Computergestützte Strukturbiologie
(Strukturelle Bioinformatik)

Fold recognition

Sommersemester 2009

Peter Güntert

Sequence identity → Structural similarity

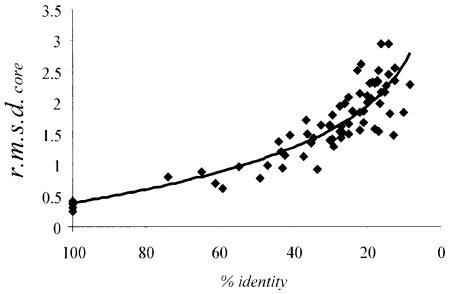


Figure 1.25 Relationships between sequence identity and structural similarity. The plot was obtained by using a larger set of proteins than in Figure 1.23, but the trend is essentially the same.

Structural similarity ↛ Sequence identity

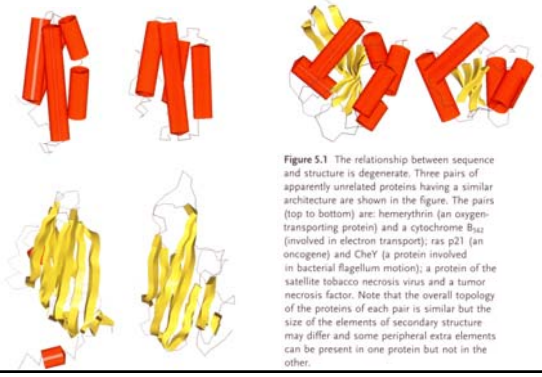


Figure 5.1 The relationship between sequence and structure is degenerate. Three pairs of apparently unrelated proteins having a similar architecture are shown in the figure. The pairs (top to bottom) are: hemerythrin (an oxygen-transporting protein) and a cytochrome B_{562} (involved in electron transport); ras p21 (an oncogene) and CheY (a protein involved in bacterial flagellum motion); a protein of the satellite tobacco necrosis virus and a tumor necrosis factor. Note that the overall topology of the proteins of each pair is similar but the size of the elements of secondary structure may differ and some peripheral extra elements can be present in one protein but not in the other.

Non-uniform distribution of folds

- **Few (~10) folds are shared by a large number (~30%) of known proteins**
- **Large diversity in sequences and functions among members of these "superfolds"**

Examples:

- **Immunoglobulin fold**
- **Rossman fold**
- **TIM barrel fold**
- **Globin fold**

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modeling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

Scheme of protein structure predicition

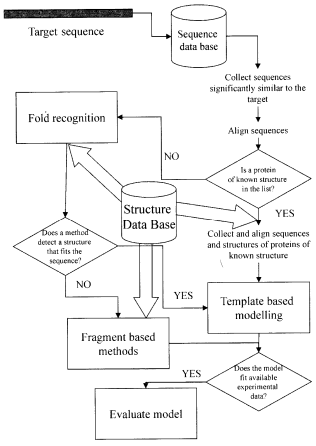


Figure 4.1 A guide to protein-structure prediction. The first step is always a search in the protein sequence database. Comparative modeling should be used when a protein of known structure sharing sequence similarity with the protein under examination is present in the database. If this is not so, fold-recognition methods should be applied and, should they fail, the user should resort to new fold or fragment-based methods. Note the central role played by the structure database in all these heuristic methods.

Inverse protein folding problem

Which amino acid sequences fold into a known three-dimensional structure?

Protein folding problem

Which three-dimensional structure is adopted by a given amino acid sequence?

Fold recognition methods

- **3D profile methods**
- **Threading**

Profile-based fold recognition

- **Physico-chemical properties of the amino acids of the target protein must "fit" with the environment in which they are placed in the modeled structure.**

Profile method for fold recognition

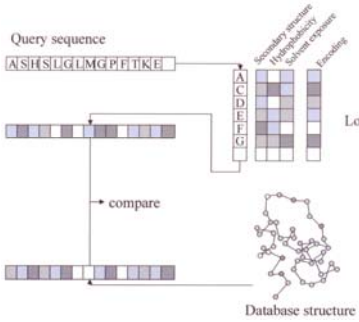


Figure 5.2 Schematic diagram of a possible profile-based method for fold recognition. The amino acids of the query sequence are replaced by a code that summarizes their hydrophobicity and their propensity for secondary structure type and solvent exposure. Each structure in the database is also encoded as a string by assigning a code to each of its amino acid positions. The code reflects their structural environment (secondary structure, solvent accessibility, and hydrophobicity of their environment). This does not depend on the actual amino acid present in the position analyzed. The string encoding the query sequence and each of the strings encoding the database structures are aligned and compared.

Sequence-structure alignment

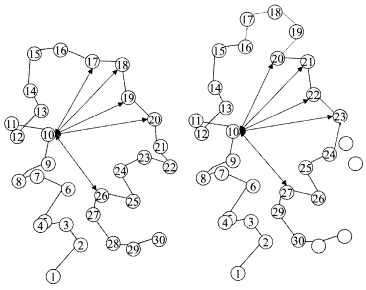


Figure 5.3 A query sequence can be positioned in a database structure in several ways, because there can be inserted and deleted residues, as shown in the right side of the figure. The interactions made by one amino acid, for example the one indicated with "10" in the figure, depend on the alignment of the rest of the sequence – the interactions of this amino acid (some of which are shown as arrows) are different in the two examples, reflecting two different alignments.

Frozen approximation

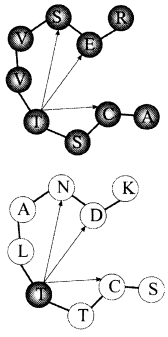


Figure 5.4 Schematic explanation of the frozen approximation. On the left, a database structure is shown with its original sequence (indicated by dark circles). In the right, the query sequence is positioned in the database structure in one of the many possible alignments. Calculation of the score should take into account which residues of the target sequence are in contact with, say, the threonine in the final alignment. In the frozen approximation, the interactions are computed by leaving the original sequence in every position of the database structure, except for the position occupied by the threonine. The procedure is repeated by substituting, in turn, each amino acid of the query sequence into a position of the target structure.

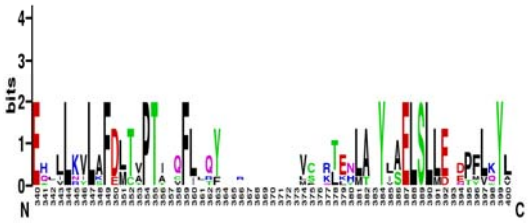
Position Specific Iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

PSI-BLAST is performed in five steps

1. Select a query sequence and search it against a protein sequence database: regular BLAST
2. PSI-BLAST constructs a multiple sequence alignment, then creates a "profile" or specialized position-specific scoring matrix (PSSM).
3. The PSSM is used as a query against the database.
4. PSI-BLAST estimates statistical significance (*E* values)
5. Repeat steps 2-4 iteratively, typically 5 times.
   At each new search, a new profile is used as the query.

Position-specific scoring matrix (PSSM)



PSI-BLAST: self-positives

- PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.
- The main source of false positives is the erroneous amplification of sequences not related to the query. For instance, a query with a coiled-coil motif may detect thousands of other proteins with this motif that are not homologous.
- Once even a single non-related protein is included in a PSI-BLAST search above threshold, it will not go away.
- One way to check results: take newly found sequences and perform PSI-BLAST using them, then examine whether we 'fish' original sequence (reciprocal identification)

A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure

James U. Bowie, Roland Lüthy, David Eisenberg

Science 253, 164-170 (1991)

3D profile method

Find sequences that are most compatible with the environments of residues in the 3D structure.

The environments are described by:

1. The area of the residue buried in the protein and inaccessible to solvent
2. The fraction of side-chain area that is covered by polar atoms (O and N)
3. The local secondary structure

Bowie, Lüthy & Eisenberg. *Science* 253, 164-170 (1991)

Sequences Structures Folds

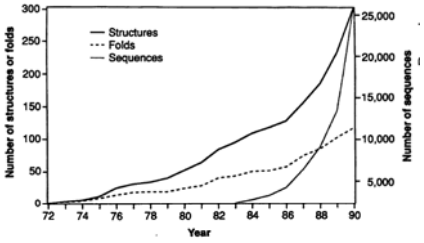


Fig. 1. The determination of amino acid sequences (right-hand scale) is outpacing the determination of 3D structures (left-hand scale) by a factor of 50. Also the number of structures is increasing faster than the number of folds: the cumulative number of structures deposited through 1990 is roughly twice the number of distinctly different protein folds. The number of sequences is the number deposited in the PIR database (*57*). The number of structures is the number of coordinate sets deposited in the Brookhaven Protein Data Bank (*58*), eliminating structures that differ only by a bound ligand, mutation, or space group. The number of folds is a subjective estimate of the number of "distinctly different structures," and should be regarded as having an uncertainty of at least ±20 in 1990.

Construction of 3D structure profile

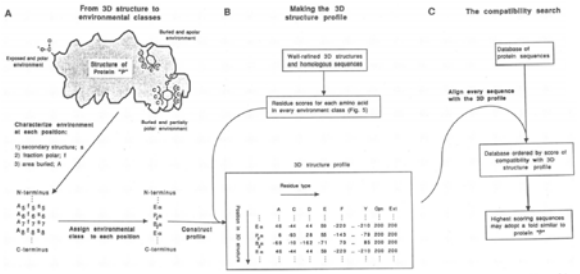


Fig. 2. Schematic description of the construction of a 3D structure profile (**A** and **B**) and of a 3D compatibility search of the sequence database (**C**). The 3D structure profile shown at the bottom of (B) is a portion of the profile for sperm whale myoglobin (Fig. 3), giving scores for only four positions of the structure (corresponding to residues 5, 6, 7, and 8) and for only 6 of the 20 amino acids.

Side-chain environment categories

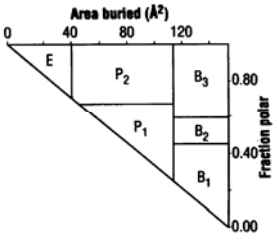


Fig. 4. The six side-chain environment categories. Two environmental characteristics were determined for each side chain: A, the total area buried in the protein structure; and f, the fraction of the side-chain area covered by polar atoms. If $A > 114$ Å^2, the residue was placed in environment class B_1 if $f < 0.45$, environment class B_2 if $0.45 \leq f < 0.58$, and environment class B_3 if $f \geq 0.58$. If $40 < A \leq 114$ Å^2, the residue was placed in environment category P_1 if $f < 0.67$ and environment class P_2 if $f \geq 0.67$. A residue was placed in the exposed environment category E if less than 40 Å^2 of the side chain was buried.

3D-1D scoring table

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	1.32	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	-1.57
$B_1\beta$	1.17	0.85	0.27	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.01	-1.67	-1.42	-1.80	-2.56	-1.81	-2.69	-1.16
B_1	1.08	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	0.83	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	-1.67	-1.41	-0.77	-1.14	-0.20
B_2	1.00	1.05	1.12	0.84	0.91	0.90	0.90	-0.86	-1.98	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\alpha$	0.80	-0.03	0.58	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	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	-0.37	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.33	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	0.68
$P_1\alpha$	-1.25	-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
$P_1\beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-0.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.31	-0.62	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.05	0.66	0.49	-0.60	-0.13	-0.61	0.28	-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	0.61	0.50
$P_2\beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.96	-1.29	-0.57	0.84	0.59	-0.06	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.56	-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	-0.20	-0.10	-1.58	-2.76	-1.10	-0.72	0.46	0.88	0.04	-0.44	-0.17	0.15	0.36	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	-0.19	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.87	0.12	1.13	0.20	-0.48	0.12	0.32	-0.03	0.41	0.05	0.22	-0.35	-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 Lüthy *et al.* (*28*). 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:

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

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.

Example of a 3D profile

Position in fold	Environment class	Amino acid type A	C	D	E	F	G	...	R	S	T	V	W	Y	Gap penalty Opn	Ext
1	E	12	-46	22	3	-190	113	...	-32	32	12	-91	-214	-94	2	0.02
2	B_2	-66	-5	-128	-135	105	-168	...	-80	-117	-76	60	102	112	2	0.02
3	$E\alpha$	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
4	$P_2\alpha$	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
5	$E\alpha$	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
6	$P_2\alpha$	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
7	$B_2\alpha$	-66	-10	-162	-71	90	-149	...	6	-147	-150	68	50	85	200	200
8	$E\alpha$	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
9	$P_2\alpha$	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
10	$B_1\alpha$	-66	-73	-197	-174	132	-253	...	-167	-273	-129	66	100	18	200	200
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Fig. 3. An example of a 3D profile. The example shows the first ten positions of the sperm whale myoglobin 3D profile (*59*). This profile was used in the compatibility search of Fig. 6. The environment group is listed for each position, followed by scores for placing each of the amino acids at that position. The actual profile is 153 positions long, the length of the sperm whale myoglobin sequence. The scores placed in each row are the 3D-1D scores of Fig. 5, multiplied by 100. The most effective gap penalties are determined empirically. In this case, gaps in helical regions were forbidden by setting very high gap penalties for the helical positions (positions 3 through 10 in the profile). In contrast, relatively low gap opening (Opn) and gap extension (Ext) penalties were used for the coil regions (positions 1 and 2).

Profile search for sperm whale myoglobin

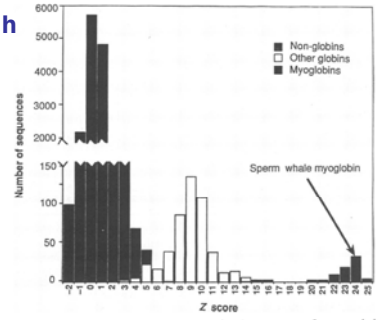


Fig. 6. Results of a compatibility search for the structure of sperm whale myoglobin. Myoglobin sequences are represented by black bars, other globin sequences are represented by white bars, and all other sequences are shown in gray bars. Sperm whale myoglobin is the eighth highest scoring protein (Z score = 23.7). Gaps were not allowed in helical regions (as defined in the protein data bank file). In nonhelical regions, a gap-opening penalty of 2.0 and a gap-extension penalty of 0.02 was used.

Comparison of sequence homology and 3D profile search

Table 1. A comparison of a sequence homology search and a compatibility search with CRP. All proteins with Z scores greater than 6.0 in either the sequence homology search or the compatibility search are listed. Z score (1D) refers to the scores obtained from a sequence homology search with a sequence profile constructed with the *Escherichia coli* CRP sequence. Z score (3D) refers to the scores obtained from a structure compatibility search with a 3D profile constructed from the *E. coli* CRP structure (*3f*). Percent identity refers to the percentage of identical amino acids in the sequences aligned with the program BESTFIT (*56*). For the sequence homology search, a gap-opening penalty of 4.5 and a gap-extension penalty of 0.05 was used. For the structure compatibility search, a gap-opening penalty of 5.0 and a gap-extension penalty of 0.05 was used. In the sequence homology search, the next highest scoring protein after fnr, Bam HI–ORF4 protein from Fowlpox virus, had an insignificant Z score of 4.90.

Protein	Z score (3D)	Z score (1D)	Percent identity
cAMP receptor protein—*E. coli* (CRP)	46.53	72.99	100.0
cAMP receptor protein—*Salmonella typhimurium* (CRP)	44.13	72.45	99.5
Hypothetical 24.1-kD protein—*Lactobacillus casei*	11.84	12.74	25.6
Regulatory protein fixK—*Rhizobium meliloti*	10.65	9.26	21.1
Regulatory protein fnr—*E. coli*	9.20	7.03	21.2
Protein kinase, cGMP-dependent—bovine	8.24	—	22.0
Protein kinase type III regulatory chain—fruit fly	6.62	—	20.9
DNA polymerase accessory protein 44—bacteriophage T4	6.58	—	19.7
Protein kinase type II regulatory chain—fruit fly	6.47	—	20.9
Protein kinase, cAMP-dependent, regulatory chain II-α—human	6.33	—	21.2
Protein kinase type I regulatory chain—fruit fly	6.15	—	20.9
Protein kinase, cAMP-dependent, type II regulatory chain—bovine	6.06	—	20.9

Comparison of a sequence homology search and a 3D profile search with ribose binding protein (RBP)

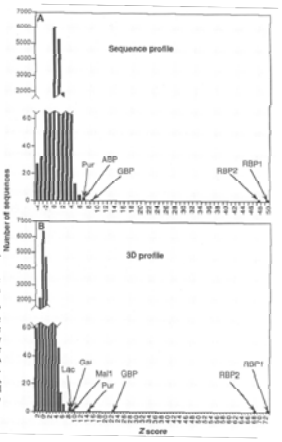


Fig. 7. Comparison of a sequence homology search and a structure compatibility search with ribose binding protein (RBP). (**A**) The results of a sequence homology search with a sequence profile constructed from the *E. coli* RBP sequence. The bar graph shows the number of sequences that give a particular Z score. A gap-opening penalty of 4.5 and a gap-extension penalty of 0.05 were used. The highest scoring proteins in (A) are RBP1 (*E. coli* RBP precursor, Z score = 49.0), RBP2 (*Salmonella typhimurium* RBP precursor, Z score = 47.9), GBP (*E. coli* galactose binding protein, Z score = 8.0), Pur (*E. coli* pur repressor, Z score = 6.1), and ABP (*E. coli* arabinose binding protein, Z score = 6.0). (**B**) The results of a structure compatibility search with a 3D profile constructed from the *E. coli* RBP structure. The bar graph shows the number of sequences that give a particular Z score. A gap-opening penalty of 5.0 and a gap-extension penalty of 0.2 were used. The highest scoring proteins labeled in (B) are RBP1 (Z score = 72.2), RBP2 (Z score = 68.9), GBP (Z score = 22.2), Pur (Z score = 14.2), Mal (*E. coli* Mal I protein, Z score = 9.0), Gal (*E. coli gal* repressor, Z score = 8.5), and Lac (*Klebsiella pneumoniae lac* repressor, Z score = 8.1).

Sequence compatibility search with a 3D structure profile for actin

Protein	Z score
69 of 71 Actin Sequences	88.11 ↕ 21.22
Kinase-related transforming protein (fgr)- feline sarcoma virus	17.47
Actin 5C - fruit fly	9.29
68-kD Heat shock protein - mouse	8.12
70-kD Heat shock protein - frog	7.95
70-kD Major heat shock - fruit fly	7.03
70-kD Heat shock cognate protein-bovine	**6.99**
HNRNP complex, protein C - frog	6.74
70-kD Heat shock cognate protein - human	6.31

Fig. 8. Sequence compatibility search with a 3D structure profile for actin (*47*). All sequences that received a Z score of 6.0 or greater are listed. A gap-opening penalty of 5.0 and a gap-extension penalty of 0.2 were used. The fgr protein is the result of a gene fusion between actin and a tyrosine-specific protein kinase (*63*). The bovine HSC70 protein, known to have a similar structure to actin, received a Z score of 6.99 and is shown in bold type.

Threading

- Sequences are fitted directly onto the backbone coordinates of known protein structures.
- Matching of sequences to backbone coordinates is performed in 3D space, incorporating specific pair interactions explicitly.

A new approach to protein fold recognition

D. T. Jones*†, W. R. Taylor† & J. M. Thornton*

* Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK
† Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs[1] or folds[2–4] in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

Nature 358, 86-89 (1992)

Threading

- A library of different protein folds is derived from the database of protein structures.
- Each fold is considered as a chain tracing through space; the original sequence being ignored completely.
- The test sequence is then optimally fitted to each library fold, allowing for relative insertions and deletions in loop regions.
- The 'energy' of each possible fit (or threading) is calculated by summing the proposed pairwise interactions and the solvation energy.
- The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match.

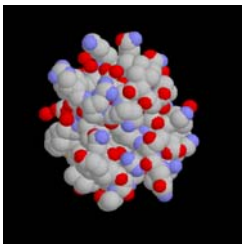
Knowledge-based (pair) potentials

$$E(r) = -kT \ln[f(r)]$$

r	distance between two atoms (or some other parameter, like dihedral angles or solvent accessible surface)
$E(r)$	is the energy at r
$f(r)$	is the probability density at r
k	is the Boltzmann constant
T	is the absolute temperature

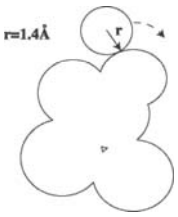
Solvent accessible surface

Represent atoms as spheres with appropriate radii and eliminate overlapping parts.



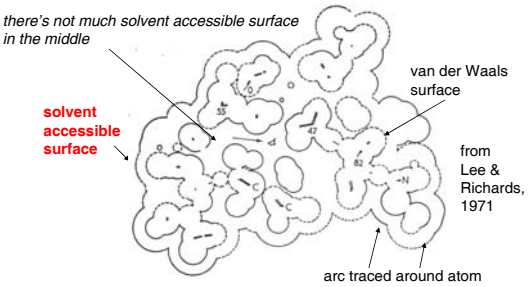
Mathematically roll a sphere all around that surface.

The sphere's center traces out a surface as it rolls.

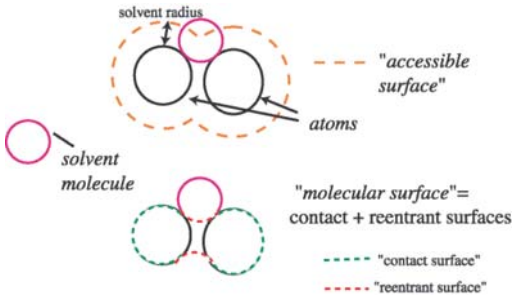


Lee & Richards, 1971
Shrake & Rupley, 1973

Cross-section (slice) of a protein structure:
Inner surfaces here are van der Waals. Outer surface is that *traced out by the center of the sphere* as it rolls around the van der Waals' surface. If any part of the arc around a given atom is traced out, that atom is accessible to solvent. The *solvent accessible surface* of the atom is defined as the sum the arcs traced around an atom.

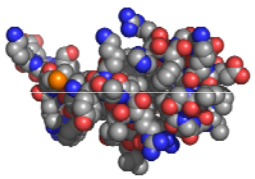


Accessible surface / Molecular surface

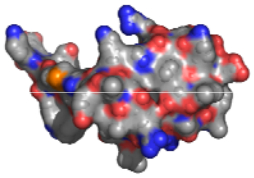


These are *alternative* ways of representing the surface which is essentially in contact with solvent.

Molecular surface of proteins



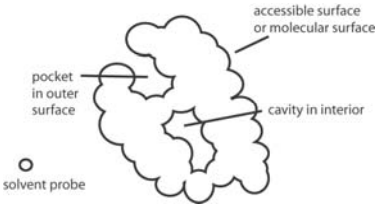
depiction of heavy atoms (O, N,C, S) in a protein as van der Waals spheres



depiction of the corresponding "molecular surface"--volume contained by this surface is vdW volume plus "interstitial volume"--spaces in between

The irregular surface of proteins: pockets and cavities

- A **pocket** is an empty concavity on a protein surface which is accessible to solvent from the outside.
- A **cavity** or **void** in a protein is a pocket which has no opening to the outside. It is an interior empty space inside the protein.

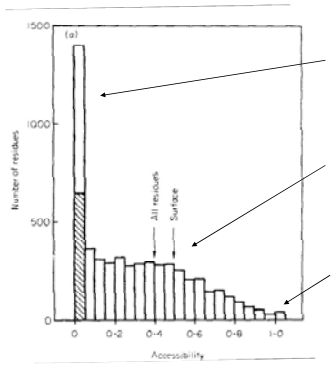


Pockets and cavities can be critical features of proteins in terms of their binding behavior, and identifying them is usually a first step in structure-based ligand design etc.

Fractional accessibility

- Calculate total solvent accessible surface of protein structure (also can calculate solvent accessible surface for individual residues/sidechains within the protein) .
- Model the accessible surface area in a *disordered or unfolded* protein using accessible surface area calculations on *model tripeptides* such as Ala-X-Ala or Gly-X-Gly.
- From these we can calculate *what fraction* of the surface is buried (inaccessible to solvent) by virtue of being within the folded, native structure of the protein.
- *Fractional accessibility* is computed by dividing the accessible surface area in the native protein structure by the accessible surface in the modelled unfolded protein. The *residue fractional accessibility* and *side-chain fractional accessibility* are calculated for individual residues or side-chains in the structure.

Distribution of residue fractional accessibilities



A sizeable group are completely buried (hatched) or nearly completely buried

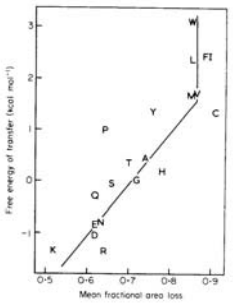
Broad distribution among non-buried residues; mean fractional accessibility for non-buried residues of around 0.5

Few residues are completely exposed to solvent, but that fractional accessibility >1 is possible

Residue fractional accessibility correlates with free energies of transfer for amino acids between water and organic solvents

The interior of a protein is akin to a nonpolar solvent in which the nonpolar sidechains are buried. Polar sidechains, on the other hand, are usually on the surface.
However, some polar side chains do get buried, and it must also be remembered that the backbone for every residue is polar, including those with nonpolar side chains. So a lot of polar moieties do get buried in proteins.

Miller, Janin, Lesk & Chothia (1987)
Fauchere & Pliska (1983)



Solvation potential

Similarly, the solvation potential for an amino-acid residue a is defined as

$$\Delta E^{a}_{solv}(r) = -RT \ln \left[\frac{f^{a}(r)}{f(r)} \right]$$

where r is the % residue accessibility (relative to residue accessibility in GGXGG fully extended pentapeptide), $f^{a}(r)$ is the frequency of occurrence of residue a with accessibility r, and $f(r)$ is the frequency of occurrence of all residues with accessibility r.

Pairwise pseudo-energy terms

For specified atoms (C$\beta \rightarrow$Cβ for example) in a pair of residues ab, topological level (sequence separation) k and distance interval s, the potential is given by the following expression

$$\Delta E^{ab}_{k} = RT \ln [1 + m_{ab}\sigma] - RT \ln \left[1 + m_{ab}\sigma \frac{f^{ab}_{k}(s)}{f_{k}(s)} \right]$$

where m_{ab} is the number of pairs ab observed at topological level k, σ is the weight given to each observation, $f_k(s)$ is the frequency of occurrence of all residue pairs at topological level k and separation distance s, and $f^{ab}_{k}(s)$ is the equivalent frequency of occurrence of residue pair ab. RT is taken as 0.582 kcal mol^{-1}. Short- (sequence separation, $k \leq 10$), medium- ($11 \leq k \leq 30$) and long- ($k > 30$) range potentials have been calculated between the following atom pairs: C$\beta \rightarrow$Cβ, C$\beta \rightarrow$N, C$\beta \rightarrow$O, N$\rightarrow$Cβ, N$\rightarrow$O, O$\rightarrow$Cβ and O$\rightarrow$N.

Statistically derived potentials

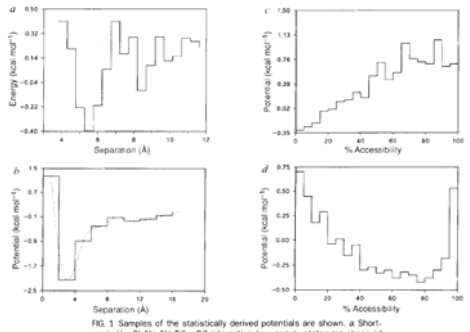
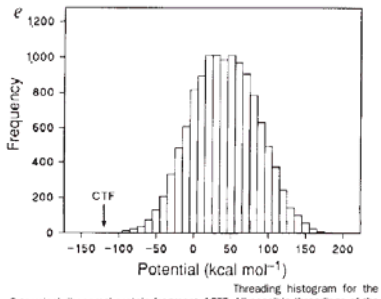


FIG. 1 Samples of the statistically derived potentials are shown. a Short-range (k = 3) Ala-Ala Cβ → Cβ interaction. Low-energy states are observed for distances around 6 Å, corresponding mainly to α-structure, and 9 Å, corresponding mainly to β-structure. b Long-range (k > 30) Cys-Cys Cβ → Cβ interaction. The most significant energy minimum around 4 Å corresponds to disulphide bridge formation. c Solvation potential for leucine, and d solvation potential for glutamic acid.

Threading histogram for 1CTF



Threading histogram for the C-terminal ribosomal protein fragment, 1CTF. All possible threadings of the CTF sequence on the CTF structure were computed (secondary structure gaps disallowed) and the energies of each threading calculated. The native threading is indicated, and was found to be the lowest energy threading.

Threading results

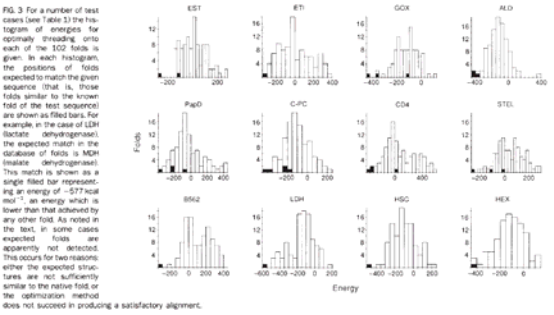


FIG. 3 For a number of test cases (see Table 1) the histogram of energies for optimally threading onto each of the 102 folds is given. In each histogram, the positions of folds expected to match the given sequence (that is, those folds similar to the known fold of the test sequence) are shown as filled bars. For example, in the case of LDH (lactate dehydrogenase), the expected match in the database of folds is MDH (malate dehydrogenase). This match is shown as a single filled bar representing an energy of −577 kcal mol^{-1}, an energy which is lower than that achieved by any other fold. As noted in the text, in some cases expected folds are apparently not detected. This occurs for two reasons: either the expected structures are not sufficiently similar to the native fold, or the optimization method does not succeed in producing a satisfactory alignment.

Summary of trial fold-recognition searches

TABLE 1 Summary of trial fold-recognition searches

Test protein	Source	Fold	Best match	ΔE	% Sequence identity	Matches
C-phycocyanin β (C-PC)	Red algae	Globin	1MBA	101	7	1, 2, 9, 18, 25
Glycolate oxidase (GOX)	Spinach	TIM barrel	1WSY(A)	52	10	1, 3, 49
Muscle aldolase (ALD)	Human	TIM barrel	4XIA(A)	80	6	1, 2, 3
Lactate dehydrogenase (LDH)	Dogfish	Rossmann	4MDH(A)	87	15	1*
Elastase (EST)	Pig	Trypsin	4PTP	110	35	1, 14
CD4	Human	Ig	2FB4(H)	87	10	1, 2, 31
Stellacyanin (STEL)	Varnish tree	Cu binding	2AZA(A)	18	14	1, 6, 20
Cytochrome B562 (B562)	*E. coli*	4-helix bundle	2MHR	78	6	1
Trypsin inhibitor DE-3 (ETI)	Kaffir tree	Interleukin 1β	1I1B	14	5	1
PapD-chaperonin	*E. coli*	Ig	2FB4(L)	64	15	1, 5, 9, 35
70K, Heat-shock cognate (HSC)	Cow	Actin	1ATN(A)	94	9	1
Hexokinase B (HEX)	Yeast	Actin	1ATN(A)	0	12	1

In each case the database included 102 protein chains, except where the test protein was itself in the database, in which case it was excluded. Templates for each chain were constructed as described in the text, with residues not in helices or strands (as calculated by DSSP[19]) assigned as loop residues. For the 70K heat-shock cognate protein and hexokinase searches, the coordinates for actin were also included (coordinates deposited under the code 1ATN, but not yet released). Proteins with > 25% sequence identity to the test protein were also excluded from the calculation of potentials. The pairwise and solvation terms were summed and stored separately, and standard deviations ($s.d._{pair}$ and $s.d._{solv}$) for the two contributing factors calculated over the set of 102 folds. To balance the contributions of the pairwise and solvation terms, the final energy was taken as $E = E_{pair} + WE_{solv}$, where $W = (s.d._{pair}/s.d._{solv})$. The 'confidence' of the match (ΔE) is given in terms of the absolute energy difference between the top scoring fold and the next highest scoring, different, fold. The 'best match' column gives the Brookhaven ID of the best matching chain fold (including chain identity where appropriate), along with the sequence identity between the best matching chain and the test protein. Positions in the sorted list of threading energies of similar folds are also shown. A constant set of alignment parameters (gap penalty for example) was used for all databank searches shown. Typical execution times for a single search of 102 chains are around 100 minutes on a Unix workstation (Sidbourne 5/6020). The 102 chains used were as follows: 351C, 4256B, 2AAT, 1ABP, A5ACN, 8ADH, 3ADK, A8ATC, B8ATC, A2AZA, 3BLM, 1BP2, 2CA2, A7CAT, 1CC5, 1CDR, A2CCY, 1CD4, 2CDY, 3CLA, 2CNA, I4CPA, 5CPA, 2CPP, 4CPV, 1CRN, 2CRO, E1CSE, I1CSE, 1CTF, 1CY3, 2CYP, 3DFR, A4DFR, A1DHF, 1ECA, E2ER7, H2FB4, L2FB4, 1FD2, 1FX1, 3FXC, 4FXN, A3GAP, 2GBP, 1GCR, 01GD1, 3GRS, A3HHB, 1I4P, A2HLA, B2HLA, 1HDE, 1I1B, 3ICB, 3ICD, 1LO1, 2LBP, 6LDH, 1LH1, 351RD, A2LTN, 1LZ1, 1MBA, 1MBD, A4MDH, 2MHR, 2OVO, A3PAB, 9PAP, 1PAZ, 1PCY, A1PPX, 3PGK, 3PGM, 1PHH, 5PTI, 4PTP, 1RHD, 2RHE, 2RNT, 7RSA, 4RXN, 2SGA, I45GB, 1SN3, 2SNS, 02SOD, 2SSI, 2STV, I1TGS, E2TMN, 4TNC, A1TNF, 1UBQ, 1UTG, A9WGA, R2WRP, A1WSY, B1WSY, A4XIA, A1YPI.

* Other topologically similar (yet structurally different) parallel αβ folds were positioned at 3, 7, 11, 12, 13, 17, 19, 31, 34, 82.

Critical components of fold recognition techniques

- Techniques producing useful alignments between sequences and structures
- Criteria for identifying native-like sequence/structure combinations
- Energy functions or parameter sets providing a reasonable description of protein-solvent systems

Fold recognition results from CASP

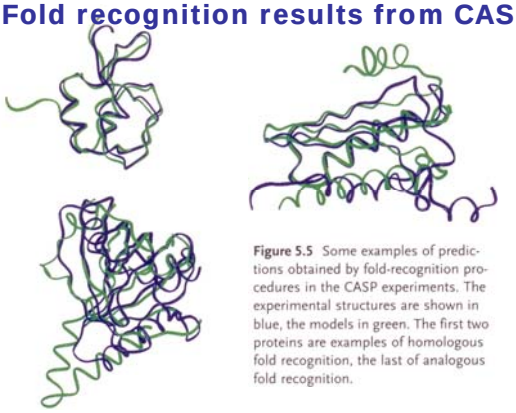


Figure 5.5 Some examples of predictions obtained by fold-recognition procedures in the CASP experiments. The experimental structures are shown in blue, the models in green. The first two proteins are examples of homologous fold recognition, the last of analogous fold recognition.

Literatur

- Anna Tramontano: *Protein Structure Prediction, Wiley-VCH*, 2006.
- J. U. Bowie, R. Lüthy & D. Eisenberg. *Science* 253, 164-170 (1991) [3D profiles]
- D. T. Jones, W. R. Taylor & J. M. Thornton. *Nature* 358, 86-89 (1992) [Threading]
- M. J. Sippl. *Current Opinion in Structural Biology* 5, 229-235 (1995) [Database-derived potentials]