searches. *Science* 227:1435-1441.

Livingstone, C. D. & Barton, G. J. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *CABIOS* 9:745-756.

Livingstone, C. D. & Barton, G. J. 1994. Secondary Structure Prediction from Multiple Sequence Data: Blood Clotting Factor XIII and Yersinia Protein-Tyrosine Phosphatase. *Int. J. Peptide Protein Res.* submitted.

Low, B. W.; Lovell, F. M. and Rudko, A. D. 1968. Prediction of a-helical regions in proteins of known sequence. *Proc. Natl. Acad. Sc. U.S.A.* 60:1519-1526.

Lüthy, R.; Bowie, J. U. and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356:83-85.

Lüthy, R.; McLachlan, A. D. and Eisenberg, D. 1991. Secondary Structure-Based Profiles: Use of Structure-Conserving Scoring Tables in Searching Protein Sequence Databases for Structural Similarities. *Proteins* 10:229-239.

Maclin, R. & Shavlik, J. W. 1993. Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou-Fasman Algorithm for Protein Folding. *Machine Learning* 11:195-215.

Macura, S. & Ernst, R. R. 1980. Elucidation of cross-relaxation in liquids by two-dimensional N.M.R. spectroscopy. *Mol. Phys.* 41:95-117.

Maiorov, V. N. & Crippen, G. M. 1992. A contact potential that recognises correct folding of globular proteins. *J. Mol. Biol.* 227:876-888.

Maiorov, V. N. & Crippen, G. M. 1994. Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins. *J. Mol. Biol.* 235:625-634.

Manoil, C. & Beckwith, J. 1986. A genetic approach to analyzing membrane protein topology. *Science* 233:1403-1408.

Martin, J. & Hartl, F. U. 1993. Protein folding in the cell: molecular chaperones pave the way. *Structure* 1:161-164.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Ac.* 405:442-451.

Maxfield, F. R. & Scheraga, H. A. 1976. Status of Empirical Methods for the Prediction of Protein Backbone Topography. *Biochem.* 15:5138-5153.

Maxfield, F. R. & Scheraga, H. A. 1979. Improvements in the Prediction of Protein Topography by Reduction of Statistical Errors. *Biochem.* 18:697-704.

May, A. C. W. & Blundell, T. L. 1994. Automated comparative modelling of protein structures. *Curr. Opin. Biotech.* 5:355-360.

May, A. C. W. & Johnson, M. S. 1994. Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Prot. Engin.* 7:475-485.

McGregor, M. J.; Flores, T. P. and Sternberg, M. J. E. 1989. Prediction of beta-turns in proteins using neural networks. *Prot. Engin.* 2:521-526.

McLachlan, A. D. 1971. Tests for comparing related amino acid sequences. *J. Mol. Biol.* 61:409-424.

Melzler, W. J.; Hare, D. R. & Pardi, A. 1989. Limited sampling of conformational space by the distance geometry algorithm: implications for structures generated from NMR data. *Biochemistry* 28:7045-7052.

Mertz, J. E.; Güntert, P.; Wüthrich, K. & Braun, W. 1991. Complete relaxation matrix refinement of NMR structures of proteins using analytically calculated dihedral angle derivatives of NOE intensities. *J. Biomol. NMR* 1:257-269.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. & Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087-1092.

Miyazawa, S. & Jernigan, R. L. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534-552.

Miyazawa, S. & Jernigan, R. L. 1993. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Prot. Engin.* 6:267-278.

Momany, F. A.; McGuire, R. F.; Burgess, A. W. and Scheraga, H. A. 1975. Energy parameters in polypeptides. *J. Phys. Chem.* 79:2361-2381.

Monod, J. 1970. *Le hasard et la nécessité*. Paris: Seuil.

Muggleton, S.; King, R. D. and Sternberg, M. J. E. 1992. Protein secondary structure prediction using logic-based machine learning. *Prot. Engin.* 5:647-657.

Murzin, A. G. & Chothia, C. 1992. Protein architecture: new superfamilies. *Curr. Biol.* 2:895-903.

Murzin, A. G. 1994. New protein folds. *Curr. Opin. Str. Biol.* 4:441-449.

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

Muskal, S. M. & Kim, S.-H. 1992. Predicting Protein Secondary Structure Content. A Tandem Neural Network Approach. *J. Mol. Biol.* 225:713-727.

Muskal, S. M.; Holbrook, S. R. and Kim, S.-H. 1990. Prediction of the disulfide-bonding state of cysteine in proteins. *Prot. Engin.* 3:667-672.

Nagano, K. & Hasegawa, K. 1975. Logical Analysis of the Mechanism of Protein Folding. *J. Mol. Biol.* 94:257-281.

Nagano, K. 1973. Logical Analysis of the Mechanism of Protein Folding. *J. Mol. Biol.* 75:401-420.

Nagano, K. 1977. Triplet Information in Helix Prediction Applied to the Analysis of Super-secondary Structures. *J. Mol. Biol.* 109:251-274.

Nayal, M. & Di Cera, E. 1994. Predicting $Ca^{2+}$-binding sites in proteins. *Proc. Natl. Acad. Sc. U.S.A.* 91:817-821.

Needlman, S. B. & Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-53.

Neher, E. 1994. How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sc. U.S.A.* 91:98-102.

Nilges, M. & Brünger, A. T. 1993. Successful Prediction of Coiled Coil Geometry of the GCN4 Leucine Zipper Domain by Simulated Annealing: Comparison to the X-Ray Structure. *Proteins* 15:133-146.

Nilges, M. 1993. A Calculation Strategy for the Structure Determination of Symmetric Dimers by 1H NMR. *Proteins* 17:297-309.

Nilges, M. 1993. A calculation stategy for the structure determination of symmetric dimers by $^{1}$H-NMR. *Proteins* 17:297-309.

Nilges, M. 1994. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* 245:645-660.

Nilges, M.; Clore, G. M. & Gronenborn, A. M. 1988a. Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Lett.* 239:129-136.

Nilges, M.; Clore, G. M. & Gronenborn, A. M. 1988b. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett.* 229:317-324.

Nilges, M.; Gronenborn, A. M.; Brünger, A. T. & Clore, G. M. 1988c. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato caroxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Engng* 2:27-38.

Nilges, M.; Habazettl, J.; Brünger, A. T. & Holak, T. A. 1991a. Relaxation matrix refinement of the solution structure of squash trypsin inhibitor. *J. Mol. Biol.* 219:499-510.

Nilges, M.; Kuszewski, J. & Brünger, A. T. 1991b. Sampling properties of simulated annealing and distance geometry. In *Computational aspects of the study of biological macromolecules by nuclear magnetic resonance* J. C. Hoch; F. M. Poulsen & C. Redfield eds. New York: Plenum Press.

Nishikawa, K. & Matsuo, Y. 1993. Development of pseudo-energy potentials for assessing protein 3-D-1D compatibility and detecting weak homologies. *Prot. Engin.* 6:811-820.

Nishikawa, K. & Ooi, T. 1982. Correlation of the amino acid composition of a protein to its structural and biological characteristics. *J. Biochem.* 91:1821-1824.

Novotny, J.; Bruccoleri, R. E. and Karplus, M. 1984. An analysis of incorrectly folded models. Implications for structure prediction. *J. Mol. Biol.* 177:787-818.

Novotny, J.; Rashin, A. A. and Bruccoleri, R. E. 1988. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* 4:19-30.

O'Donoghue, S. I.; Junius, F. K. & King, G. F. 1993. Structure determination of symmetric coiled coils by

NMR; applications to the leucine zipper proteins jun and GCN4. *Protein Engng* 6:557-564.

Oliver, S., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357:38-46.

Orengo, C. A.; Flores, T. P.; Jones, D. T.; Taylor, W. R. and Thornton, J. M. 1993. Recurring structural motifs in proteins with different functions. *Curr. Biol.* 3:131-139.

Ouzounis, C.; Sander, C.; Scharf, M. and Schneider, R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.* 232:805-825.

Overington, J. P. 1992. Comparison of three-dimensional structures of homologous proteins. *Curr. Opin. Str. Biol.* 2:394-401.

Overington, J.; Donnelly, D.; Johnson, M. S.; Sali, A. and Blundell, T. L. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Prot. Sci.* 1:216-226.

Overington, J.; Johnson, M. S.; Sali, A. and Blundell, T. L. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Royal Soc. Lond.* B 241:132-145.

Pain, R. H. & Robson, B. 1970. Analysis of the Code Relating Sequence to Secondary Structure in Proteins. *Nature* 227:62-63.

Pancoska, P.; Blazek, M. and Keiderling, T. A. 1992. Relationships between Secondary Structure Fractions for Globular Proteins. Neural Network Analyses of Crystallographic Data Sets. *Biochem.* 31:10250-10257.

Park, K.; Perczel, A. and Fasman, G. D. 1992. Differentiation between transmembrane helices and peripheral helices by the deconvolution of circular dichroism spectra of membrane proteins. *Prot. Sci.* 1:1032-1049.

Pastore, A. & Lesk, A. M. 1990. Comparison of the Structures of Globins and Phycocyanins: Evidence for Evolutionary Relationship. *Proteins* 8:133-55.

Pauling, L. & Corey, R. B. 1951. Configurations of Polypeptide Chains with Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc. Natl. Acad. Sc. U.S.A.* 37:729-740.

Pauling, L. & Corey, R. B. 1953a. Two Pleated-sheet Configurations of Polypeptide Chains Involving Both cis and trans Amide Groups. *Proceedings National Academy of Science* 39:247-252.

Pauling, L. & Corey, R. B. 1953b. Two Rippled-sheet Configurations of Polypeptide Chains, and a Note About the Pleated Sheets. *Proceedings National Academy of Science* 39:253-256.

Pauling, L.; Corey, R. B. and Branson, H. R. 1951. The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain. *Proc. Natl. Acad. Sc. U.S.A.* 37:205-234.

Pearson, W. R. & Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sc. U.S.A.* 85:2444-2448.

Periti, P. F.; Quagliarotti, G. and Liquori, A. M. 1967. Recognition of a-Helical Segments in Proteins of Known Primary Structure. *J. Mol. Biol.* 24:313-322.

Persson, B. & Argos, P. 1994. Prediction of Transmembrane Segments in Proteins Utilising Multiple Sequence Alignments. *J. Mol. Biol.* 237:182-192.

Perutz, M. F. 1940. "Unboiling" an egg. *Discovery* May:reprint in Jaenicke, Rainer: Protein Folding. Amsterdam, New York: Elsevier, 1980, p. 14.

Perutz, M. F. 1980. In Jaenicke, R. eds. *Protein Folding.* Amsterdam, New York: Elsevier.

Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, G.; Will, G. and North, A. T. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* 185:416-422.

Pickett, S. D. & Sternberg, M. J. E. 1993. Empirical Scale of Side-chain conformational Entropy in Protein Folding. *J. Mol. Biol.* 231:825-839.

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

Presnell, S. R. & Cohen, F. E. 1993. Artificial Neural Networks for Pattern Recognition in Biochemical Sequences. *Annu. Rev. Biophys. Biomol. Struct.* 22:283-298.

Presnell, S. R.; Cohen, B. I. and Cohen, F. E. 1992. A Segment-Based Approach to Protein Secondary Structure Prediction. *Biochem.* 31:983-993.

Prothero, J. W. 1966. Correlation between the distribution of amino acids and alpha helices. *Biophys. J.* 6:367-370.

Ptitsyn, O. B. & Finkelstein, A. V. 1983. Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* 22:15-25.

Ptitsyn, O. B. 1969. Statistical Analysis of the Distribution of Amino Acid Residues among Helical and Non-helical Regions in Globular Proteins. *J. Mol. Biol.* 42:501-510.

Ptitsyn, O. B. 1992. Secondary structure formation and stability. *Curr. Opin. Str. Biol.* 2:13-20.

Qian, N. & Sejnowski, T. J. 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.* 202:865-884.

Rao, S.; Zhu, Q.-L.; Vajda, S. and Smith, T. 1993. The local information content of the protein structural database. *FEBS Lett.* 322:143-146.

Read, R. J. & Moult, J. 1992. Fitting electron density by systematic search. *Acta Crystallogr. A* 48:104-113.

Rees, A. R.; Sternberg, M. J. E. and Wetzel, R. 1992. *Protein Engineering* . Oxford: IRL Press.

Richardson, J. 1981. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* 34:168-339.

Richardson, J. S. & Richardson, D. C. 1989. Principles and patterns of protein conformation. In Fasman, G. D. eds. *Prediction of protein structure and the principles of protein conformation.* New York, London: Plenum Press.

Richardson, J. S. 1985. Describing patterns of protein tertiary structure. *Meth. Enzymol.* 115:349-358.

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

Robson, B. & Pain, R. H. 1971. Analysis of the Code Relating Sequence to Conformation in Proteins: Possible Implications for the Mechanism of Formation of Helical Regions. *J. Mol. Biol.* 58:237-259.

Robson, B. & Pain, R. H. 1974a. Analysis of the Code Relating Sequence to Conformation in Globular Proteins - An Informational Analysis of the Residue in Determining the Conformation of Its Neighbours in the Primary Sequence. *Biochem. J.* 141:883-897.

Robson, B. & Pain, R. H. 1974b. Analysis of the Code Relating Sequence to Conformation in Globular Proteins - Development of a Stereochemical alphabet on the Basis of Intra-Residue Information. *Biochem. J.* 141:869-882.

Robson, B. & Pain, R. H. 1974c. Analysis of the Code Relating Sequence to Conformation in Globular Proteins - The Distribution of Residue Pairs in Turns and Kinks in the Backbone Chain. *Biochem. J.* 141:899-904.

Robson, B. 1974. Analysis of the Code Relating Sequence to Conformation in Globular Proteins - Theory and Application of expected Information. *Biochem. J.* 141:853-867.

Robson, B. 1976. Conformational Properties of Amino Acid Residues in Globular Proteins. *J. Mol. Biol.* 107:327-56.

Rooman, M. & Wodak, S. J. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature* 335:45-49.

Rooman, M. J. & Wodak, S. 1991. Weak Correlation Between Predictive Power of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins. *Proteins* 9:69-78.

Rooman, M. J.; Kocher, J. P. and Wodak, S. J. 1991. Prediction of Protein Backbone Conformation Based on Seven Structure Assignments: Influence of Local Interactions. *J. Mol. Biol.* 221:961-979.

Rossmann, M. G. & Blow, D. M. 1962. The detection of sub-units within the crystallographic assymetric unit. *Acta Crystallogr. A* 15:24-31.

Rost, B. & Sander, C. 1992a. Exercising Multi-layered Networks on Protein Secondary Structure. In Neural Networks: From Biology to High Energy Physics, 209-220. Elba, Italy: International Journal of Neural Systems.

Rost, B. & Sander, C. 1992b. Jury returns on structure prediction. *Nature* 360:540.

Rost, B. & Sander, C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sc. U.S.A.* 90:7558-7562.

Rost, B. & Sander, C. 1993b. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584-599.

Rost, B. & Sander, C. 1993c. Secondary structure prediction of all-helical proteins in two states. *Prot. Engin.* 6:831-836.

Rost, B. & Sander, C. 1994a. 1D secondary structure prediction through evolutionary profiles. In Bohr, H. and Brunak, S. eds: *Protein Structure by Distance Analysis.* Amsterdam, Oxford, Washington: IOS Press.

Rost, B. & Sander, C. 1994b. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55-72.

Rost, B. & Sander, C. 1994c. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20:216-226.

Rost, B. & Sander, C. 1994d. Protein Structure Prediction by Neural Networks. In Arbib, M. eds. *The Handbook of Brain Theory and Neural Networks.* Boston: Bradford Books/The MIT Press.

Rost, B. & Sander, C. 1994e. Structure prediction of proteins - where are we now? *Curr. Opin. Biotech.* 5:372-380.

Rost, B. & Sander, C. 1995. Progress of 1D protein structure prediction at last. *Proteins* x.x:submitted March, 1995.

Rost, B. & Vriend, G. 1993. Neural Networks in Chemistry. *CDA News* 8:24-27.

Rost, B. 1993. Neural networks and evolution - advanced prediction of protein secondary structure. Ph.D. diss., Dep. of Physics and Astronomy, University of Heidelberg, F.R.G.

Rost, B. 1995a. Fitting 1D predictions into 3D structures. In Bohr, H. and Brunak, S. eds. *Protein structure by distance analysis.* CRC Press.

Rost, B. 1995b. PHD: predicting 1D protein structure by profile based neural networks. *Meth. Enzymol.* to submit June 1995.

Rost, B. 1995c. TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures. In The Third International Conference on Intelligent Systems for Molecular Biology, submitted. Cambridge, U. K., July 16-19, 1995: AAAI Press.

Rost, B.; Casadio, R.; Fariselli, P. and Sander, C. 1995. Prediction of helical transmembrane segments at 95% accuracy. *Prot. Sci.* 4:521-533.

Rost, B.; Sander, C. and Schneider, R. 1993. Progress in protein structure prediction? *TIBS* 18:120-123.

Rost, B.; Sander, C. and Schneider, R. 1994a. Evolution and Neural Networks - Protein secondary structure prediction above 71% accuracy. In 27th Hawaii International Conference on System Sciences, 385-394. Wailea, Hawaii, U.S.A.: IEEE Society Press.

Rost, B.; Sander, C. and Schneider, R. 1994b. PHD - an automatic server for protein secondary structure prediction. *CABIOS* 10:53-60.

Rost, B.; Sander, C. and Schneider, R. 1994c. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13-26.

Russell, R. B. & Barton, G. J. 1992. Multiple Protein Sequence Alignment From Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels. *Proteins* 14:309-323.

Russell, R. B. & Barton, G. J. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* 234:951-957.

Saibil, H. & Wood, S. 1993. Chaperonins. *Curr. Opin. Str. Biol.* 3:207-213.

Saitoh, S.; Nakai, T. and Nishikawa, K. 1993. A Geometrical Constraint Approach for Reproducing the Native Backbone Conformation of a Protein. *Proteins* 15:191-204.

Salamov, A. A. & Solovyev, V. V. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* 247:11-15.

Sali, A. & Blundell, T. 1994. Comparative Protein Modelling by Satisfaction of Spatial Restraints. In Bohr, H. and Brunak, S. eds. *Protein Structure by Distance Analysis.* Amsterdam, Oxford, Washington: IOS Press.

Sali, A.; Overington, J. P.; Johnson, M. S. and Blundell, T. L. 1990. From comparisons of protein sequences and structures to protein modelling and design. *TIBS* 15:235-240.

Salzberg, S. & Cost, S. 1992. Predicting Protein Secondary Structure with a Nearest-neighbor Algorithm. *J. Mol. Biol.* 227:371-374.

Sander, C. & Schneider, R. 1991. Database of homology-derived structures and the structurally meaning of sequence alignment. *Proteins* 9:56-68.

Sander, C. & Schneider, R. 1994. The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* 22:3597-3599.

Sander, C.; Scharf, M. and Schneider, R. 1992. Design of protein structures. In Rees, A. R., Sternberg, M. J. E. and Wetzel, R. eds. *Protein Engineering*. Oxford: IRL Press.

Sasagawa, F. & Tajima, K. 1993. Prediction of protein secondary structures by a neural network. *CABIOS* 9:147-152.

Scheraga, H. A. 1960. Structural studies of ribonuclease III. A model for the secondary and tertiary structure. *J. Am. Chem. Soc.* 82:3847-3852.

Schneider, R. 1989. Sekundärstrukturvorhersage von Proteinen unter Berücksichtigung von Tertiärstrukturaspekten. Ph.D. diss., Department of Biology, Univ. Heidelberg, FRG.

Schneider, R. 1994. Sequenz und Sequenz-Strukture Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen. Ph.D. diss., Univ. of Heidelberg.

Schulz, G. E. & Schirmer, R. H. 1979. *Principles of Protein Structure* . New York et al.: Springer.

Schulz, G. E. 1988. A Critical Evaluation of Methods for Prediction of Protein Secondary Structures. *Annu. Rev. Biophys. Biophys. Chem.* 17:1-21.

Sheek, R. M.; Torda, A. E.; Kemmink, J. & van Gunsteren, W. F. 1991. Structure determination by NMR: the modelling of NMR parameters as ensemble averages. In *Computational aspects of the study of biological macromolecules by nuclear magnetic resonance spectroscopy* J. C. Hoch; R. M. Pulsen & C. Redfield eds. New York: Plenum Press.

Sheridan, R. P.; Dixon, J. S. and Venkataraghavan, R. 1985. Generating plausible protein folds by secondary structure similarity. *Int. J. Peptide Protein Res.* 25:132-143.

Shindyalov, I. N.; Kolchanov, N. A. and Sander, C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Prot. Engin.* 7:349-358.

Shortle, D. 1995. Protein fold recognition. *Nature Struct. Biol.* 2:91-92.

Sipos, L. & von Heijne, G. 1993. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* 213:1333-1340.

Sippl, M. J. & Jaritz, M. 1994. Predictive Power of Mean Force Pair Potentials, In Bohr, H. and Brunak, S. eds. *Protein Structure by Distance Analysis*. Amsterdam, Oxford, Washington DC: IOS Press.

Sippl, M. J. & Lackner, P. 1992. Mean Field Energy Analysis of Experimentally Determined Protein Folds, Center for Applied Molecular Engineering, Inst. f. Chemistry and Biochemistry, Univ. Salzburg, Jakob Haringer Str. 1, A-5020 Salzburg, Austria.

Sippl, M. J. & Weitckus, S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* 13:258-271.

Sippl, M. J. 1982. On the Problem of Comparing Protein Structures. *J. Mol. Biol.* 156:359-388.

Sippl, M. J. 1990. The Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-based Prediction of Local Structures of Globular Proteins. *J. Mol. Biol.* 213:859-883.

Sippl, M. J. 1993a. Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Design* 7:473-501.

Sippl, M. J. 1993b. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355-362.

Sippl, M. J.; Hendlich, M. and Lackner, P. 1992. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin $\beta_4$. *Prot. Sci.* 1:625-640.

Sippl, M. J.; Jaritz, M.; Hendlich, M.; Ortner, M. and Lackner, P. 1994. Applications of Knowledge Based Mean Fields in the Determination of Protein Structures. In Doniach, S. eds. *Statistical Mechanics, protein structure and protein-substrate interactions*. New York, London: Plenum Press.

Sippl, M. J.; Weitckus, S. and Flöckner, H. 1994. In search of protein folds. In Merz, K. H. and LeGrand, S. eds. *The Protein Folding Problem and Tertiary Structure prediction*. Boston, MA, U.S.A.: Birkhäuser Boston Inc.

Smith, H. O.; Annau, T. M. and Chandrasegaran, S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proc. Natl. Acad. Sc. U.S.A.* 87:826-830.

Smith, R. F. & Smith, T. F. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sc. U.S.A.* 87:118-122.

Smith, T. F. & Waterman, M. S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Solovyev, V. V. & Salamov, A. A. 1994. Predicting α-helix and β-strand segments of globular proteins. *CABIOS* 10:661-669.

Sternberg, M. J. E. 1992. Secondary structure prediction. *Curr. Opin. Str. Biol.* 2:237-241.

Stigter, D.; Alonso, D. O. V. and Dill, K. A. 1991. Protein stability: Electrostatics and compact denatured states. *Proceedings National Academy of Science* 88:4176-4180.

Stolorz, P.; Lapedes, A. and Xia, Y. 1992. Predicting Protein Secondary Structure Using Neural Net and Statistical Methods. *J. Mol. Biol.* 225:363-377.

Stultz, C. M.; White, J. V. and Smith, T. F. 1993. Structural analysis based on state-space modeling. *Prot. Sci.* 2:305-314.

Subbarao, N. & Haneef, I. 1991. Defining topological equivalences in macromolecules. *Prot. Engin.* 4:877-884.

Summers, N. L. & Karplus, M. 1989. Construction of Side-chains in Homology Modelling. *J. Mol. Biol.* 210:785-811.

Summers, N. L. & Karplus, M. 1990. Modeling of Globular Proteins. *J. Mol. Biol.* 216:991-1016.

Suzuki, E. & Robson, B. 1977. Relationship between helix-coil transition parameters for synthetic polypeptides and helix conformation parameters for globular proteins. A simple model. *J. Mol. Biol.* 899-904.

Sweet, R. M. & Eisenberg, D. 1983. Correlation of Sequence Hydrophobicities Measures Similarity in

Three-dimensional Protein Structure. *J. Mol. Biol.* 171:479-488.

Szent-Györgyi, A. G. & Cohen, C. 1957. Role of Proline in Polypeptide Chain Configuration of Proteins. *Science* 126:697.

Tanaka, S. & Scheraga, H. A. 1975. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc. Natl. Acad. Sc. U.S.A.* 72:3802-3806.

Taylor, W. 1992. New paths from dead ends. *Nature* 356:478-480.

Taylor, W. R. & Hatrick, K. 1994. Compensating changes in protein multiple sequence alignments. *Prot. Engin.* 7:341-348.

Taylor, W. R. & Orengo, C. A. 1989. Protein Structure Alignment. *J. Mol. Biol.* 208:1-22.

Taylor, W. R. & Thornton, J. M. 1983. Prediction of super-secondary structure in proteins. *Nature* 301:540-542.

Taylor, W. R. & Thornton, J. M. 1984. Recognition of Super-secondary Structure in Proteins. *J. Mol. Biol.* 173:487-514.

Taylor, W. R. 1984. An Algorithm to Compare Secondary Structure Predictions. *J. Mol. Biol.* 173:512-521.

Taylor, W. R. 1988. Pattern matching methods in protein sequence comparison and structure prediction. *Prot. Engin.* 2:77-86.

Taylor, W. R. 1991. Towards Protein Tertiary Fold Prediction Using Distance and Motif Constraints. *Prot. Engin.* 4:853-870.

Taylor, W. R. 1993. Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. *Prot. Engin.* 6:593-604.

Taylor, W. R.; Jones, D. T. and Green, N. M. 1994. A Method for α-Helical Integral Membrane Protein Fold Prediction. *Proteins* 18:281-294.

Tchoumatchenko, I.; Vissotsky, F. and Ganascia, J.-G. 1993. How to Make Explicit A Neural Network Trained to Predict Proteins Secondary Structure, ACASA, LAFORIA-CNRS, Université Paris VI, 4 Place Jussieu, 75 252 Paris, CEDEX 05, France.

Thompson, J. D.; Higgins, D. G. and Gibson, T. J. 1994. Improved sensitivity of profile searches through the use of sequence weights and gab excision. *CABIOS* 10:19-29.

Thornton, J. M.; Flores, T. P.; Jones, D. T. and Swindells, M. B. 1992. Prediction of progress at last. *Nature* 354:105-106.

Topham, C. M.; McLeod, A.; Eisenmenger, F.; Overington, J. P.; Johnson, M. S. and Blundell, T. L. 1993. Fragment Ranking in Modelling of protein structure: conformationally-constrained environmental amino acid substitution Tables. *J. Mol. Biol.* 229:194-220.

Torda, A. E.; Scheek, R. M. & van Gunsteren, W. F. 1989. Time-dependent distances restraints in molecular dynamics simulations. *Chem. Phys. Lett.* 157:289-294.

Torda, A. E.; Scheek, R. M. & van Gunsteren, W. F. 1990. Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J. Mol. Biol.* 214:223-235.

Totrov, M. & Abagyan, R. 1994. Detailed ab initio prediction of lysozyme-antibody complex with 1.6Å accuracy. *Nature Struct. Biol.* 1:259-263.

van Gunsteren, W. F. & Mark, A. E. 1992. On the interpretation of biochemical data by molecular dynamcis computer simulations. *Eur. J. Biochem.* 204:947-961.

van Gunsteren, W. F. 1988. The role of computer simulation techniques in protein engineering. *Prot. Engin.* 2:5-13.

van Gunsteren, W. F. 1993. Molecular dynamics studies of proteins. *Curr. Opin. Str. Biol.* 3:167-174.

van Schaik, R. C.; van Gunsteren, W. F. & Berendsen, H. J. C. 1992. Conformational search by potential energy annealing: algorithm and application to cyclosporin A. *J. Comput.-Aided Mol. Design* 6:97-112.

Verlet, L. 1967. Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* 159:98-105.

Villar, H. O. & Kauvar, L. M. 1994. Amino acid preferences at protein binding sites. *FEBS Lett.* 349:125-130.

Vingron, M. & Argos, P. 1989. A fast and sensitive multiple sequence alignment algorithm. *CABIOS* 5:115-121.

Vingron, M. & Argos, P. 1991. Motif Recognition and Alignment for Many Sequences by Comparison of Dot-matrices. *J. Mol. Biol.* 218:33-43.

Vingron, M. & Waterman, M. S. 1994. Sequence alignment and penalty choice. *J. Mol. Biol.* 235:1-12.

Viswanadhan, V. N.; Denckla, B. and Weinstein, J. N. 1991. New Joint Prediction Algorithm (Q7-JASEP) Improves the Prediction of Protein Secondary Structure. *Biochem.* 30:11164-11172.

von Heijne, G. & Gavel, Y. 1988. Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* 174:671-678.

von Heijne, G. & Manoil, C. 1990. Membrane proteins - from sequence to structure. *Prot. Engin.* 4:109-112.

von Heijne, G. 1981. Membrane proteins-the amino acid composition of membrane-penetrating segments. *Eur. J. Biochem.* 120:275-278.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.* 14:4683-4690.

von Heijne, G. 1991. Computer analysis of DNA and protein sequences. *Eur. J. Biochem.* 199:253-256.

von Heijne, G. 1992. Membrane Protein Structure Prediction. *J. Mol. Biol.* 225:487-494.

Vriend, G. & Eijsink, V. 1993. Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases. *J. Comput. Aided Mol. Design* 7:367-396.

Vriend, G. & Sander, C. 1991. Detection of Common Three-Dimensional Substructures in Proteins. *Proteins* 11:52-58.

Vriend, G. & Sander, C. 1993. Quality of Protein Models: Directional Atomic Contact Analysis. *J. Appl. Cryst.* 26:47-60.

Vriend, G. 1990. Parameter relation rows: a query system for protein structure function relationships. *Prot. Engin.* 4:221-223.

Vriend, G.; Sander, C. and Stouten, P. F. W. 1994. A novel search method for protein sequence-structure relations using property profiles. *Prot. Engin.* 7:23-29.

Wagner, G. & Wüthrich, K. 1982. *J. Mol. Biol.* 155:347.

Wako, H. & Blundell, T. L. 1994a. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from

aligned sequences of homologous proteins I. Solvent accessibility classes. *J. Mol. Biol.* 238:682-692.

Wako, H. & Blundell, T. L. 1994b. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins II. Secondary structures. *J. Mol. Biol.* 238:693-708.

Wang, Z.-X. 1994. Assessing the accuracy of protein secondary structure. *Nature Struct. Biol.* 1:145-146.

Waterman, M. S. 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl. Acad. Sc. U.S.A.* 80:3123-3124.

Weiss, M. S. & Schulz, G. E. 1992. Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.* 227:493-509.

Wilbur, W. J. & Lipman, D. J. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sc. U.S.A.* 80:726-730.

Wilmanns, M. & Eisenberg, D. 1993. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. *Proc. Natl. Acad. Sc. U.S.A.* 90:1379-1383.

Wodak, S. & Janin, J. 1980. Analytical approximation to the accessible surface area of proteins. *Proc. Natl. Acad. Sc. U.S.A.* 77:1736-1740.

Wodak, S. J. & Janin, J. 1978. Computer Analysis of Protein-Protein Interaction. *J. Mol. Biol.* 124:323-342.

Wodak, S. J. & Rooman, M. J. 1993. Generating and testing protein folds. *Curr. Opin. Str. Biol.* 3:247-259.

Wu, C.; Whitson, G.; McLarty, J.; Ermongkonchai, A. and Chang, T.-C. 1992. Protein classification artificial neural system. *Prot. Sci.* 1:667-677.

Wüthrich, K. 1984. NMR of proteins and nucleic acids. New York: Wiley International.

Yang, J. X. & Havel, T. F. 1993. SESAME - A Least-Squares Approach to the Evaluation of Protein Structures Computed from NMR Data. *J. Biomol. NMR* 3:355-360.

Yi, T.-M. & Lander, E. S. 1993. Protein Secondary Structure Prediction Using Nearest-neigbor Methods. *J. Mol. Biol.* 232:1117-1129.

Yip, P. & Case, D. A. 1989. A new method for refinement of macromolecular structures based on nuclear overhauser effect spectroscopy. *J. Biomol. NMR* 83:643-648.

Zabin, H. B.; Horvath, M. P. and Terwilliger, T. C. 1991. Approaches to Predicting Effects of Single Amino Acid Substitutions on the Function of a Protein. *Biochem.* 30:6230-6240.

Zhang, C.-T. & Chou, K.-C. 1992. An optimization approach to predicting protein structural class from amino acid composition. *Prot. Sci.* 1:401-408.

Zhang, X.; Mesirov, J. P. and Waltz, D. L. 1992. Hybrid System for Protein Secondary Structure Prediction. *J. Mol. Biol.* 225:1049-63.

Zimm, B. H. & Bragg, J. K. 1959. Theory of the phase transition between the helix and random chain in ploypeptide chains. *J. Chem. Phys.* 31:526-535.

Zuckerkandl, E. & Pauling, L. 1965. Evolutionary Divergence and Convergence in Proteins. In Bryson, V. and Vogel, H. J. eds. *Evolving Genes And Proteins*. New York and London: Academic Press.

Zvelebil, M. J.; Barton, G. J.; Taylor, W. R. and Sternberg, M. J. E. 1987. Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* 195:957-961.