Reprinted from *J. Mol. Biol.* (1993) **232**, 805–825

Prediction of Protein Structure by Evaluation of Sequence-structure Fitness

Aligning Sequences to Contact Profiles Derived from Three-dimensional Structures

Christos Ouzounis, Chris Sander, Michael Scharf and Reinhard Schneider

J. Mol. Biol. (1993) **232**, 805–825

Prediction of Protein Structure by Evaluation of Sequence-structure Fitness

Aligning Sequences to Contact Profiles Derived from Three-dimensional Structures

Christos Ouzounis, Chris Sander, Michael Scharf and Reinhard Schneider

Protein Design Group, EMBL, D-6900 Heidelberg, Germany

(Received 21 April 1992; accepted 17 March 1993)

The problem of protein structure prediction is formulated here as that of evaluating how well an amino acid sequence fits a hypothetical structure. The simplest and most complicated approaches, secondary structure prediction and all-atom free energy calculations, can be viewed as sequence-structure fitness problems. Here, an approach of intermediate complexity is described, which involves; (1) description of a protein structure in terms of contact interface vectors, with both intra-protein and protein-solvent contacts counted, (2) derivation of sequence preferences for 2 up to 29 contact interface types, (3) generation of numerous hypothetical model structures by placing the input sequence into a large set of known three-dimensional structures in all possible alignments, (4) evaluation of these models by summing the sequence preferences over all structural positions and (5) choice of predicted three-dimensional structure as that with the best sequence-structure fitness. Evolutionary information is incorporated by using position-dependent core weights derived from multiple sequence alignments. A number of tests of the method are performed: (1) evaluation of cyclic shifts of a sequence in its native structure; (2) alignment of a sequence in its native structure, allowing gaps; (3) alignment search with a sequence or sequence fragment in a database of structures; and (4) alignment search with a structure in a database of sequences. The main results are: (1) a native sequence can very well find its native structure among a large number of alternatives, in correct alignment; (2) substructures, such as $(\beta\alpha)_n$ units, can be detected in spite of very low sequence similarity; (3) remote homologues can be detected, with some dependence on the set of parameters used; (4) contact interface parameters are clearly superior to classical secondary structure parameters; (5) a simple interface description in terms of just two states, protein-protein and protein-water contacts, performs surprisingly well; (6) the use of core weights considerably improves accuracy in detection of remote homologues; (7) based on a sequence database search with a myoglobin contact profile, the C-terminal domain of a viral origin of replication binding protein is predicted to have an all-helical fold. The sequence-structure fitness concept is sufficiently general to accommodate a large variety of protein structure prediction methods, including new models of intermediate complexity currently being developed.

Keywords: protein structure prediction; sequence-structure alignment; computer algorithm; database; evolutionary information.

1. Introduction

(a) *A unified view of protein structure prediction: evaluation of sequence-structure fitness*

The prediction of the three-dimensional structure of a protein from its one-dimensional amino acid sequence is an unsolved problem, in spite of much effort, using a variety of approaches. The problem can be cast in the following form: given a protein sequence, generate and evaluate all possible three-dimensional structures and then choose the structure (or structures) that best fits the sequence. We call this the sequence-structure fitness approach to protein structure prediction.

(b) *The problem: generation and evaluation of hypothetical structures*

In generating hypothetical structures, one must explore all reasonable alternatives, e.g. by

0022–2836/93/150805–21 $08.00/0

constructing explicit three-dimensional models or enumerating secondary structure states. In evaluating a hypothetical structure, i.e. the fitness of the structure for the sequence or the fitness of the sequence for the structure, one must be able to distinguish between correct and incorrect structures, i.e. identify those structural states that have a high probability of being observed (in given environmental conditions). Success or failure depends crucially on the underlying description of structural states and on the evaluation scheme of sequence-structure fitness.

(c) *Most complicated: three-dimensional structures and free energy estimates*

At one extreme, in the most complicated description, one can generate all-atom co-ordinates for every conceivable structure and evaluate its fitness by calculation of the free energy of unfolding. This fully three-dimensional approach is impractical. It takes too much computer time, so only a small fraction of the available conformational space can be explored this way. Furthermore, even for a single structure, current estimates of free energy differences between significantly different conformational states are not sufficiently accurate (Novotny *et al.*, 1984, 1988).

(d) *Simplest: one-dimensional secondary structure strings and preferences*

At the other extreme, one of the simplest descriptions of protein structure is in terms of secondary structure, e.g. helix, β-strand and loop. All possible structures for a given sequence can be simply enumerated as combinations of secondary structure states for single residues and the fitness of structure for sequence can be evaluated by statistical preference parameters of single residues in secondary structure states (Garnier *et al.*, 1978). This inherently one-dimensional approach can be executed rapidly, but fails to solve the prediction problem. The description in terms of secondary structure alone is much too simple and ignores important physical effects. Furthermore, even with a 100% successful prediction of secondary structure one would still lack the three-dimensional view that is essential for a full understanding of function.

(e) *Intermediate: description of protein structure in terms of residue-residue contacts*

For these reasons we believe that an intermediate description of protein structure is a key requirement for further progress with structure prediction, a description between all-atom three-dimensional models and one-dimensional strings of secondary structure symbols. To be of practical value, the description should not be so complicated as to be intractable and not so simple as to neglect important effects. Accordingly, we have developed a description of protein structure in terms of contacts between atoms, including contacts with solvent atoms. In this, we follow an idea developed in the context of analyzing contacts in parallel and antiparallel β-sheets: "As more protein structures become available, further distinctions of secondary structure elements according to the type of tertiary contacts should be made. For example, one can distinguish between different hydrogen bonding positions in β-sheets, solvent-exposed and interior faces of sheets and helices, segments in tertiary contacts with sheets compared to those in contact with helices. Such distinctions are likely to lead to more clear-cut statistical preferences, and also serve as a starting point for predicting tertiary structure" (Lifson & Sander, 1979). The first results from our approach are reported in a diploma thesis (Scharf, 1989).

The contact description of protein structure is a two-dimensional reduction of the full complexity of three-dimensional structure, with the aim of capturing the physically relevant effects. The basic element in this description is the enumeration of atomic contacts made by a residue. Contacts are best represented by a "contact map" (Fig. 1a,b), similar to a "distance plot" used in distance geometry calculations and in protein structure determination by nuclear magnetic resonance spectroscopy. The description is of size $N(N+1)$, where N is the number of residues in the proteins and the slot number $(N+1)$ is allocated to residue-solvent contacts.

(f) *Generation of hypothetical structures: use of the Protein Data Bank*

With a contact description of protein structure, how do we, in a sequence-structure fitness approach, solve the two problems of generating all reasonable structures and evaluating each of them? The universe of all possible protein structures, even in the contact description, is very large. One very powerful and practical way of circumventing this enormous complexity is by working within the much smaller universe of all known protein structure types, as deposited, for the most part, in the Protein Data Bank (Bernstein *et al.*, 1977). For a given sequence, a set of alternative structures is then generated by implicitly trying out all possible alignments of the sequence in each of the known structures, with gaps permitted.

(g) *Evaluation in two-dimensions: residue-residue contacts*

If one represents the known protein structure "templates" by two-dimensional contact maps, the evaluation of alternate sequence alignments in each structure becomes a two-dimensional alignment problem. Algorithmically, the problem of finding the alignment with the highest score, assuming additivity of fitness values, can be solved in a number of ways, e.g. by a two-level dynamic

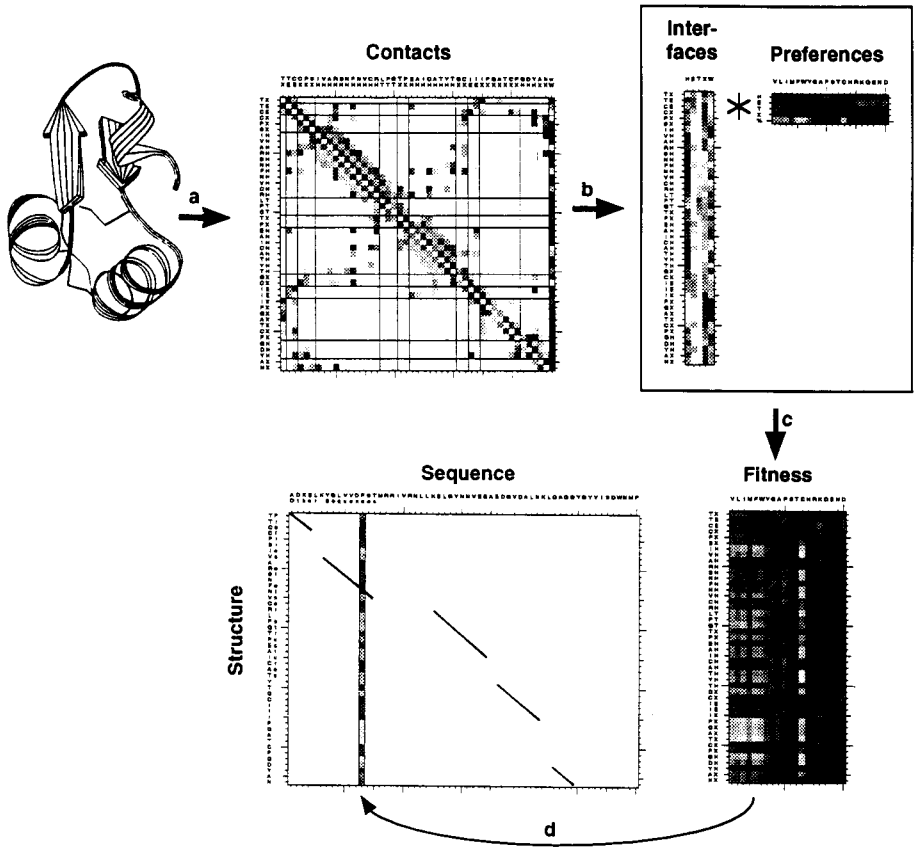
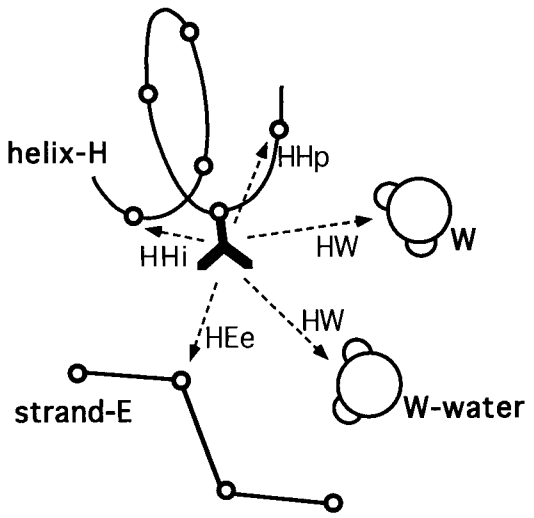


Figure 1. Scheme of the basic procedure of structure-sequence alignment. a. From 3-dimensional structure to a 2-dimensional residue-residue contact map. Starting from the 3-dimensional structure, the strength of residue-residue and residue-solvent contacts in the known 3-dimensional structure (Contacts) is calculated. Contact strength is indicated by grey levels; contacts with the solvent are in right-most column (W). b. From contact map to contact interface profile. For each residue, i.e. for each row in the contact map, contacts are summed over all contacting residues in each interface type (5 types in this Fig.). The resulting set of 5 interface strengths is called the contact interface vector and describes the structural environment of a residue. The array of interface vectors, 1 for each residue, is a simplified representation of protein structure, called the interface contact profile (Interfaces). This description is 1-dimensional in the sense that each interface vector describes the local structural environment, independently of the type of the contacting residues. Preferences for a residue type in each of the interface states (Preferences) are derived from the interface profiles of all known structures in the database, removing from the database the protein to be aligned. c. From contact interface profile to fitness profile. In preparation of alignment, the fitness of each of the 20 amino acids at each structural position is evaluated by simply summing over the preference of that residue type for each interface, weighted with the strength of the interface at the structural position. The resulting table (Fitness profile) represents the fitness of each of the 20 amino acid residue types for this structural position. Mathematically, the fitness profile $f(R,j)$ is simply the matrix product of the interface profile $\mathbf{c}(j,I)$ with the preference table $p(R,I)$. Such fitness profiles can also be derived from other types of structural preferences, e.g. those for secondary structure, as well as from multiple sequence alignments (Gribskov *et al.*, 1987, 1990). d. From fitness profile to structure-sequence alignment. The fitness profile is just the right form of input to an alignment problem, here that of aligning some amino acid sequence to the given structure. The local similarity value for the alignment problem, at a given sequence position, is simply the appropriate column copied from the fitness table, representing the fitness of that residue type for each position in the structure. A dynamic programming algorithm then finds the optimal trace, such that the sum of fitness values along the trace is optimal. The result is an alignment of the sequence with the given structure. The alignment is the basis of an explicit 3-dimensional model, and the total fitness value quantifies how well the sequence fits into this structure.

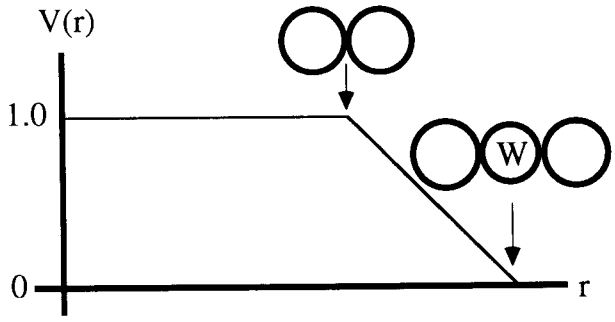
programming approach (Taylor & Orengo, 1989). Statistical preference parameters that can be used in this type of alignment problem are residue-residue contact preferences, pseudo-energies, or potentials of mean force: typically a set of one or more tables of dimension 20 by 20 (Crippen, 1977; Tanaka & Scheraga, 1975; Warme & Morgan 1978; Lifson & Sander, 1980; Galaktionov & Rodionov, 1981; Miyazawa & Jernigan, 1985; Scharf, 1989; Sippl, 1990; Sander & Vriend, 1991; Sander *et al.*, 1992), representing average interactions between the 20 different amino acid types. Only one of these (Sippl, 1990) has been used in full two-dimensional sequence-structure alignment (Jones *et al.*, 1992), as far as we know. Here, we use an approximate one-dimensional reduction of the problem (Fig. 1).



(0.1HHp, 0.2HHi, 0.3HEe, 0.4HW)

Contact interface vector of a residue describes its 3D environment

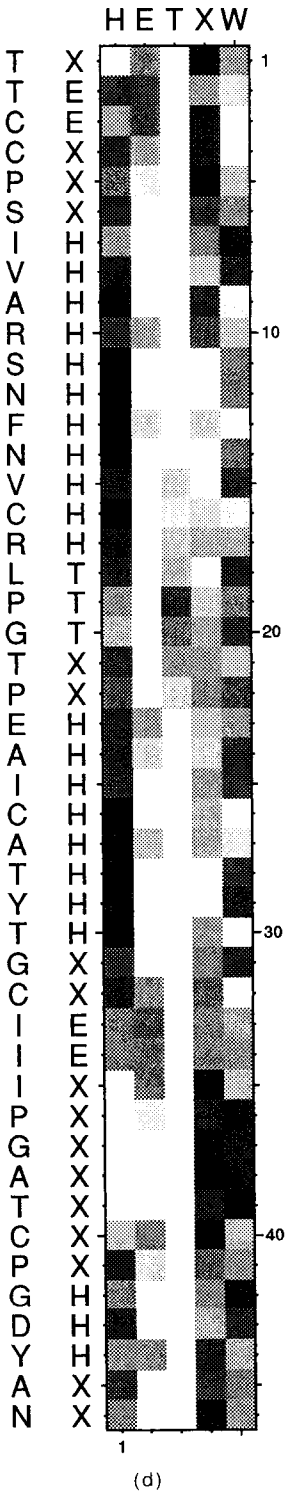
(a)



(b)

	Helix	Beta	Turn	Loop	Water
Helix	HHa HHi HHe	HEe	HTe	HXe	HW
Beta	HEe	EEa EEi EEs EEe	ETe	EXe	EW
Turn	HTe	TEe	TTa TTi TTe	TXe	TW
Loop	XTe	XEe	XTe	XXa XXi XXe	XW

(c)



(d)

Figure 2. Description of 3-dimensional protein structure by a 1-dimensional array of contact interface vectors. (a) Cartoon of a residue contact environment. In this example, a valine residue (Y-shaped side-chain) in a helix makes contacts (arrows) with residues (circles) adjacent in the same helix (interface HHp), on the same helix 1 turn away (interface HHi), on a beta-strand (interface HEe) and with solvent (interface HW). Not all contacts are shown, for clarity of the Figure, nor are side-chains. The extent to which a residue participates in various interfaces is calculated in terms of interatomic contacts. Only contacts with nearest neighbours are counted and summed for all atoms in the residue. A given atom makes contacts either with other protein atoms or with solvent atoms (substrate or co-factor atoms are treated here as solvent, for simplicity). When solvent contacts are expressed in the appropriate units (Colonna-Cesari & Sander, 1990), the total number of contacts an atom makes is approximately constant and proportional to the volume of a spherical shell. This allows an estimate of water contacts from the calculation of protein-protein contacts alone. The total number of contacts of a residue can then be partitioned according to the type and state of the contacting atoms. The result is a normalized contact interface vector for each residue (bottom). (b) Contacts are

(h) *Evaluation in one dimension: residue-interface contacts*

The two-dimensional contact view can be simplified by describing the contacts made by a single residue with its environment, i.e. by averaging over the amino acid type of the second contact partner. For example, a two-dimensional contact map enumerated in terms of pair contacts, e.g. Ala-Val, Ala-Ile, Ala-Thr, Ala-Water, etc. would be collapsed, e.g. to Ala–Protein, Ala–Water (with relative strengths summed up). In this way, a structure can be represented as a string of structural contact vectors attached to residue positions (Fig. 1; Fig. 2(d)). The description is of size NK, where K is the number of distinct contact interface types ($K-1$ intra-protein interfaces and one protein-water interface).

The representation of protein structures as one-dimensional strings of contact descriptors has the practical advantage that sequence-structure alignment can now be performed by "standard" dynamic programming algorithms developed for string comparisons (Levenshtein 1966; Smith & Waterman, 1981). The corresponding fitness parameters express the preference (information value, potential of mean force) of single residues for a particular type of contact interface.

Here, we describe and parametrize the sequence-structure fitness approach in the one-dimensional approximation. The full two-dimensional view in terms of explicit residue-residue contacts will be dealt with in a subsequent manuscript. Our initial approach (Scharf, 1989) is conceptually similar to methods developed independently by other groups (Bashford *et al.*, 1987; Bowie *et al.*, 1990, 1991; Hendlich *et al.*, 1990; Sali *et al.*, 1990; Sali & Blundell, 1990; Sippl & Weitckus, 1992; Jones *et al.*, 1992; Godzik & Skolnick, 1992; Godzik *et al.*, 1992; Luethy *et al.*, 1992; Goldstein *et al.*, 1992*b*).

2. Methods

Central to our approach is the method of optimally aligning a single protein sequence ("input sequence": a string of amino acid types), and a single protein structure ("template structure": a string of contact descriptors). When the single sequence is aligned against a representative database of protein folds, the highest score corresponds to a predicted 3-dimensional structure for the input sequence (Fig. 1).

(a) *Residue-interface contact vector*

Starting from the co-ordinates of the template structure, intramolecular contacts and solvent contacts are calculated for each residue, as a sum over the contacts made by its constituent atoms (Fig. 2(a) and (b)). Solvent contacts are always included in the description. The nature of the contacting partners is noted, e.g. the secondary structure of the residues, the polar/non-polar or protein/solvent nature of the contacting atoms. The residue contact vector $\boldsymbol{c}(j)$ describes the contact environment of a particular residue.

We define the contact strength between 2 atoms such that 2 atoms are counted as being in contact if they are so close that a water molecule cannot fit between them: the linear-square box potential function $V(r)$ (Fig. 2(b)) is equal to 1·0 until the interatomic distance r equals the sum of the van der Waals radii (here 3·6 Å) and decreases linearly until it reaches 0·0, when a water molecule just fits between the 2 atoms, i.e. at $r = 3{\cdot}6+2{\cdot}8$ Å.

The contact strength $\boldsymbol{c}(j,I)$ of a residue of type R at position j in the structure is the sum over all interatomic contacts the residue makes in interface I. In this paper, only contacts made by the side-chain or C^{α} of residue j with side-chain or backbone of other residues are counted, i.e. contacts made by backbone N-H or C-O of residue j

calculated using a simple atom-atom potential $V(r)$ that is fairly insensitive to errors in atomic co-ordinates: a linear-square box. A single contact event is counted with a strength of 1·0 when 2 atoms are "touching" (interatomic distance r less or equal to the sum of the van der Waals radii). With increasing interatomic distance the strength decreases linearly and reaches 0·0 when a water molecule (W) just fits between 2 atoms. No contacts are counted when the interatomic distance is larger than the sum of the van der Waals radii plus the diameter of a water molecule. The contact strength of a residue is the sum of the contact strengths over all its atoms. Here, to emphasize specificity, a contact of a residue with its environment excludes contacts made by its backbone (i.e. only contacts of side-chain plus C^{α} with all other atoms are counted). c. Definition of 29 different contact interface types. A residue on a helix (H), β-strand (E), turn (T) or loop (X) (rows) can be in contact with other residues on a helix, β-strand, turn or loop or with water (columns). In the simplest description, with 2 interface types (used for parameter set AM2, 4c,e), the contacts in the rightmost column are grouped together as protein-water contacts (PW), the rest as protein-protein contacts (PP). Notation: interface type is $S_1S_2p_{12}$ (HHa, HHi, HHe, HEe, etc.), where S_1 = H,E,T, or X is the secondary structure of a residue; S_2 = H,E,T, or X is the secondary structure of the contacting residue(s) or S_2 = W for contacts with water; and p_{12} = a,i,s,e is the chain distance (proximity) of any 2 contacting residues, defined as: a (adjacent: the 2 residues are adjacent in sequence), i (internal: the 2 residues are on the same element of secondary structure, i.e. on the same helix or strand), s (strand-strand: the 2 residues are on adjacent strands in the same beta sheet), e (external: the 2 residues are on 2 different elements of secondary structure; for strands, in different sheets). (d) Description of the 3-dimensional structure of crambin (PDB (Protein Data Bank) data set 1crn, Hendrickson & Teeter, 1981) in terms of contact interface vectors $\boldsymbol{c}(j)$. For each residue we have, left to right, the amino acid residue (1-letter code), the secondary structure state, according to the DSSP summary column, except that H,G,I→H and S,B,blank→X (Kabsch & Sander, 1983), and the contact strength in each of the 5 interface types (AS5), H,E,T,X,W, on a scale of 0·0 (white square) to 100·0 (black square), with grey squares for intermediate values (see bottom legend). For aligning a given sequence with this structural template, one needs contact interface preference parameters for residue types in these states. Their derivation is explained in Methods and numerical values are given in Fig. 4.

are not counted. In general, a residue participates in K different interfaces (Fig. 2(a) and (c)). For the purposes of this paper, the vectors $\boldsymbol{c}(j)$ with the components $\boldsymbol{c}(j,I)$, $I = 1..K$ are normalized to unit length for each residue j. This is done to prevent sequence "read-through", i.e. to eliminate information about residue size. In the current implementation, the structure at position j is represented by the contact vector derived from a single native structure. In future implementations, $\boldsymbol{c}(j)$ can be an average over all equivalent positions in structurally similar proteins.

The simplest contact interface classification is in terms of 2 types, i.e. contacts made with other protein parts ("inside" or I = PP for protein-protein) and contacts made with solvent ("outside" or I = PW for protein-water). Note that the water contact strength $\boldsymbol{c}(j,\mathrm{PW})$ of a residue can have any value between 0 and 1, a distinctive advantage relative to methods that have to classify a residue as either being exposed or "buried" in binary fashion. If the secondary structure of the contacting partner is noted, one has 5 contact types, i.e. contact with a helix (I = PH for protein-helix), with a strand (I = PE), with a turn (I = PT), with a loop (I = PX) or with water (I = PW). The most complicated classification (in this paper) is in terms of the secondary structure state of both contacting partners and of chain proximity, for a total of 29 contact interface types (Fig. 2(c)).

The string $\boldsymbol{c}(j)$, $j = 1..N$ of contact vectors (Fig. 2(d)), also called the contact profile, is an approximate description of the 3-dimensional structure of the desired intermediate complexity. The profile has length N, where N is the number of residues in the protein structure, and width K, where K is the number of different types of contact interfaces. An example of a 5-state contact profile for the small protein crambin is in Fig. 2(c). With this description of protein structures we now derive preference parameters for residue participation in different interface types.

(b) *Residue-interface contact preferences*

To quantify how well particular residues are adapted to particular interfaces, we choose a statistical approach. Preference parameters are defined as the logarithm of the ratio of observed over expected contact counts. These are derived from a database of 64 mutually non-homologous proteins (Fig. 3).

Scanning over the database, total contact counts $C(R,I)$ for each residue type R in each interface type I are calculated by summing single residue contacts $c(R(j),I)$ over all residues j of type R in all protein chains:

$$C(R,I) = \sum_j \boldsymbol{c}(j,I), \text{ residue at } j \text{ is of type } R. \qquad (1)$$

For example, C(Ala, HEe) = 1121·0 is the total contact strength of all alanine residues in helix-sheet interfaces.

From the raw contact counts $C(R,I)$, preferences are derived as:

$$p(R,I) = ld\left(\frac{C(R,I)C}{C(R)\,C(I)}\right), \qquad (2)$$

where $C(R)$, $C(I)$ and C are the partial sums:

$$C(R) = \sum_I C(R,I),\ C(I) = \sum_R C(R,I),\ C = \sum_{R,I} C(R,I), \qquad (3)$$

and ld is the logarithm base 2. The set of preferences $p(R,I)$, $I = 1..K$ can also be interpreted as a vector, written $\boldsymbol{p}(R)$.

These formulas are identical to those used in, e.g. secondary structure prediction, except that the counts C used here are contact strengths (sums over atomic contacts) rather than residue occurrences. The preferences can be interpreted as the logarithm of the ratio of observed counts, $C(R,I)$, over expected counts, $E(R,I) = C(R)C(I)/C$, or as the information in units of bits that a residue of type R has about the structural state I (e.g. R = Ala, I = HEe). Low counts can introduce serious fluctuations. This problem is dealt with here by scaling the preference values $p(R,I)$ with a damping factor, $\min(1{\cdot}0, E(R,I)/100)$. The factor is 1·0 if $E(R,I)$ is 100 or more and linearly decreases for lower expected counts until it reaches 0·0 for $E(R,I) = 0$. Thus, states with very low expected counts have "neutral" preference values near 0. This is an important technical detail. Other forms of damping noise from low counts have been used (Sippl, 1990). The set of preferences $p(R,I)$, $I = 1..K$ of 1 residue type for the various contact interface types is called here the preference vector for that residue type (Figs 1 and 4). Numerical values are tabulated in Sander & Vriend (1991) and Sander *et al.* (1992).

(c) *Evaluation of sequence-structure fitness*

Given a position j in the template structure and a residue of a certain type R in the input sequence, one can now quantify how well the residue is adapted to this position, i.e. one can calculate the sequence-structure fitness at this position. We simply need to accumulate the preferences of residue type R for the various interface types, weighted with the extent to which interface I is present in the contact environment at position j. Mathematically, this weighted sum is equivalent to the scalar (inner) product of the contact interface vector $\boldsymbol{c}(j,I)$ and the contact preference vector $p(R,I)$. For each protein structure template we precalculate and store the fitness of all 20 amino acid types at each structural position j as sequence-structure fitness profiles:

$$f(R,j) = \sum_I p(R,I)\,\boldsymbol{c}(j,I), \qquad (4)$$

where R covers all 20 amino acid types, I all interface types (Fig. 5(a)). The underlying assumption is that of statistical independence of the individual residue terms. An example of a fitness profile for crambin is given in Fig. 5(b). It has 20 real numbers, one for each residue type (R = VL..ND), at each sequence position. The profiles $f(R,j)$ are analogous to profiles used in sequence database searches (Gribskov *et al.*, 1987, 1990), and can be used as input to profile alignment software, e.g. MaxHom (Sander & Schneider, 1991; and unpublished). However, here they are derived from quantification of residue contacts in the 3-dimensional structure, rather than from the amino acid sequence. The profiles can be used to evaluate different arrangements of an input sequence in a template structure.

(d) *Sequence-structure alignment algorithm*

A dynamic programming algorithm provides an efficient way of determining the best arrangement of a particular sequence in a particular structure. We use a straightforward adaptation of the alignment algorithm for pairs of sequences described by Smith & Waterman (1981). Sequence similarity between residues at position i in one protein and j in the other protein is simply replaced by the local sequence-structure fitness $f(R,j)$, where R is the residue type at i. The global similarity, or fitness, is

#PID	C	SIZ	RES	%H	%B	%BP	%BA	SID	ORIGIN	PROTEIN_NAME
351C	_	82	1.6	50	4	0	100	C551$PSEAE	PSEUDOMONAS AERUGINOSA	CYTOCHROME C 551
256B	A	106	1.4	79	0	0	0	C562$ECOLI	ESCHERICHIA COLI	CYTOCHROME B 562
8ADH	_	374	2.4	28	24	45	55	ADHE$HORSE	EQUUS CABALLUS	ALCOHOL DEHYDROGENASE
8ATC	A	310	2.5	40	15	100	0	PYRB$ECOLI	ESCHERICHIA COLI	ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)
8ATC	B	146	2.5	15	34	1	98	PYRI$ECOLI	ESCHERICHIA COLI	ASPARTATE CARBAMOYLTRANSFERASE (ASPARTATE TRANSCARBAMYLASE)
2AZA	A	129	1.8	16	35	36	63	AZUR$ALCDE	ALCALIGENES DENITRIFICANS	AZURIN
3B5C	_	85	1.5	31	23	25	75	CYB5$BOVIN	BOS TAURUS	CYTOCHROME B 5
3BLM	_	257	2.0	42	17	0	100	BLAC$STAAU	STAPHYLOCOCCUS AUREUS	BETA-LACTAMASE
2CA2	_	256	1.9	16	30	23	76	CAH2$HUMAN	HOMO SAPIENS	CARBONIC ANHYDRASE II (CARBONATE DEHYDRATASE)
1CCR	_	111	1.5	42	1	0	100	CYC$ORYSA	ORYZA SATIVA	CYTOCHROME C
2CCY	A	127	1.7	74	1	0	100	CYCP$RHOMO	RHODOSPIRILLUM MOLISCHIANUM	CYTOCHROME C'
1CD4	_	173	2.3	5	41	11	88	CD4$HUMAN	HOMO SAPIENS, recombinant	T-CELL SURFACE GLYCOPROTEIN CD4 (N-TERMINAL FRAGMENT)
3CLA	_	213	1.8	29	28	23	76	CAT3$ECOLI	ESCHERICHIA COLI, engineered	CHLORAMPHENICOL ACETYLTRANSFERASE TYPE III
5CPA	_	307	1.5	38	16	63	36	CBPA$BOVIN	BOS TAURUS	CARBOXYPEPTIDASE A
2CPP	_	405	1.6	51	10	11	88	CPXA$PSEPU	PSEUDOMONAS PUTIDA	CYTOCHROME P450CAM (CAMPHOR MONOOXYGENASE)
4CPV	_	108	1.5	56	1	0	100	PRVB$CYPCA	CYPRINUS CARPIO	CALCIUM-BINDING PARVALBUMIN
1CSE	E	274	1.2	30	20	73	26	SUBT$BACLI	BACILLUS SUBTILIS	SUBTILISIN
1CSE	I	63	1.2	22	33	44	55	ICIC$HIRME	HIRUDO MEDICINALIS	EGLIN-C
1CTF	_	68	1.7	55	26	0	100	RL7$ECOLI	ESCHERICHIA COLI	50S RIBOSOMAL PROTEIN L7/L12 (C-TERMINAL DOMAIN)
2CYP	_	293	1.7	50	7	8	91	CCPR$YEAST	SACCHAROMYCES CEREVISIAE	CYTOCHROME C PEROXIDASE
8DFR	_	186	1.7	23	33	57	42	DYR$CHICK	GALLUS GALLUS	DIHYDROFOLATE REDUCTASE
1ECN	_	136	1.4	75	0	0	0	GLB3$CHITH	CHIRONOMOUS THUMMI)	HEMOGLOBIN (ERYTHROCRUORIN) (FRACTION III)
2ER7	E	330	1.6	11	45	13	86	CARP$CRYPA	ENDOTHIA PARASITICA	ASPARTIC PROTEINASE (ENDOTHIAPEPSIN)
4FD1	_	106	1.9	33	14	0	100	FER1$AZOVI	AZOTOBACTER VINELANDII	FERREDOXIN
4FXN	_	138	1.8	36	22	95	4	FLAV$CLOSP	CLOSTRIDIUM MP	FLAVODOXIN
3GAP	A	208	2.5	30	14	0	100	CRP$ECOLI	ESCHERICHIA COLI	CATABOLITE GENE ACTIVATOR PROTEIN
2GBP	_	309	1.9	43	19	90	10	DGAL$ECOLI	ESCHERICHIA COLI	D-GALACTOSE/D-GLUCOSE BINDING PROTEIN
1GCR	_	174	1.6	7	46	0	100	CRGB$BOVIN	BOS TAURUS	CRYSTALLIN GAMMA-II
1GD1	O	334	1.8	29	29	52	47	G3P$BACST	BACILLUS STEAROTHERMOPHILUS	D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
1GOX	_	350	2.0	44	13	78	21	2HAO$SPIOL	SPINACIA OLERACEA	GLYCOLATE OXIDASE
1GP1	A	183	2.0	32	18	47	52	GSHP$BOVIN	BOS TAURUS	GLUTATHIONE PEROXIDASE
2HLA	B	99	2.6	0	49	0	100	HA1H$HUMAN	HOMO SAPIENS	HISTOCOMPATIBILITY CLASS I ANTIGEN
1HOE	_	74	2.0	0	48	0	100	IAA$STRTE	STREPTOMYCES TENDAE	ALPHA-AMYLASE INHIBITOR
1I1B	_	151	2.0	5	47	0	100	IL1B$HUMAN	HOMO SAPIENS, recombinant	INTERLEUKIN-1 BETA
4ICD	_	414	2.5	39	18	52	47	IDH$ECOLI	ESCHERICHIA COLI	ISOCITRATE DEHYDROGENASE
1IL8	A	71	NMR	26	25	0	100	IL8$HUMAN	HOMO SAPIENS, recombinant	INTERLEUKIN 8
1L13	_	164	1.7	64	9	0	100	LYCV$BPT4	BACTERIOPHAGE T4, mutant	LYSOZYME
6LDH	_	329	2.0	43	17	51	48	LDHM$SQUAC	SQUALUS ACANTHIAS	LACTATE DEHYDROGENASE
2LIV	_	344	2.4	44	19	73	26	LIVJ$ECOLI	ESCHERICHIA COLI	LEU/ILE/VAL-BINDING PROTEIN
2LTN	A	181	1.7	1	43	0	100	LEC$PEA	PISUM SATIVUM, recombinant	LECTIN
2LTN	B	47	1.7	8	63	0	100	LEC$PEA	PISUM SATIVUM, recombinant	LECTIN
1LZ1	_	130	1.5	39	12	11	88	LYC$HUMAN	HOMO SAPIENS	LYSOZYME
1MBD	_	153	1.4	77	0	0	0	MYG$PHYCA	PHYSETER CATODON	MYOGLOBIN
2MHR	_	118	1.7	70	0	0	0	HEMM$THEZO	THEMISTE ZOSTERICOLA	MYOHEMERYTHRIN
2PAB	A	114	1.8	7	51	16	83	TTHY$HUMAN	HOMO SAPIENS	PREALBUMIN
1PAZ	_	120	1.6	16	37	35	64	AZUP$ALCFA	ALCALIGENES FAECALIS	PSEUDOAZURIN
4PTP	_	223	1.3	10	34	2	97	TRYP$BOVIN	BOS TAURUS	BETA TRYPSIN
1R69	_	63	2.0	63	0	0	0	RPC1$BP434	PHAGE 434	434 REPRESSOR (N-TERMINAL DOMAIN)
1RHD	_	293	2.5	29	13	87	12	THTR$BOVIN	BOS TAURUS	RHODANESE
7RSA	_	124	1.3	20	35	3	96	RNP$BOVIN	BOS TAURUS	RIBONUCLEASE A
2RSP	A	115	2.0	5	41	17	82	GAG$RSVP	ROUS SARCOMA VIRUS	RSV PROTEASE
5RXN	_	54	1.2	16	22	0	100	RUBR$CLOPA	CLOSTRIDIUM PASTEURIANUM	RUBREDOXIN
2SGA	_	181	1.5	9	55	6	93	PRTA$STRGR	STREPTOMYCES GRISEUS	PROTEINASE A
4SGB	I	51	2.1	0	29	11	88	IPR2$SOLTU	SOLANUM TUBEROSUM	SERINE PROTEINASE B INHIBITOR PCI-I
2SNS	_	141	1.5	20	22	15	85	NUC$STAAU	STAPHYLOCOCCUS AUREUS	STAPHYLOCOCCAL NUCLEASE
2SOD	O	151	2.0	1	42	2	97	SODC$BOVIN	BOS TAURUS	CU,ZN SUPEROXIDE DISMUTASE
2SSI	_	107	2.6	15	28	5	95	ISUB$STRAO	STREPTOMYCES ALBOGRISEOLUS	SUBTILISIN INHIBITOR
2STV	_	184	2.5	11	47	1	98	COAT$STNV	SATELLITE TOBACCO NECROSIS VIRUS	COAT PROTEIN
2TMN	E	316	1.6	40	17	26	73	THER$BACTH	BACILLUS THERMOPROTEOLYTICUS	THERMOLYSIN
1TNF	A	152	2.6	1	44	0	100	TNFA$HUMAN	HOMO SAPIENS, recombinant	TUMOR NECROSIS FACTOR-ALPHA
2TS1	_	317	2.3	54	10	85	14	SYY$BACST	BACILLUS STEAROTHERMOPHILUS	TYROSYL-tRNA SYNTHETASE
1UBQ	_	76	1.8	23	34	25	75	UBIQ$HUMAN	HOMO SAPIENS	UBIQUITIN
1UTG	_	70	1.3	75	0	0	0	UTER$RABIT	ORYCTOLAGUS CUNICULUS	UTEROGLOBIN
2WRP	R	104	1.6	78	0	0	0	TRPR$ECOLI	ESCHERICHIA COLI	TRP REPRESSOR
1WSY	A	248	2.5	50	13	100	0	TRPB$SALTY	SALMONELLA TYPHIMURIUM	TRYPTOPHAN SYNTHASE
4XIA	A	393	2.3	47	10	85	14	XYLA$ARTS7	ARTHROBACTER SP	D-XYLOSE ISOMERASE
1YPI	A	247	1.9	43	17	96	3	TPIS$YEAST	SACCHAROMYCES CEREVISIAE	TRIOSE PHOSPHATE ISOMERASE

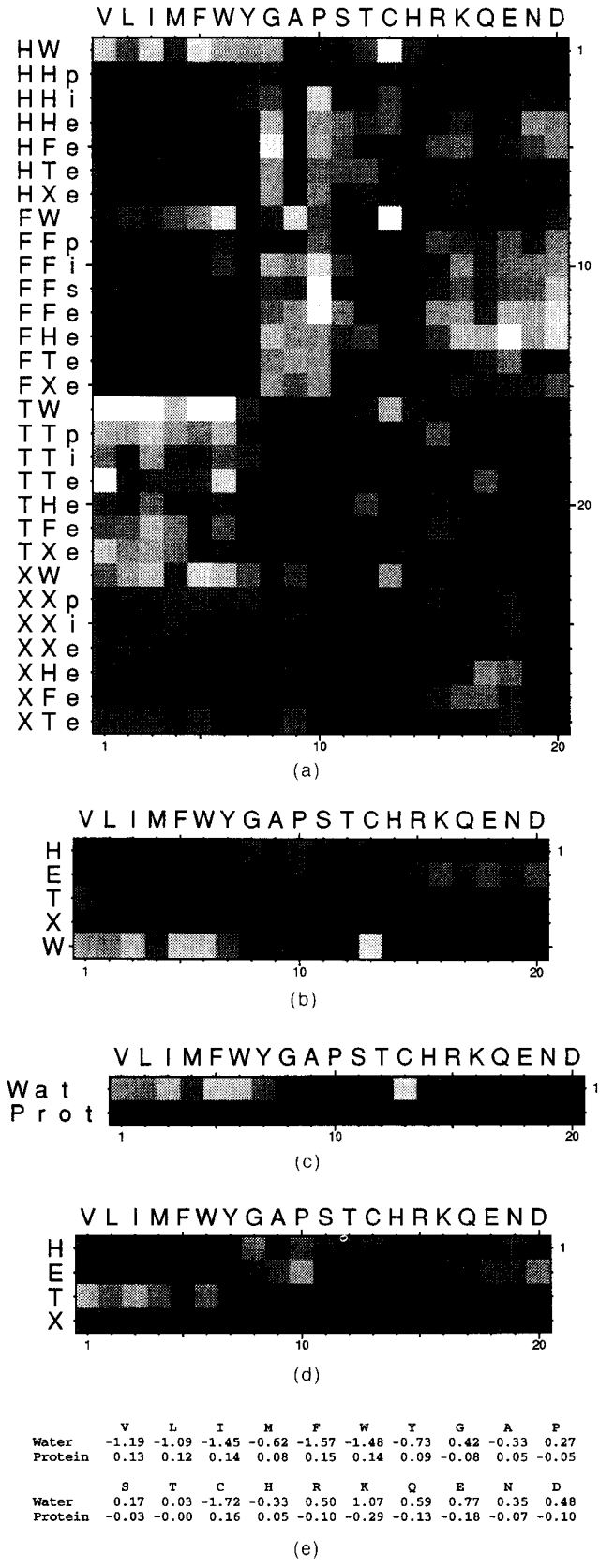
Figure 3. Representative subset of proteins of known 3-dimensional structure used as the database for the derivation of contact preference parameters. A total of 67 chains in 64 proteins with 12,460 amino acid residues were selected from the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). No 2 protein chains in the selection are homologous to each other, using as a criterion the threshold for structural homology (Sander & Schneider, 1991). For protein chains longer than 80 residues this translates to a sequence identity of maximally 24·8% for any 2 proteins in the list. The threshold is higher for shorter chains, e.g. 32·3% for length 50. The shortest chain has 47 residues, the longest 414. Steps in the selection of the non-homologous set were as follows (Hobohm *et al.*, 1992): Of the more than 400 data sets of experimental protein structures in PDB, 111 mutually non-homologous chains were selected by performing all pair comparisons using a dynamic programming alignment algorithm, and then selecting unique data sets one by one, going down a list sorted according to resolution and not using chains homologous to a previously selected one. A subsequent exclusion filter imposed the following requirements: crystallographic resolution as listed in PDB 2·6 Å or better (RES); chain length 40 or more residues (SIZ); number of Cys residues involved in disulfide bonds less than 8% of the chain length (CSS); well-formed secondary structure, i.e. 35 or more hydrogen bonds involved in helix (%H) or β-strand (%B) secondary structure per 100 residues (SEC) (sum of hydrogen bonds in parallel (%BP) and antiparallel (%BA) bridges, and in $(i,i+4)$ and $(i,i+3)$ type H-bonds (%H), as defined in Kabsch & Sander, 1983); number of heteroatoms (non-protein atoms) less than 8% of protein atoms (HET); not membrane protein (MEM). Protein chains are identified by the PDB 4 letter code (PID), the chain identifier (C), the name of the corresponding Swiss-Prot (Bairoch & Boeckmann, 1991) protein sequence entry (SID) (identified by: Swiss-Prot line DR; or, sequence identity between Swiss-Prot and PDB entry more than 98% and total sequence length within 3 residues), the species (ORIGIN) and the protein name.

calculated by summing over all pairs (i,j) in the alignment.

The dynamic programming algorithm requires additivity and independence. Both are fulfilled: the fitness values are additive because the fitness values are logarithms of probabilities, assumed to be independent of one another, which are multiplied along the alignment trace; and, the best path ending at positions i and j is independent of subsequent choices, as the contact profile is a 1-dimensional string of contact interface vectors derived solely from the template structure, with no dependence on the amino acid type of the 2nd contact partner that results from fitting the input sequence into the template structure.

(e) *Alignment parameters*

In the dynamic programming algorithm, 3 parameters have to be set at the appropriate values in order to obtain realistic alignments. The 1st parameter determines the



	V	L	I	M	F	W	Y	G	A	P
Water	-1.19	-1.09	-1.45	-0.62	-1.57	-1.48	-0.73	0.42	-0.33	0.27
Protein	0.13	0.12	0.14	0.08	0.15	0.14	0.09	-0.08	0.05	-0.05

	S	T	C	H	R	K	Q	E	N	D
Water	0.17	0.03	-1.72	-0.33	0.50	1.07	0.59	0.77	0.35	0.48
Protein	-0.03	-0.00	0.16	0.05	-0.10	-0.29	-0.13	-0.18	-0.07	-0.10

(e)

Figure 4. Contact preference parameters extracted from the non-redundant dataset of 67 protein chains. The parameters can be used to evaluate how well a particular sequence fits into a particular 3-dimensional structure, e.g. in sequence-structure alignment. For example, line HHe contains the single residue preference for helix-helix interfaces, line HEe those for helix-sheet interfaces. The 20 standard amino acids are given in 1-letter code. Secondary structure notation is: H (helix), E (extended or beta sheet), T (hydrogen bonded turn), X (everything else, called loop). Contacts with water are labelled W. Chain distance (proximity) of 2 contacting residues is: a = adjacent, i = internal to strand or helix, s = strand-strand, e = different segments, different sheets. Numerical values for the parameter sets are given by Sander & Vriend (1991) and Sander *et al.* (1992). (a) AInt29: Preferences of amino acid side-chains for contacts in 29 contact interface types, as defined in Fig. 2(c). For example, Pro has a strong preference for TTP (Pro located in a turn makes contacts with other residue(s) in the same turn); Lys has clear preference for HW, EW, TW, and XW (Lys located on any element of secondary structure make contacts with water); the strongest preferences for EHe are expressed by Ile and Phe (Ile or Phe in a beta-strand make helix-strand contacts). (b) AS5: Preferences of amino acid side-chains for contacts in 5 contact interface types, e.g. contact with a helix, contact with a sheet, or contact with water. The counts in these simpler "anything-structure" interface types are derived from those in the 29 interface types (Fig. 2(c)) used in (a), by summing over the secondary structure state S_1 of the central residue. So the 5 "contact-with" interface types are: H = residue contact of the 1st residue (in any secondary structure state) with a helix residue, E = contact with a β-sheet residue, T = contact with a turn residue, X = contact with a loop residue, W = contact with solvent. Examples of the resulting preferences (see the text for definition): Val and Ile have a clear preference to be in contact with β-sheet residues; Ala, for making contacts with helix residues; Lys, for making contact with water; and so on. It appears that these preferences are dominated by the secondary structure state of the central residue, simply because residues in a helix are likely to make contacts with other helix residues. The AS5 parameters are primarily used for comparison with other sets, not for production runs. (c) AM2: Preferences of amino acid side-chains for contacts in 2 contact interface types, i.e. contact with protein atoms (Prot) or contact with water molecules (Wat). These very simple protein-water interface types are derived from the AInt29 (Fig. 2(c)) parameters used in (a), by summing protein-protein contacts over the secondary structure states of both participating residues or from the AS5 parameters used in (b), by summing over the secondary structure state S_2 of the 2nd residue. The only remaining distinction is that between contacts of a protein atom with other protein atoms or with solvent atoms. These parameters resemble a hydrophobicity scale, e.g. Lys has the strongest preference for water contacts while Ile, Phe, Trp and Cys have the strongest preference for contacts with protein atoms. The apparently weaker contrast in line Protein is a numerical effect, due to the fact that, in the database used, the total number of contacts with protein atoms exceeds by far that with water atoms. (d). S4: Preferences of amino acid side-chains for residue occurrence in 4 secondary structure states, given for comparison. These are the classical occurrence residue preferences used in secondary structure prediction methods (Chou & Fasman, 1978; Garnier *et al.*, 1978; Maxfield & Scheraga, 1979). There is significant correlation of these parameters with those in (b). The AS5 parameters in (b), however, include water contacts in a natural and consistent fashion. e. AM2 preference parameters, numerical values. The parameters are calculated as logarithms of ratios of probabilities and are thus additive along a-chain (see Methods).

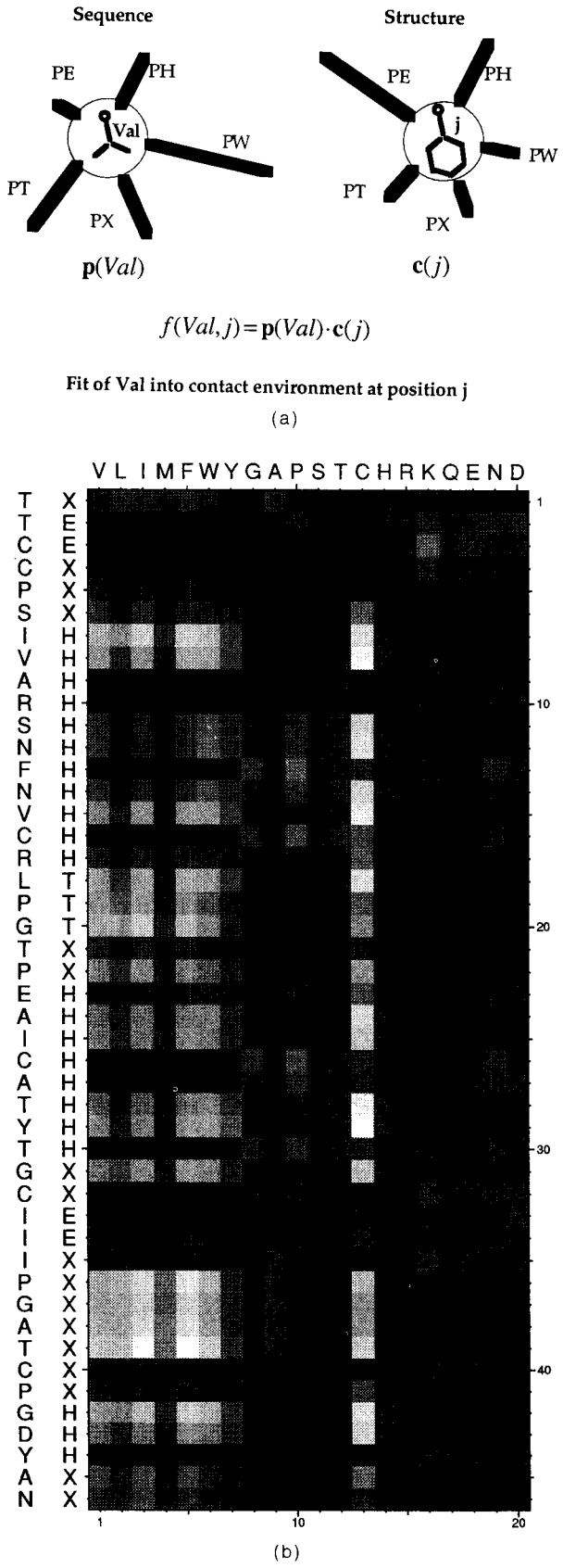


Figure 5. Sequence-structure fitness. (a) Calculation of sequence-structure fitness of residue $R(i)$ for the contact environment $c(j)$ in terms of the 5 contact interface types PX, PT, PE, PH and PW. The fitness quantifies the match between the interface vector at position j of the template structure and the preference vector for residue type R at position i of the input sequence. The numbers $f(R,j)$ constitute the structure-derived profile in (b). (b) Structure-derived profile for crambin, using 5-state (AS5) preference parameters. The amino acid type and secondary structure is given in the left margin, as in Fig. 2(d). At each position in the structure (rows $j = 1$–46), the fitness $f(R,j)$ of each amino acid type R (columns V to D) is represented by the grey level of the squares, on a scale of $-1{\cdot}0$ (white) to $+1{\cdot}0$ (black). The fitness of any amino acid sequence for this structural template is obtained by looking up the fitness of the particular amino acid (column) for a particular structural position (row) and summing these values over all positions in the sequence-structure alignment. For a given set of contact preference parameters, the sequence-structure profiles can be precalculated for all database proteins and reused in all subsequent sequence-structure alignments. Other preference parameters, such as secondary structure preferences, can also be represented in this way.

distribution of similarity values relative to zero and, as a result, the average length of alignments. Here, to establish a link with experience accumulated in the practice of sequence alignments, we scale the local similarity values $f(R,j)$ to new values $s(i,j)$ using a linear transformation. The transformation is such that the mean value of $f(R,j)$ minus 1 standard deviation (S.D.†) maps to a similarity $s_{min} = -0{\cdot}5$, while the mean value plus 1 standard deviation maps to a similarity $s_{max} = 1{\cdot}0$. The averages are taken over all positions j in the database of 64 proteins and over all residue types R. Their values are for AInt29: mean $= -0{\cdot}125$, S.D. $= 0{\cdot}490$; for AM2: mean $= -0{\cdot}032$, S.D. $= 0{\cdot}138$; for AS5: mean $= -0{\cdot}071$, S.D. $= 0{\cdot}262$; for S4: mean $= -0{\cdot}076$, S.D. $= 0{\cdot}486$. As a result, most of the transformed similarity values $s(i,j)$ lie between $-0{\cdot}5$ and $1{\cdot}0$, for all parameter sets. By adjusting the constant s_{min}, with fixed s_{max}, the average length of alignments can be adjusted. The other 2 parameters are penalties for gap opening and gap elongation. For the results reported here we chose a gap open penalty of $4{\cdot}0$ or $3{\cdot}0$ and gap elongation penalty of $0{\cdot}1$ per residue, so that typically a gap can be compensated for by 3 or 4 (or more) optimal residue fits.

(f) *Core weights*

In order to explore the importance of the protein core in fitting sequence to structure, we have introduced "core weights" in the alignment procedure. For this purpose we define the protein core as consisting of residues that are conserved in sequence and not exposed to solvent. Core weights are calculated as follows: from the multiple sequence alignment, sequence variability v is calculated as described by Sander & Schneider (1991) and linearly scaled to weights w, which for a protein family, as defined by a multiple sequence alignment, average to $1{\cdot}0$. So conserved residues have $w > 1{\cdot}0$, variable residues have $0{\cdot}1 < w < 1{\cdot}0$ (Sander & Schneider, unpublished results). These weights are only used for residues that have a relative solvent accessibility of less than 30%, relative to a fully extended chain, and a variability of $v < 25$; all other residues are given a weight of $w = 0{\cdot}5$. When evaluating the sequence-structure fit at positions i and j in the alignment, weighted similarities $w \times s(i,j)$ are used

† Abbreviations used: S.D., standard deviation; r.m.s., root-mean-square.

in place of $s(i,j)$, with an appropriate position-dependent rescaling of gap weights. In this way, more importance is given to core residues. Here, to introduce the concept, we explored the use of core weights in the most demanding test, i.e. when searching with one structure against a large database of sequences (see below). The precise form of the core weights has not been extensively optimized.

(g) *Pitfalls: unrealistic gap penalties and missing jack-knife tests*

There are 2 classical pitfalls in this type of approach. First, in aligning a sequence against its own structure as a template it is important to set realistic values of gap parameters. If gap placement is prohibitively expensive, the native full-length alignment, which has no gaps, is heavily favored (Hendlich *et al.*, 1990). To avoid this pitfall we use realistic gap penalties, chosen such that for homologous proteins of known structure an approximately correct number of gaps is introduced. Second, it is important to remove from the database the input protein and its homologues before parameters are derived that are applied to the input protein. Otherwise, scores may be unrealistically high for some test proteins and then disappointingly low for completely new test proteins. The larger the number of parameters derived from the database, the larger the danger of reproducing protein-specific structural information in the parameters while losing predictive power. On occasion, it is difficult to tell from published work whether this effect played a role. Bowie *et al.* (1991), for example, used parameters derived from 16 proteins including myoglobin (1mbo) and haemoglobin alpha-chain (3hhb) and then tested the ability of a profile derived from the structure of sperm whale myoglobin to detect globins in the sequence database. Jack-knife testing is less important when the number of parameters used in the scoring tables or potentials of mean force is smaller.

(h) *Removing sequence information from contact profiles*

To determine the information content in structure-derived profiles, we remove direct sequence information from the search profiles in all tests. The reason for this is methodological. If the size of residues or their chemical character is used in defining the structural states in the template protein, then any sequence-structure alignment contains elements of sequence-sequence alignment and it becomes very difficult to assess genuine improvement relative to pure sequence alignment methods. Here, we have removed any direct information about the size of a residue by normalizing the contact interface vectors $\boldsymbol{c}(j)$ attached to structural position j. So all tests reported here can be used to evaluate the contribution of structural information to the success of sequence-structure alignment. This is in contrast to Bowie *et al.* (1991), who use the total accessible surface area of a residue as 1 of 2 criteria for representing the structural state of a residue (larger residues, on average, have larger accessible surface areas). After assessing the contribution from structural information alone, one can then determine the proportion of sequence information to be combined directly or indirectly with structure-derived information in order to optimize the performance of sequence-structure searches. This is left to future work.

3. Results

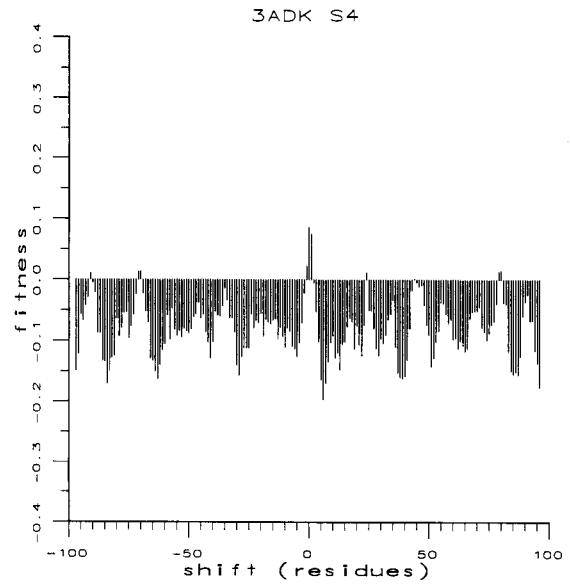
To test the power of the method, proteins of known structure are represented by their contact profiles and the fitness of different sequences for these structures is evaluated. Different types of tests are performed, in order of increasing difficulty: shift self test, align self test, search for structures, search for folding units and search for sequences. In each case and in all tests the input protein is explicitly removed from the database of 64 non-homologous proteins (Fig. 3) and all preference parameters are recalculated (jack-knife test, see Methods).

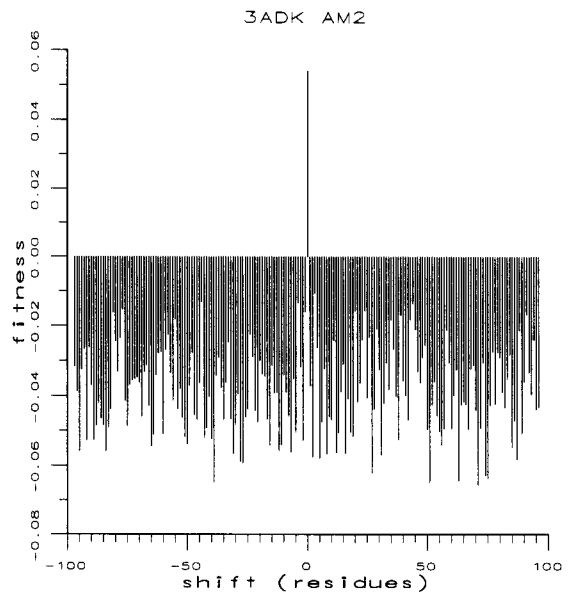
(a) *Shift self test: can the correct start position of a sequence in its own structure be identified?*

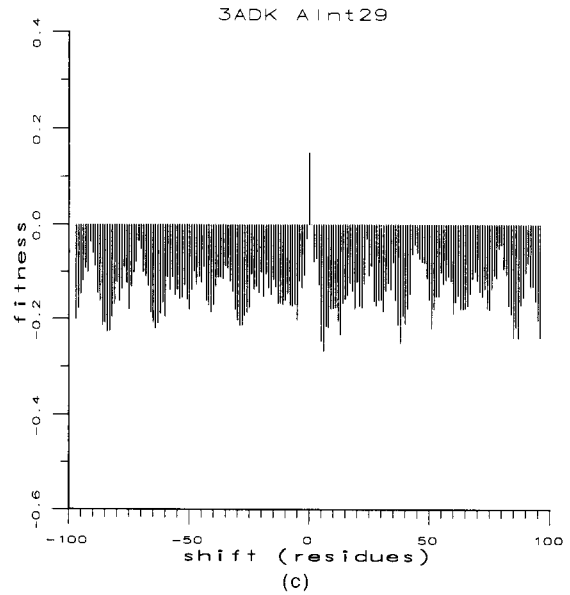
In this test, the sequence of a protein, without insertions or deletions, is started at all possible different offsets relative to its own native structure, with residues that would extend beyond the end of the structure wrapped to the beginning in cyclic fashion. For each such arrangement, the sequence-structure fitness is evaluated (see Methods). Which set of parameter produces a clear maximum at the origin, i.e. for the native sequence-structure arrangement?

The result is astonishingly clear cut. The fitness of the native arrangement is clearly superior to any other shifted arrangement, with by far the highest peak at the origin (zero shift, Fig. 6(a) to (d)). This result holds for all 64 proteins tested here and for all sets of contact preference parameters, with only a single exception: α-amylase inhibitor 1hoe, using AM2 and AS5 contact parameters. Typically, the peak at the origin is 4 to 7 standard deviations above background, defined as all shifted arrangements in the same protein. The second highest peak is typically 2 to 3 standard deviations above background. When the same test is performed with classical secondary structure preference parameters (S4), (Fig. 6a), analogous to those of Chou & Fasman (1978), Garnier *et al.* (1978), or Maxfield & Scheraga (1979), the number of proteins with a highest peak at the origin drops to 46 out of 64 proteins (although 60 out of 64 have the maximum within 1 residue of the origin) and the resolution is inferior, i.e. the signal (standard deviation above background) of the strongest peak relative to the second best peak is not as strong. We conclude that contact preferences, even in the simplest two-state description (AM2), outperform classical (S4) secondary structure preferences, primarily because they include the effect of solvent. Apparently, hydrophobic inside/outside preferences (protein-protein contacts *versus* protein-water contacts) carry more information than secondary structure preferences (α, β, turn or loop).

Co-operativity appears to be the reason why the shift self test is so selective and the native optimum so sharp. In typical globular proteins, with secondary structure segments and loops of varying lengths, only the native arrangement has correct phasing in which all single residue preferences on average add up to a positive number. Shifts of only one residue can switch many side-chains from







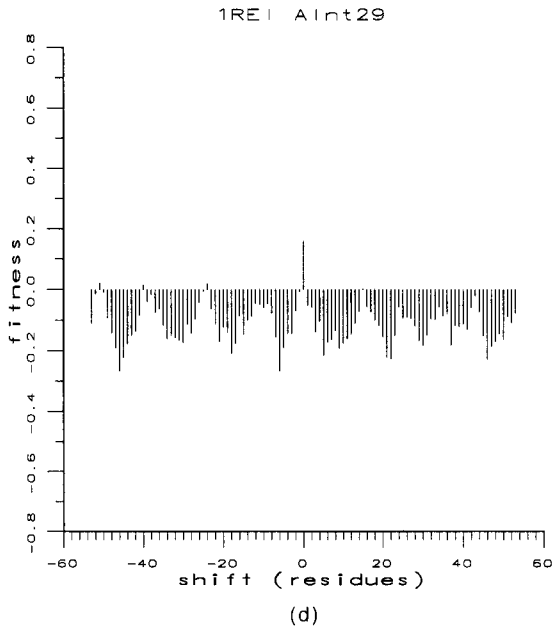


Figure 6. Shift test: detection of the correct alignment of a sequence in its own structure, allowing no gaps. In this shift test, the sequence of a protein is shifted along its native structure, in cyclical fashion. Residues that leave at the C terminus re-enter at the N terminus. For a given shift (in units of number of residues), preference parameters are used to evaluate how well the shifted sequence is adapted to the structure. A shift of zero residues corresponds to the native structure. In all cases, the native arrangement has the highest preference, with a large positive peak, but interface parameters have a sharper maximum than classical secondary structure preference parameters. To 1st approximation, the different interface parameters (protein-water, etc.), work equally well in this shift test. (a) Adenylate kinase, a mixed α/β structure, evaluated with 4-state (S4) secondary structure occurrence parameters, i.e. not using the interface approach and not including the effect of water contacts. Note the ambiguity of ± 1 residue near the native arrangement (shift = 0); 194 residues, PDB code 3adk (Schulz *et al.*, 1974). (b) Adenylate kinase evaluated with 2-state (AM2) protein-water contact preference parameters. Note the improved performance relative to (a), in spite of the smaller number of interface states. (c) Adenylate kinase evaluated with 29-state (AInt29) contact preference parameters. (d) Immunoglobulin-like REI protein, a homodimeric antiparallel beta sandwich structure with 4 sheets, evaluated with AInt29 preference parameters; 107 residues per chain, PDB code 1rei. Weak periodicity (4-on-4 beta strands?) appears to be reflected in the sequence-structure fitness signal.

solvent contact to contact in the protein interior, and *vice versa*. Larger shifts put loop sequences onto secondary structure and *vice versa*. Shifts large enough to place one loop sequence onto the structure of a subsequent loop lead to mismatches elsewhere (unless the structure is periodic, e.g. Fig. 6(d)). So only in the native arrangement do many positive contributions add up co-operatively to give a dominant peak.

The results of the shift test indicate that evaluation of sequence-structure alignments can be a powerful filter for rejecting incorrect sequence placements in a structural framework and that surface/interior preferences of amino acids are the most important single factor. A similar conclusion was drawn from testing protein models using solvation preference parameters (Holm & Sander, 1992). Whether this can be exploited for structure prediction depends on the influence of alignment gaps. The effect of gaps is tested below.

(b) *Align self test: can the sequence be aligned correctly to its own structure?*

In a more demanding align self test, a sequence is aligned with its own structure, allowing gaps (but not cyclic wrapping), i.e. all sequentially ordered arrangements of a sequence in its own structure are evaluated. The alignment is performed using a dynamic programming algorithm. In the matrix of local similarities, the algorithm finds the globally best trace among a very large number of possible traces (Fig. 7). The local similarity is evaluated using the sequence-structure fitness profile (Fig. 5(b)). The best trace is the one with the highest cumulative similarity score.

The result is positive for almost all proteins in the database. The native alignment is identified as best, although there are many different reasonable ways of fitting sequence pieces into structure segments (off-diagonal traces in Fig. 7). The test is more severe for shorter sequences, as they can be fit into many of the larger proteins over their entire length, while longer sequences have few sufficiently long realistic templates in the database. The quality of the resulting alignment can be measured by the resulting effective sequence similarity, as measured, e.g. by the percentage of identical residues (see columns in Fig. 8). For example, the sequence identity of 76% for trypsin (4ptp) means at least one insertion or deletion causes a shift of 24% of the residues away from their correct position. The three different sets of contact parameters tested perform as follows. With 29-state parameters (AInt29) 63/64 proteins are 100% correctly aligned (except for trypsin with 24% mismatch); with five-state parameters (AS5), 63/64 are correct (except virus coat protein 2stv with 14% mismatch); with two-state parameters (AM2), 60/64 are correct (except 1gd1, 13%, 6ldh, 15%, 2tmn, 4%; 4xia, 20%). For classical four-state secondary structure parameters, which lack solvent contact information, about one third of the alignments contains errors of up to 42% mismatch.

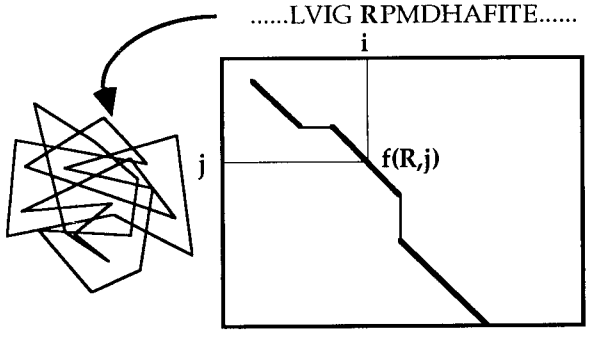
Aligning a sequence with its own structure is a non-trivial exercise when all direct sequence information has been removed and a residue position is only characterized by its (normalized) contact environment. However, the correct alignment of a sequence in its own structure does have a non-specific competitive advantage compared to shorter, off-diagonal, alignments. This occurs in any alignment procedure that aims to maximize the total alignment score, i.e. the sum of local similarity along the entire alignment trace (Smith & Waterman, 1981). Theoretically speaking, the main diagonal would be the one with the highest score, because of its length, even if the entire matrix of local similarities were uniform (but not if the matrix were noisy). So success in the align self test is a necessary, but not sufficient, condition for the overall success of any sequence-structure alignment method.

(c) *Sequence in search of structures: can native and homologous structures be identified in a database search?*

In this test, a given protein sequence is aligned to a dataset of 64 different structures, one at a time, presented as contact profiles. Can the native structure be identified, i.e. does it have the highest score? If not, what is the highest scoring structure and does it reflect an essentially correct structure for this sequence? For each sequence-structure pair, the five best alternative alignments (not sharing any subtrace) are evaluated. Thus, the native structure has to compete with more than 300 alternatives.

The result of this test is very encouraging (Fig. 8). For 58 out of 64 proteins the native structure scores highest, i.e. at position 1 out of 320, using the 29-state contact interface description (AInt29). The exceptions are: 1ctf (rank 2/320), 1hoe (rank 11), 5rxn (rank 2), 2sga (rank 2), 4sgb (rank 2) and 1utg (rank 12). The quality of the sequence-native structure alignments is excellent (100% sequence identity with the native alignment in most cases), in spite of the fact that insertions or deletions are allowed. Using AInt29 parameters, only 4ptp has a 26% alignment error.

This result is almost independent of the type of contact parameters used, i.e. parameters based on



(a)

Figure 7. (a) Self-alignment. The matrix of similarity values reflects the alignment of residue R at position i in the input sequence (horizontal axis) with position j in the template structure (vertical axis). The alignment algorithm finds the best path through the matrix such that the sum over the fitness values f is maximized.

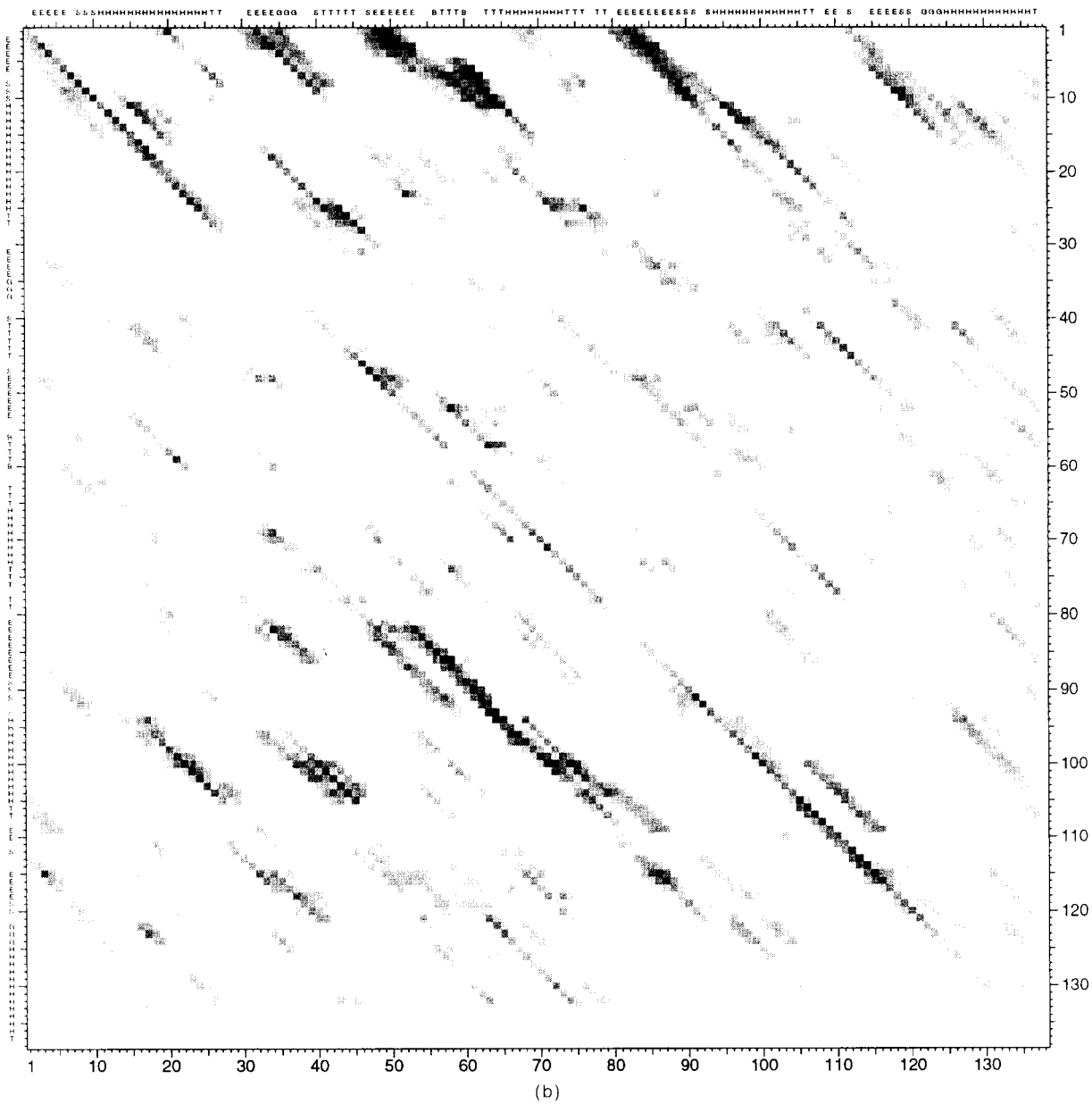


Figure 7. (b) Self-alignment. Alignment similarity matrix for flavodoxin, a mixed β-α structure of 138 residues (PDB data set 3fxn). Residue numbers and secondary structure are shown in margins. The fitness values are averaged over windows of length 12 (value centered at the beginning of the window), representing the fit of any sequence fragment of length 12 into any structure fragment of this length. The grey level indicates the sequence-structure fitness. Note on the one hand that the preferences are clear for certain segment matches (diagonal and off-diagonal traces), capable of distinguishing between segments that prefer an α-helix or β-strand, but that preferences are not sufficiently strong to locally indicate the correct alignment. Only the co-operative effect of summing along the long trace can lead to a high fitness score, as along the main diagonal. Note also that clear off-diagonal traces in some cases correctly indicate the preference of a sequence segment for several of the structurally similar β-α-β units in the structure, e.g. the N-terminal 20 to 25 residues.

29 (AInt29) and five (AS5) contact states perform equally well and those based on two states (AM2) almost as well. Remarkably, however, four-state secondary structure prediction parameters perform less well, as already seen in the cyclic shift tests: about one half of the native structures do not rank as number one (Fig. 8, column S4).

The practical aim of the one-sequence against many-structures search mode is the prediction of three-dimensional structure, given a sequence. It answers the following question: if the protein has one of the known folds, which is the correct one? The predicted fold is that of the top-scoring structure. If the top-scoring fold has a relatively low score, in absolute terms, one rejects the hypothesis that this sequence has a known fold. Work to determine an appropriate absolute criterion is in progress. This predictive procedure works very well when the interface contact profile is derived from the *native* structure: the native structure has a high

protein	self rank with AInt	% identity for AInt	self rank with AM	% identity for AM	self rank with AS	% identity for AS	self rank with SA	% identity for SA
351c	1	100	9	100	1	100	3	100
256b	1	100	1	100	1	100	9	100
8adh	1	100	1	100	1	100	-	
8atc	1	100	1	100	1	100	1	100
2aza	1	100	1	100	1	100	51	93
3b5c	1	100	1	100	1	100	-	
3blm	1	100	1	100	1	100	-	
2ca2	1	100	1	100	1	100	1	100
1ccr	1	100	1	100	1	100	4	100
2ccy	1	100	1	100	1	100	2	100
1cd4	1	100	1	100	1	100	1	100
3cla	1	100	1	100	1	100	4	100
5cpa	1	100	1	100	1	100	1	61
2cpp	1	100	1	100	1	100	1	80
4cpv	1	100	1	100	1	100	-	
1cse	1	100	1	100	1	100	1	76
1ctf	2	100	1	100	2	100	134	100
2cyp	1	100	1	100	1	100	-	
8dfr	1	100	1	100	1	100	1	100
1ecn	1	100	1	100	1	100	141	100
2er7	1	100	1	100	1	100	1	97
4fd1	1	100	1	100	1	100	1	100
4fxn	1	100	1	100	1	100	1	76
3gap	1	100	1	100	1	100	-	
2gbp	1	100	1	100	1	100	1	100
1gcr	1	100	1	100	1	100	3	88
1gd1	1	100	1	87	1	100	1	85
1gox	1	100	1	100	1	100	2	100
1gp1	1	100	2	100	1	100	2	100
2hla	1	100	1	100	1	100	1	100
1hoe	11	100	269	100	-		32	87
1il1	1	100	1	100	1	100	10	100
4icd	1	100	1	100	1	100	1	58
1il8	1	100	1	100	1	100	9	100
1il3	1	100	1	100	1	100	1	84
6ldh	1	100	1	85	1	100	1	90
2liv	1	100	1	100	1	100	1	78
2ltn	1	100	1	100	1	100	-	
1lz1	1	100	1	100	1	100	-	
1mbd	1	100	1	100	1	100	4	100
2mhr	1	100	1	100	1	100	131	100
2pab	1	100	1	100	1	100	36	100
1paz	1	100	1	100	1	100	1	64
4ptp	1	76	1	100	1	100	-	
1r69	1	100	1	100	1	100	2	100
1rhd	1	100	1	100	1	100	1	100
7rsa	1	100	1	100	1	100	1	100
2rsp	1	100	1	100	1	100	2	100
5rxn	2	100	1	100	1	100	80	85
2sga	2	100	1	100	1	100	-	
4sgb	2	100	39	100	2	100	28	100
2sns	1	100	1	100	1	100	-	
2sod	1	100	1	100	1	100	-	
2ssi	1	100	2	100	1	100	4	100
2stv	1	100	29	100	3	86	12	100
2tmn	1	100	1	96	1	100	-	
1tnf	1	100	1	100	1	100	3	100
2ts1	1	100	1	100	1	100	4	63
1ubq	1	100	1	100	1	100	-	
1utg	12	100	-		18	100	-	
2wrp	1	100	13	100	1	100	2	100
1wsy	1	100	1	100	1	100	1	100
4xia	1	100	1	80	1	100	-	
1ypi	1	100	1	100	1	100	1	100

Figure 8. Sequence in search of a structure: rank of the native structure. Results of an alignment search with an input sequence against all structures in the selected database of 64 proteins, using contact interface parameters and, for comparison, secondary structure parameters. For each pair comparison, the 5 best alignments are noted and the $5 \times 64 = 320$ scores are sorted. The native structure is then ranked in the sorted list (self rank) and the quality of the corresponding alignment is reported in terms of percentage identical residues. Rank 1 and sequence identity 100% implies that the sequence-structure search was highly successful, both in finding the native structure and in aligning the sequence to it. Contact parameter sets (AInt29, AS5, AM2) outperform pure secondary structure preferences (S4). A dash indicates that the sequence identity was below 50%, interpreted as a negative result. Alignment parameters: gap open = 4·0, gap elongation = 0·1 (units of sigma relative to 0 = 1·0), $s_{min} = -0{\cdot}5$, $s_{max} = 1{\cdot}0$, nbest = 5, maxdel = 10.

score and the alignment is correct (see previous section). For useful practical applications detection of the correct fold also needs to work for structures *remotely homologous* to the native fold, e.g. the pair immunoglobulin/CD4 receptor, globin/phycocyanin; this is not consistently achieved in the current implementation.

(d) *Sequence in search of structural folding units: can correct structural domains or folding units be identified in a database search?*

A more demanding test is the search for substructures, rather than entire structures. Can correct structural domains or folding units be identified even when a native-like structure is not among the structures searched? Down to what level of structural unit does the search procedure work? An example is given in Figure 9, where the sequence of D-galactose binding protein (2gbp) is scanned against a set of structures, including that of 2gbp. The sequence fits best into the structures of 2gbp and two known homologues (1abp and 2liv, with sequence relation to 2gbp below the twilight zone), as well as into the structures of 3icd and 3pgk. What is detected in the latter two structures is not the same overall structure, but the presence of $(\beta\alpha)_n$ units (supersecondary structure), embedded in different folds. In all cases the secondary structure agreement is remarkable. This type of local structure agreement is also apparent in many other test runs (data not shown), including runs with sequences of all-β and all-α proteins. The substructure units detected by the procedure may be folding units. In the future, one may be able to use these correctly detected units to build a predicted three-dimensional structure, even in cases when a native-like structure is not in the database.

(e) *Structure in search of sequences: problems in scanning large sequence databases*

The search method can be inverted by scanning with a single structure against a database of sequences. In this inverted search one attempts to ask the ambitious question of identifying all protein sequences that contain a structural domain like that of the search protein. Does the native sequence or that of a homologue appear at rank 1? We aligned the structural template of each of the 64 structures against 640 sequences: the sequences of the 64 search structures plus 576 sequences chosen randomly from Swiss-Prot (Bairoch & Boeckmann, 1991). Parameters used were the 29-state interface preferences (AInt29, normalized, gap open = 10·0, gap elongation = 0·3, $s_{min} = -0{\cdot}7$, gaps in secondary structure not allowed).

The native sequence is detected at rank 1 in 38/64 cases out of 640 (59%), at rank 1 or 2 in 47/64 cases (73%), and among the top 10 in 55/64 cases (86%). Alignments of a structure with its native sequence, irrespective of the rank, are perfect in almost all cases, in spite of gap parameters that for bad

```
1 - 100      ....:....1....:....2....:....3....:....4....:....5....:....6....:....7....:....8....:
2gbp         ADTRIGVTIYKYDDNFMSVVRKAIEQDAKAAPDVQLLMNDSQNDQSKQNDQIDVLLAKGVKALAINLVDPAAAGTVIEKARGQNV
               EEEEEEES TT HHHHHHHHHHHHHHTT TTEEEEEEE TT HHHHHHHHHHHHHTT SEEEE  SSGGGHHHHHHHHTTT

             ********************************************************************************
2gbp  212.20 ADTRIGVTIYKYDDNFMSVVRKAIEQDAKAAPDVQLLMNDSQNDQSKQNDQIDVLLAKGVKALAINLVDPAAAGTVIEKARGQNV
               EEEEEEES TT HHHHHHHHHHHHHHTT TTEEEEEEE TT HHHHHHHHHHHHHTT SEEEE  SSGGGHHHHHHHHTTT
                  *              *      *                *  *    ** * * * *   *   **        ****
1abp  159.71 ENLKLGFLVKQPEEPWFQTEWKFADKAGKDLGFEVIKIAV..PDGEKTLNAIDSLAASGAKGFVICTPDPKLGSAIVAKARGYDM
                   EEE   SSTTHHHHHHHHHHHHHSSS   EEE .. SHHHHHHHHHHHHT      B  S SS TTHHHHHHHHH
                 *           ** * *   *                  * *          *                    *
3icd  155.26 ENPIIPYIEgdVTPAMLKVVDAAVEKAYKGERKISWMEIYTGelPAETLDLIREYRVAIKGPLTT..PVGGGIRSLNVALRQELD
             SSBEEEEE HHHHHHHHHHHHHHHHHTTTS   EEEEE  THS HHHHHHHHHHSEEEE     .. SSS    HHHHHHHHTT
              *    *   * *          *          *    *   *                 * *     *              *
2liv  151.58 EDIKVAVVGaqYGDQEFTGAEQAVADINAkgNKLQIAKYDDACDPKQAVAVANKVVNDGIKYVIGHlyEDEGILMITPAATAPer
               EEEEEEE HHHHHHHHHHHHHHHHHHHHTTB EEEEEEE TT HHHHHHHHHHHHTT  EEEE  HHHTT EEEESS   GGS
              * *                 *          *    *                  ** * *   *     *       *
3pgk  149.39 KDKRVFIRVdiTSNQRIVAALPTIKYVLEHHPRYVVLASHLGrsLAPVAKELQSLLGKDVTFLNDCVgdGQKVKASKEDVQKFRH
             SS  EEEE   S  HHHHHHHHHHHHHHHH  S  EEE      SHHHHHHHHHHH S EE   SS  SSSS     HHHHHHHH
```

Figure 9. Sequence in search of substructures (folding units). Results of an optimal alignment search with the sequence of D-galactose binding protein (2gbp) against a database of 64 protein 3-dimensional interface profiles. Only residues 1 to 85 of 2gbp are shown (residue numbers, sequence, secondary structure, sequence identities as asterisks, gaps as dots, inserts bracketed by lower case letters). The 4 top-scoring proteins (over their entire length of more than 300 residues) were arabinose binding protein (1abp), isocitrate dehydrogenase (3icd), leucine-isoleucine-valine binding protein (2liv) and phosphoglycerate kinase (3pgk). Two of these have a similar overall topology (1abp, 2liv) and are correctly detected, with good (but not perfect) alignments. The fold of the other 2 (3icd, 3pgk) is of the same general type, $(\alpha\beta)_n$, but the topology of chain connections is different. Yet, over most of the aligned length, the secondary structure of these 2 also agrees well with that of the search sequence (of which the structure, of course, was not used). This indicates that prediction of substructures or folding units using sequence-structure fitness searches may be a practical proposition.

sequence-structure fits are, realistically, accompanied by several gaps of varying length per 100 residues. For the two-state preferences (AM2) the results are comparable (rank 1: 73%, rank 1 or 2: 83% , rank top 10: 84%). The overall results of this test appear inferior to those of the sequence in search of structure tests. However, most of the low-ranking cases actually have close homologues at top ranks, e.g. the all-helical folds of globin 1ecn, globin 1mbd, uteroglobin 2utg, myohemerythrin 4mhr, calcium-binding protein 4cpv all pick up all-helical myosin or tropomyosin at top ranks, i.e. correct identification at the segment level. Note that in the independent but conceptually similar structure in search of sequence test by Bowie *et al.* (1991), the results for only a handful of proteins are given without any indication of the quality of alignments, while here we test 64 different structural templates and evaluate alignment errors.

(f) *Improvement by use of core weights*

Evolutionary information can be used to improve the performance of sequence-structure alignment. This is done (see Methods) by placing higher weights on residues in the conserved core of a structure, as derived from multiple sequence alignments and from the solvent accessibility profile. The idea is to place less emphasis on residues that are very exposed to solvent or highly variable in sequence.

Core weights are particulary valuable in the detection of remote homologues. For example, a search with the sequence of malate dehydrogenase (4mdh) against a set of structure-derived contact profiles with core weights detects the structurally related lactate dehydrogenase (6ldh) at rank 1, in spite of the very low sequence similarity. The quality of the implied three-dimensional model is quite good as judged by direct comparison of the two known structures (Fig. 10). However, because of the above-average size of the dehydrogenases (more than 300 residues), the detection at rank 1 in this and similar cases is only moderately difficult when ranking is according to absolute score: shorter contact profiles are unlikely to lead to a comparably large score. Fortuitous length match of sequence and structure may also in part be responsible for the

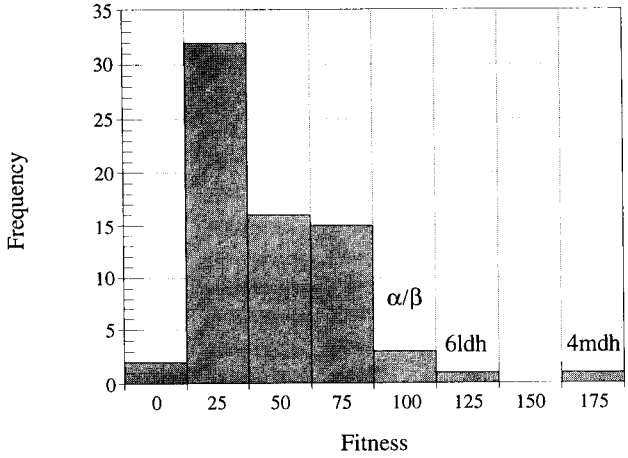
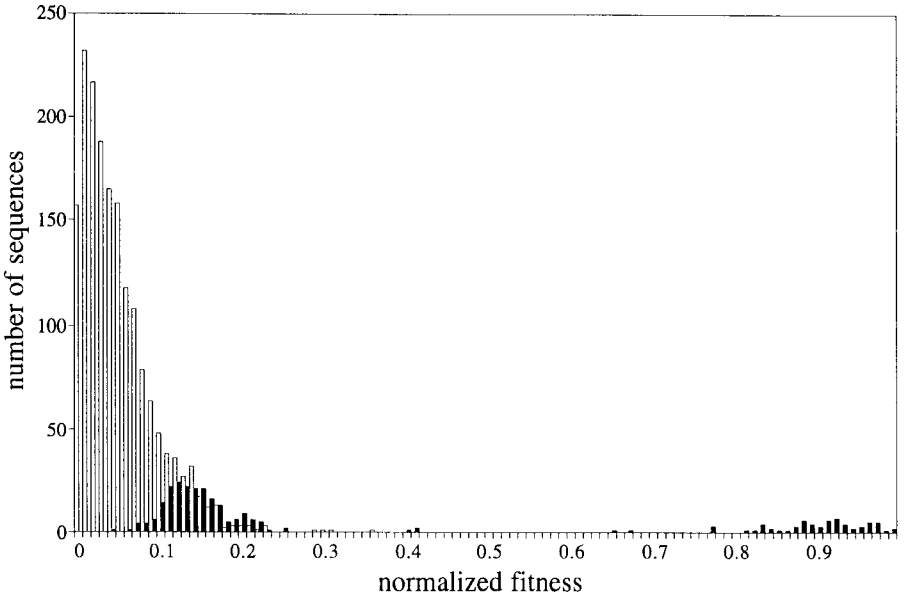


Figure 10. Application of core weights: search of remote homologues. The sequence of malate dehydrogenase (4mdh) detects the structural profile of the remote homologue lactate dehydrogenase (6ldh) at rank 2, right after the native structure, in competition with 68 alternative folds. The alignment is quite good: the sequence-structure alignment corresponds to a C^{α} positional r.m.s. deviation is 4·7 Å over 309 residues and a sequence identity of 15%, compared to the optimal 2·7 Å r.m.s. and 18% sequence identity derived from direct structure comparison (Holm *et al.*, 1992). Rank 3,4 and 5 are filled by other mixed α-β proteins. Contact interface parameters were AInt29; alignment parameters gap open = 10·0, gap elongation = 0·3, $s_{min} = -0{\cdot}7$.

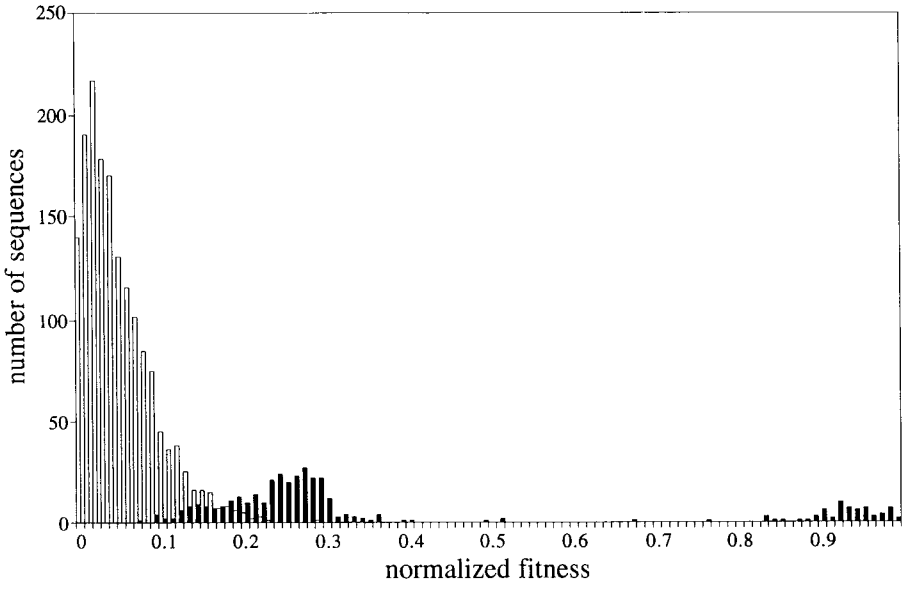
successful detection of remote homologues in similar searches by Bowie *et al.* (1991) and Jones *et al.* (1992).

Another example is the more demanding search with the myoglobin structure (1mbd) against the entire Swiss-Prot database of 25,000 proteins (Fig. 11). The overlap between the scores of globins and non-globins is significantly reduced when using core weights (Fig. 11 a,b). Moreover, the search picks up some remote homologues not detectable by sequence-sequence alignment. The discrimination between true positives, i.e. globins and the background, is fairly sharp as seen in a two-dimensional scatter plot in terms of sequence-structure fitness (length-normalized as described by Sander & Schneider, 1991) and the resulting sequence identity (Fig. 11(c)).

Sequences at the border of signal and background in this type of search have an implied predicted structure. In the myoglobin search, five sequences are tentatively identified to contain an all-helical domain (Fig. 11(c)). The src-related kinase is a false positive as judged by the known structure of a homologue, cAMP-dependent protein kinase (Knighton *et al.*, 1991). The origin of replication binding protein of herpes simplex virus (McGeoch *et al.*, 1988) is a likely true positive: the C-terminal end (residues 701 to 851) is predicted to be an all-helical

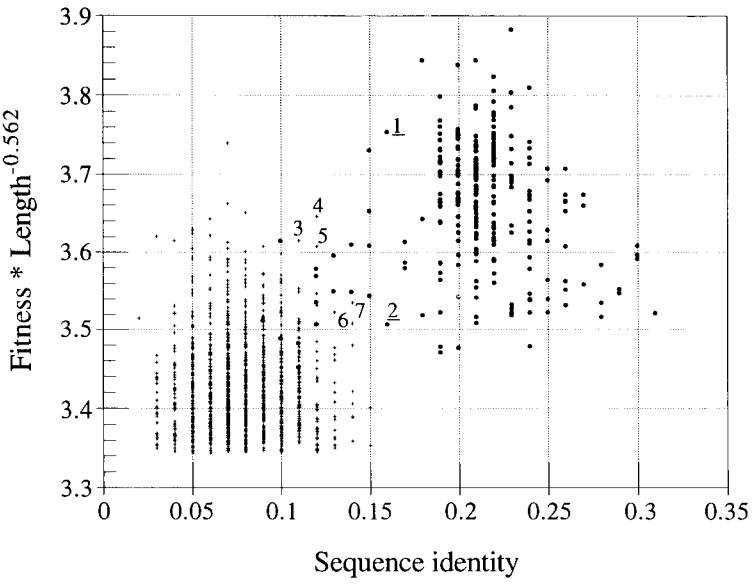


(a)



(b)

Fig. 11.



(c)

Figure 11. Application of core weights: structure in search of a sequence. Result of an alignment search with the contact interface profile of 1mbd against 25,000 Swiss-Prot sequences, (a) without and (b) and (c) with core weights. Sequence-structure fitness scores for the best 2000 hits are reported as a histogram. Separation of scores between globins (filled bars, myoglobins on the right, haemoglobins on the left) and non-globins (open bars) is superior using core weights. (a) All sequence positions carry equal weight; (b) conserved core positions have a higher weight in the alignment optimization (see methods); (c) same as (b), but with sequence identity resulting from sequence-structure alignment (as a fraction of 1·0) and length-normalized fitness values as dimensions, allowing a sharper distinction between background and noise than in (b). Non-globins with scores similar to globins either are false positives or are predicted to have the globin fold: syfb_ecoli (Phe-tRNA synthase beta-chain, 3), cyl4_neucr (sulfate permease, 4), kstk_hydat (src-related Tyr kinase, 5), ul09_hsv11 (origin of replication binding protein, 6), and bioa_bacsh (dapa aminotransferase, 7). Globins were labelled as such based on database protein names. Two remote homologues of myoglobin (16% identical residues in sequence-structure alignment) are cobra haemoglobin hbb2_najna (1) and plant globin hbpl_parad (2).Very simple fitness parameters were used (AM2), with gaps allowed in alignments (gap open = 6·0, gap elongation = 0·2). In (a) and (b) fitness is scaled to be 100 for the best reported alignment, 0 for the 2000th. Note that only the structure-derived profile and residue conservation weights, but not the sequence of 1mbd, were used in the search.

structure that possibly adopts the globin fold. The prediction is supported by additional independent evidence: a stretch of 85 residues has 29·4% sequence identity to a hemoglobin alpha-chain; and, a secondary structure prediction method, rated at 70·8% sustained three-state accuracy (Rost & Sander, 1993) predicts six helices. This prediction is open to falsification.

4. Discussion

(a) *Background: structure prediction by sequence-sequence alignment*

Prediction of protein structure by homology has been a practical proposition for some time, based on the detection of significant similarities at the sequence level. A number of methods have been developed to search for profile-sequence or template-sequence matches in a database of many sequences, using alignment algorithms (dynamic programming or block matching) with gaps allowed (Taylor, 1986; Bashford *et al.*, 1987; Gribskov *et al.*, 1987; Staden, 1988; Smith & Smith, 1990; Hennikof & Hennikof, 1991). When a significant (Sander & Schneider, 1991) match with a sequence of known structure is found in the database, then by implication the three-dimensional structure of the search protein can be modelled with reasonable accuracy. The new and simple idea here and in related approaches is to evaluate sequence-structure fitness directly, rather than indirectly *via* the comparison of sequences.

(b) *Background: measures of sequence-structure fitness*

Measures of sequence-structure fitness for entire proteins or protein domains, as opposed to single residues or secondary structure segments, have been developed over the last ten years by a number of groups, originally to distinguish between properly and improperly folded protein models. They are based on intramolecular potential energy (Novotny *et al.*, 1984, 1988), volume considerations (Gregoret & Cohen, 1990), contact counts (Bryant & Amzel, 1987), empirical solvation terms (Eisenberg & McLachlan, 1986; Baumann *et al.*, 1989; Chiche *et al.*, 1990, Holm & Sander, 1992), potentials of mean force derived from contact counts (Sippl, 1990; Hendlich *et al.*, 1990; Sippl & Weitckus, 1992), interresidue contact potential optimized to yield a global minimum for native structures (Crippen, 1991; Maiorov & Crippen, 1992), a self-consistent

hydrophobic molecular field (Finkelstein & Reva, 1991), environment-specific residue preferences (Luethy *et al.*, 1992) or alignment-based associative memory Hamiltonians (Goldstein *et al.*, 1992*a*,*b*). Effective contact energies or preferences for use in structure prediction were earlier derived by, e.g. Tanaka & Scheraga (1975), Crippen (1977), Lifson & Sander (1980), Galaktionov & Rodionov (1981), Miyazawa & Jernigan (1985), Scharf (1989). In related work, environment-specific substitution tables were derived from multiple alignment of sequences or structures (Overington *et al.*, 1990; Luethy *et al.*, 1991; Sander & Schneider, unpublished results). When summed over an entire protein, such substitution probabilities also provide a measure of sequence-structure fitness.

(c) *Different approaches to sequence-structure alignment*

Several groups have realized that sequence-structure fitness can be successfully used for structure prediction if the protein fold belongs to one of the known structural classes. The fold is identified among the many alternatives by techniques similar to those used in searching sequence databases with profiles derived from multiple sequence alignments alone (see above), or in combination with inspection of three-dimensional structures, as done for the globin fold by Bashford *et al.* (1987), and for the immunoglobulin fold by Taylor (1986). Simple profiles derived from structures alone were used by Bowie *et al.* (1990), who aligned solvent accessibility patterns (structure information) to hydrophobicity patterns (sequence information). Hendlich *et al.* (1990) used pair potentials to test various sequence arrangements of one sequence in a known structure. Luethy *et al.* (1991) combined secondary structure and solvent exposure to derive search profiles. Refinements were then made in the definition of structural states and in the definition of potentials (Bowie *et al.*, 1991; Casari & Sippl, 1992; Godzik & Skolnick, 1992; Godzik *et al.*, 1992; Goldstein *et al.*, 1992*b*). Geometrically generated hypothetical folds, rather than just known folds, were tested by Finkelstein & Reva (1991, 1992) using an iterative molecular field approach and by Taylor (1991). The optimization of fitness among various sequence-structure arrangements is usually treated as a one-dimensional alignment problem. The problem becomes two-dimensional when contact statistics for residue pairs are used without averaging over one of the contact partners. This more difficult problem has been solved by an algorithm analogous to one used in structure-structure alignment (Taylor & Orengo, 1989; Jones *et al.*, 1992).

(d) *Quality of alignments in the twilight zone of structural homology*

How do the results of these methods compare to straightforward detection of homology by sequence alignment? Sequence alignment works well down to a level of sequence similarity that corresponds to about 25 to 30% identical residues, for a length of 80 or more residues (Sander & Schneider, 1991). So the goal of new methods must be the detection of structural homologies below this level, in the twilight zone, e.g. the structural similarity between the two domains of rhodanese (sequence identity below 15%). None of the sequence-structure alignment methods has been proven to consistently and generally detect remote structural homologies with a high score and with the correct alignment. However, several interesting examples in addition to the ones given here have been reported, e.g. the relation between actin and heat shock protein hsp70 (Bowie *et al.*, 1991) or between phycocyanin and myoglobin (Jones *et al.*, 1992).

It would be interesting to perform a comprehensive test of these different methods using a large number of control examples, jack-knifing (removal of homologues of test proteins from the parameter database), predetermined gap penalties (rather than adjusting gap penalties to fit the known examples), and reporting not only the score of the best hits, but also the quality of the resulting alignments. Work along these lines is continuing in several groups. Exhaustive examples of sequence-remote homologues to known structures can be taken from the database of structurally aligned protein families (Holm *et al.*, 1992) available *via* anonymous file transfer protocol from ftp.embl-heidelberg.de in the directory /pub/databases/protein-extras/fssp.

(e) *Database searches: sequence seeks structure or structure seeks sequence?*

There is a not-so-subtle asymmetry in the direction of the search. Taking one sequence and scanning again, say, against 200 different protein structures, corresponds to testing 200 different three-dimensional structures for the input sequence, a small but reasonable exploration of the set of all possible structures for this sequence. Most groups working in this field have concentrated on this mode (sequence seeks structure), as a way of predicting three-dimensional structure for a given sequence, with the apparently best results reported by Jones *et al.* (1992). In the reverse direction (structure seeks sequence), one takes one input structure, e.g. a $(\beta\alpha)_8$ barrel, and scans a large database of 25,000 protein sequences, with the intent of identifying all known sequences compatible with the input structure (Bowie *et al.*, 1991). The sequence-structure fitness function needs to be well calibrated on an absolute scale for this to work, as the only physical competition in folding is between different structures for one sequence (only protein design or natural evolution explore different sequences for one structure!). It would be interesting to see reports of both search directions in future publication in this field.

(f) *Quest for the best parameter set*

The main conceptual difference of our approach to that of others is in the contact vector formulation

(Scharf, 1989). Rather than classifying a residue into one of several discrete environmental states (Bowie *et al.*, 1991), we retain for each residue information about its environment on a sliding scale, in terms of a set of real numbers that describe all interatomic contacts made by the residue. Evaluation of the preferences is done not by a simple table lookup, but by a vector product that effectively weights each participating interresidue contact by the corresponding contact strength.

This conceptual framework allows a large variety of different definitions of environmental preference parameters, or potentials of mean force, in a unified language, that of interatomic contacts. Solvent contacts are described in this language in a natural way. So far, the comparison of five different parameter sets has led to the conclusion that; (1) classical secondary structure preferences are inferior to contact interface preferences, (2) the best interface definition is the most complicated one, with 29 states, but only marginally so and (3) the simplest two-state preferences (protein-protein, protein-water) work almost as well as the most sophisticated 29-state description used here. Recent analysis of many-parameter mean field contact potentials confirms this view (Casari & Sippl, 1992). Future work will show which of the parameter sets currently being developed by several groups have the best performance in terms of identifying three-dimensional folds within and below the twilight zone of sequence similarity.

(g) *Anticipated improvements*

Improvement of sequence-structure alignments for protein structure prediction in the twilight zone may come from a more adequate definition of structural states (not necessarily more states), from use of homologous sequence information, e.g. in the form of conservation weights or multiple sequence alignments, from mixing sequence and structure information in correct proportions, from improved optimization methods in combining different interaction terms, from genuine two-dimensional alignment optimization that takes into account the change of contact partner with alignment, and from variation in backbone geometry and general plasticity of homologous structures. The problem is an urgent one, as genome projects will soon flood us with many protein sequences lacking functional or structural information. When developed and used properly, computational sequence analysis and structure prediction algorithms can save countless hours in the laboratory, by deriving probable three-dimensional structures.

We thank François Colonna-Cesari for his involvement in the initial definition of contact interface states and calculation of contact statistics; the members of the Protein Design Group at EMBL for constructive discussions. The work is part of the Ph.D. thesis of Christos Ouzounis. Reinhard Schneider provided the alignment algorithm, in the program MaxHom. The largest part of the work, codenamed FosFos (fit of sequence for structure), was done by Michael Scharf, part of it as a diploma thesis. Co-ordination was by Chris Sander. The order of authors is alphabetical and implies no priorities. Support from the Human Frontiers Science Program and the EC Bridge program is gratefully acknowledged.

References

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**, 2247–2250.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Baumann, G., Frömmel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329–334.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank, a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Indentification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, **7**, 257–264.

Bowie, J. U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Pept. Protein Res.* **29**, 46–52.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Chiche, L., Gregoret, L. M., Cohen, F. E. & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 3240–3243.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of protein from amino acid sequence. *Advan. Enzymol.* **47**, 45–148.

Colonna-Cesari, F. & Sander, C. (1990). Excluded volume approximation to protein-solvent interaction the solvent contact model. *Biophys. J.* **57**, 1103–1107.

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

Crippen, G. M. (1991). Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232–4237.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature (London)*, **319**, 199–203.

Finkelstein, A. V. & Reva, B. A. (1991). A search for the most stable folds of protein chains. *Nature (London)*, **351**, 497–499.

Finkelstein, A. V. & Reva, B. A. (1992). Search for the stable state of a short chain in a molecular field. *Protein Eng.* **5**, 617–624.

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. *Biophys.* **25**, 395–403. (Translation of *Biofizika*, **25**, 385–392 (1980).)

Garnier, J., Osguthorpe, D. & 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.

Godzik, A. & Skolnick, J. (1992). Sequence-structure matching in globular proteins: application to super-secondary and tertiary structure determination. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 12098–12102.

Godzik, A., Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the protein folding problem. *J. Mol. Biol.* **227**, 227–238.

Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992*a*). Optimal protein folding codes from spin-glass theory. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 4918–4922.

Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992*b*). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 9029–9033.

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

Gribskov, M., McLachlan, M. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Gribskov, M., Luethy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146–159.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & 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.

Hendrickson, W. A. & Teeter, M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)*, **290**, 107–113.

Hennikof, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565–6572.

Hobohm, U., Sander, C., Scharf, M. & Schneider, R. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698.

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

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

Knighton, D. R., Zheng, J. H., Ten-Eyck, L. F., Ashford, V. A., Xuong, N. H., Taylor, S. S. & Sowadski, J. M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, **253**, 407–414.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernet. Control Theor.* **10**, 707–710.

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

Lifson, S. & Sander, C. (1980). Specific recognition in the tertiary structure of beta-sheets of proteins. *J. Mol. Biol.* **139**, 627–639.

Luethy R., McLachlan, A. D. & 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.

Luethy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature (London)*, **356**, 83–85.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Maxfield, F. R. & Scheraga, H. A. (1979). Improvement in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry*, **18**, 697–704.

McGeoch, D. J., Dalrymple, M. A., Dolan, A., McNab, D., Perry, L. J., Taylor, P. & Challberg, M. D. (1988). Structures of herpes simplex virus type 1 genes required for replication of virus DNA. *J. Virol.* **62**, 444–453.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 535–552.

Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure prediction. *J. Mol. Biol.* **177**, 787–818.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, **4**, 19–30.

Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity, templates, key residues and structure prediction. *Proc. Roy. Soc. London*, **241**, 132–145.

Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein fold. *Protein Sci.* **1**, 216–226.

Rost, B. & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci.* **18**, 120–123.

Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. *J. Mol. Biol.* **212**, 403–428.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sander, C. & Vriend, G. (1991). Preference parameters: residues in contact interfaces. In *Protein Design on Computers. Biocomputing Technical Document*, vol. 6, pp. 202–215, Heidelberg.

Sander, C., Scharf, M. & Schneider, R. (1992). Design of protein structures. In *Protein Engineering, A Practical Approach*. (Rees, A. R., Sternberg, M. J. E. & Wetzel, R. eds), pp. 89–115, Oxford University Press, Oxford.

Scharf, M. (1989). Diploma thesis, Fakultät für Physik, University of Heidelberg.

Schulz, G. E., Elzinga, M., Marx, F. & Schirmer, R. H. (1974). Three-dimensional structure of adenylate kinase. *Nature (London)*, **250**, 120–123.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

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.

Smith R. & Smith T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 118–122.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Staden, R. (1988). Methods to define and locate patterns of motifs in sequences. *CABIOS*, **4**, 53–60.

Tanaka, S. & Scheraga, H. A. (1975). Model of protein folding inclusion of short-, medium-, and long-range interactions. *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3802–3806.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Taylor, W. R. (1991). Towards protein tertiary structure prediction using distance and motif constraints. *Protein Eng.* **4**, 853–870.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

Warme, P. K. & Morgan, R. S. (1978). A survey of amino acid side-chain interactions in 21 proteins. *J. Mol. Biol.* **118**, 289–304.

Edited by F. E. Cohen