Residue–residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds

MICHAEL A. RODIONOV^{1,2} AND MARK S. JOHNSON¹

¹ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom

²Institute of Bioorganic Chemistry, Belarus Academy of Sciences, Zhodinskaya str. 5/2, Minsk-141, Republic of Belarus 220141

(RECEIVED May 11, 1994; ACCEPTED August 17, 1994)

Abstract

We report the derivation of scores that are based on the analysis of residue–residue contact matrices from 443 3-dimensional structures aligned structurally as 96 families, which can be used to evaluate sequence–structure matches. Residue–residue contacts and the more than 3×10^6 amino acid substitutions that take place between pairs of these contacts at aligned positions within each family of structures have been tabulated and segregated according to the solvent accessibility of the residues involved. Contact maps within a family of structures are shown to be highly conserved (~75%) even when the sequence identity is approaching 10%. In a comparison involving a globin structure and the search of a sequence databank (>21,000 sequences), the contact probability scores are shown to provide a very powerful secondary screen for the top scoring sequence–structure matches, where between 69% and 84% of the unrelated matches are eliminated. The search of an aligned set of 2 globins against a sequence databank and the subsequent residue contact-based evaluation of matches locates all 618 globin sequences before the first non-globin match. From a single bacterial serine proteinase structure, the structural template approach coupled with residue–residue contact substitution data lead to the detection of the mammalian serine proteinase family among the top matches in the search of a sequence databank.

Keywords: globins; protein fold recognition; residue–contact substitution probabilities; serine proteinases; structural alignments

In comparison with the ab initio prediction of the 3-dimensional (3D) structure of a protein from its sequence alone, knowledge-based approaches attempt to relate an amino acid sequence to folds already defined by structural techniques such as NMR and X-ray crystallography. Knowledge-based modeling, for example, relies very heavily upon the data bank of known 3D structures (Bernstein et al., 1977; Abola et al., 1987) and the identification of related protein folds given 1 or more amino acid sequences (Blundell & Johnson, 1993; Johnson et al., 1994b).

In approaches introduced by Eisenberg and coworkers (Bowie et al., 1991; Lüthy et al., 1991) and our own group (Sali et al., 1990; Overington et al., 1990, 1992; Johnson et al., 1993, 1994a), 3D information is converted to a 1-dimensional (1D) description of the protein fold. A comparison then results from the matching of an amino acid sequence with these 1D representa-

tions of the 3D fold. In the former case, the fold is encoded as 18 sets of environmental propensities (Bowie et al., 1991); given the environment of each residue, these propensities are used to score amino acid matches from positions in the sequence of unknown fold. In the latter procedure, both the identity of the amino acid and its local structural environment are used to specify 1 of up to 766 sets of scores for matching of any amino acid at a site along the fold (Johnson et al., 1993, 1994a).

Phillips (1970) introduced the description of a protein fold in the form of a 2-dimensional matrix that characterizes the proximity of residues—local in 3D space, but not necessarily local in terms of their position along the sequence. Such contact matrices have subsequently been used to define elements of secondary structure, the accessibility or packing of residues, and the location of domains in protein structures (Ooi & Nishikawa, 1974; Galaktionov et al., 1975; Kuntz, 1975). The internal regularities of contact matrices have been investigated in proteins with known 3D structures, including the distribution of residue–coordination numbers, the estimation of residue–residue affinities, and residue–packing criteria (Rodionov & Galaktionov,

Reprint requests to: Mark S. Johnson at his present address: Centre for Biotechnology, P.O. Box 123, FIN-20521 Turku, Finland; e-mail: mark@hops.btk.utu.fi.

1992b; Rodionov et al., 1993). The application of residue-contact matrices as an object for the direct prediction of 3D structure was first formulated in works by Galaktionov and Rodionov (Galaktionov & Rodionov, 1980; Rodionov et al., 1981).

Several publications have described the incorporation of residue-residue contact data into alignment procedures. Ouzounis et al. (1993) have converted residue-residue contact matrices into contact-strength vectors and then applied the corresponding probability tables within dynamic programming alignment procedures. Johnson et al. (1994a) and Edwards et al. (1994) have reported their preliminary results at introducing information from residue-residue contacts within a set of computer programs designed to match structure with sequence. In particular, Edwards et al. (1994) have investigated side-chain to main-chain interactions and the substitutions that take place given a main-chain polar group (NH and/or CO), the number of contacts, and the identity of the amino acid side chains in contact with the main chain.

Sippl (1990) has introduced the concept of residue-residue potentials that are derived from the distribution of distances between pairs of amino acids in protein structures. These potentials were then used to evaluate the conformations of structure fragments with regard to a specific sequence. Jones et al. (1992) have used similar potentials within the Taylor and Orengo (1989) double dynamic programming procedure to find distant matches between sequence and a structure likely to share a common fold. Godzik et al. (1992) have compared a protein 3D structure and a sequence with the help of the structure's residue-residue contact matrix: random alignments between the structure and the sequence are generated using a Monte Carlo procedure, which are then evaluated in terms of 1-, 2-, and 3-residue interaction potentials.

In this study we derive a new set of amino acid substitution parameters and apply them to the discrimination of au-

thentic distant matches in the comparison between structure and sequence.

Results

Contact matrices as a measure of structural difference

For nearly identical 3D structures from a family, one would expect their individual contact matrices to be nearly identical (Table 1; Fig. 1). As both their sequences and their corresponding structures deviate from one another through the evolutionary process, there are some contacts that will change and the correspondence between contact maps will not be exact. Nevertheless, those contacts that are conserved in the entire family or a subset of the family can be useful in the recognition of the family as a whole. Thus, on the basis of a contact map constructed for 1 structure, or the average map derived from a number of aligned family members, we should be able to distinguish between other family members and any unrelated proteins.

There is only a minimal loss in the conservation of residue-residue contacts between structures within a protein family that are nevertheless very distant in terms of their sequence relationships. For example, 76% of the residue-residue contacts are conserved between 2 folds, the erythrocyruorin (Brookhaven code: 1ecd) and the α -chain of the hemoglobin of an Antarctic fish (1pbx), although the sequence identity upon structural alignment is only 11.8%. The tree shown in Figure 1, based on differences between pairs of contact maps (Table 1), reflects this behavior: thus, the myoglobins and the α -chains and β -chains of the hemoglobins cluster as individual groups that are themselves clustered together. The globins (lamprey, invertebrate, and plant globins) that are more distantly related to the members of these 3 clusters are also more different on the basis of their contact maps (Fig. 1).

Table 1. Loss of correspondence (percentage noncorrespondence) between residue-residue contact matrices in pairs of structurally aligned globins (upper triangle) and pairwise amino acid identities (lower triangle)^a

	1ecd	1lth	1lh1	1mba	1mbs	1pbx	1pmb	1sdh	2hbg	2hhba	2hhbb	2lhb	2mhba	2mhbb	2mm1	4mbn
1ecd		22.7	24.9	21.5	23.5	23.7	21.9	25.1	25.5	23.1	22.6	23.9	22.9	22.2	22.2	21.5
1lth	17.6		23.2	19.2	20.9	19.7	19.8	18.9	21.4	19.6	20.0	22.1	19.5	19.0	19.9	19.1
1lh1	15.4	17.7		21.7	21.5	22.5	19.9	24.7	21.9	22.1	22.7	23.9	22.4	21.7	20.5	19.3
1mba	22.0	16.3	21.2		21.3	21.2	19.4	19.6	20.9	21.2	21.3	19.9	21.3	20.6	19.6	18.4
1mbs	22.8	17.0	16.3	22.6		20.2	13.6	23.2	21.4	18.2	18.4	21.2	18.7	16.8	13.5	13.5
1pbx	11.8	14.2	18.3	15.5	22.5		17.1	21.2	17.7	9.3	14.2	21.6	8.9	14.6	17.5	17.0
1pmb	22.8	16.3	15.0	23.3	88.2	22.5		22.6	18.6	15.8	14.6	19.0	16.4	14.1	5.0	6.0
1sdh	13.2	16.3	13.0	21.2	21.2	19.0	22.5		22.9	20.3	22.5	22.0	20.9	20.9	22.4	21.8
2hbg	19.1	23.4	19.0	19.2	18.4	18.3	19.0	15.1		17.1	19.0	22.3	17.1	19.3	19.6	19.1
2hhba	16.9	12.0	16.3	19.1	28.4	49.6	27.0	18.4	19.8		13.3	20.7	6.7	13.8	16.2	16.4
2hhbb	19.1	18.4	17.1	17.1	26.7	42.2	26.7	21.2	21.9	44.7		19.9	14.5	6.9	15.7	14.8
2lhb	17.6	15.6	15.4	30.1	20.8	22.5	20.1	15.1	17.7	32.6	24.6		20.8	18.9	18.9	18.5
2mhba	19.8	14.2	16.3	18.4	25.5	51.8	27.0	18.4	20.6	87.2	44.0	31.2		14.1	16.6	16.6
2mhbb	17.6	19.1	18.5	17.8	25.3	44.4	26.0	19.9	23.3	44.0	83.6	23.3	45.4		14.6	13.7
2mm1	22.0	17.0	15.7	21.9	84.3	22.5	93.5	20.5	20.4	27.0	24.6	20.1	27.0	24.6		5.9
4mbn	23.5	14.9	17.0	24.0	83.0	22.5	86.3	20.5	19.7	27.0	24.6	22.1	26.2	24.6	84.3	

^a Brookhaven structure code designations: 1ecd, erythrocyruorin, midge; 1lth, innkeeper worm globin; 1lh1, leghemoglobin; 1mba, sea hare myoglobin; 1mbs, seal myoglobin; 1pbx, α -chain of the Antarctic fish *Pagothenia bernacchii*; 1pmb, porcine myoglobin; 1sdh, globin, *Scapharca inequivalvis*; 2hbg, globin, marine bloodworm; 2hhba, α -chain, human hemoglobin; 2hhbb, β -chain, human hemoglobin; 2lhb, lamprey globin; 2mhba, α -chain, horse hemoglobin; 2mhbb, β -chain, horse hemoglobin; 2mm1, human myoglobin; 4mbn, sperm whale myoglobin.

Fig. 1. Tree constructed (KITSCH; Felsenstein, 1985) on the basis of the pairwise contact-map differences reported in Table 1 for 16 globins after their structural alignment. Scale (percentage difference): the bottom horizontal branch leading to the erythrocyruorin, 1ecd, is 11.5%. See Table 1 for identification of the 16 globin 3D structures.

Contact maps from a databank of aligned 3D structures

Residue-residue contacts from individual structures and the substitutions that occur for these positions across an entire aligned family were determined for 96 families of 3D structures. These families (443 total structures) were structurally aligned (Overington et al., 1990, 1992, 1993). The family of globin structures was not used for the derivation of the contact substitution matrices.

From all pairs of proteins in 96 families and their individual contact matrices, we have tabulated the substitutions from all of the interacting residue pairs (see Table 2). To make the contact substitutions more sensitive to the local environment surrounding residue positions, the residue-pair replacements were segregated according to the corresponding solvent accessibility: where both residues are completely buried, both residues are exposed to solvent, or where one of the pair is buried and the other is not. (Inaccessible positions, defined as residues with

less than 7% solvent-accessible surface area, have been observed to form the common hydrophobic core of protein structure families [Hubbard & Blundell, 1987].) This results in $3\,400 \times 400$ matrices of residue-pair substitutions tabulating more than 3 million substitutions.

Residue-residue contact maps have been calculated from 3D coordinates of proteins in the Brookhaven Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987). Elements of the contact matrices, $a_{i,j}$, are set equal to "1" if residues i and j in the protein are in contact with each another (Table 2); otherwise, $a_{i,j}$ is set equal to "0." Two residues are said to be in contact when the C β atoms (C α , in the case of glycine) of these residues are less than the distance $r_{i,j}$:

$$r_{i,j} = r_i + r_j + 3(\sigma_i + \sigma_j). \quad (1)$$

The effective residue radii (see Table 3), r_i , for residue i and the corresponding standard deviations, σ_i , were obtained from the optimization of the correspondence between authentic X-ray structures and models reconstructed using distance geometry and contact maps derived from the structures (Rodionov & Galaktionov, 1992a).

Tabulation of residue-pair substitutions

A contact matrix that describes the residues that interact in each protein in the family forms the basis of our method. Residue-residue contact matrices for 2 aligned homologous fictitious proteins are depicted in Table 2. Thus, residue number 1 (serine) in the first protein is aligned with residue number 1 (alanine) in the second protein, and so on. The residue-pair substitutions recorded from these 2 proteins would consist of the serine-threonine contact in the first protein and the alanine-threonine match in the second, plus all other contacts at equivalent matrix positions in these aligned structures (Table 2).

An element of the substitution matrix, $S_{k,l \rightarrow x,y}$ (a contact between amino acids k and l in the first protein [p_1] substituted by a contact between residues x and y in the second protein [p_2]) is computed from the summation over all pairs of structures (Pr) within each of the families (F).

Table 2. Residue-residue contact maps for portions of 2 fictitious proteins (*A* and *B*) that have been aligned

	Protein A: Contacts								Protein B: Contacts							
	S _{1A}	T _{2A}	H _{3A}	Q _{4A}	V _{5A}	L _{6A}	K _{7A}	G _{8A}	A _{1B}	T _{2B}	H _{3B}	Q _{4B}	V _{5B}	F _{6B}	K _{7B}	G _{8B}
S _{1A}	—	1		1				1	A _{1B}	—	1	1				1
T _{2A}	1	—	1	1					T _{2B}	1	—	1	1			
H _{3A}		1	—	1	1				H _{3B}		1	—	1	1		
Q _{4A}	1	1	1	—	1	1			Q _{4B}	1	1	1	—	1	1	
V _{5A}			1	1	—	1	1		V _{5B}			1	1	—	1	
L _{6A}				1	1	—	1	1	F _{6B}			1	1	—	1	1
K _{7A}					1	1	—	1	K _{7B}				1	1	—	1
G _{8A}	1					1	1	—	G _{8B}	1				1	1	—

^a The pattern of contacts is the same for both proteins and serves to illustrate the tabulation of residue-residue pair substitutions. Matrix values equal to "1" represent residue contacts within the protein according to residue-residue contact distances using the values from Table 3 and Equation 1. Recorded substitutions on the basis of this matrix are S_{1A}-T_{2A} ↔ A_{1B}-T_{2B}, S_{1A}-Q_{4A} ↔ A_{1B}-Q_{4B}, S_{1A}-G_{8A} ↔ A_{1B}-G_{8B}, T_{2A}-H_{3A} ↔ T_{2B}-H_{3B}, ..., K_{7A}-G_{8A} ↔ K_{7B}-G_{8B}. Subscripts refer to the alignment position in protein A or B. The single-letter amino acid code is used: S, serine; T, threonine; H, histidine; Q, glutamine; V, valine; L, leucine; K, lysine; G, glycine; A, alanine; F, phenylalanine.

Table 3. Table 3. Amino acid side-chain radii r_i and standard deviations σ_i chosen for detection of contacts (adapted from Rodionov & Galaktionov, 1992a)^a

Residue type	r_i (Å)	σ_i (Å)
Ala	2.45	0.28
Cys	2.36	0.32
Asp	2.69	0.29
Glu	2.78	0.28
Phe	2.92	0.33
Gly	2.01	0.18
His	2.88	0.38
Lys	2.88	0.29
Ile	2.87	0.31
Leu	2.86	0.35
Met	2.87	0.36
Asn	2.76	0.35
Pro	2.76	0.35
Gln	2.80	0.28
Arg	2.89	0.28
Ser	2.45	0.28
Thr	2.67	0.28
Val	2.76	0.31
Trp	2.96	0.34
Tyr	2.90	0.39

^a Contact atom: C β ; C α for glycine.

$$S_{k,l \rightarrow x,y} = \sum_{l=1}^F \sum_{p_2=p_1}^{Pr} \sum_{p_1=1}^{Pr} \sum_{k,l} a_{k,l}^{p_2}. \quad (2)$$

From the 96 families (443 protein structures) and the 2,986 individual contact maps, we have tabulated a total of 3,299,027 residue-residue contact substitutions or approximately 2,500 data points for each set of residue-pair substitutions. Thus, it is possible that the statistics do not properly estimate some of the 400 data points under the 3 solvent-accessibility conditions, especially when they involve rare amino acids. As a consequence, our strategy is to compare distributions of residue-pair substitutions rather than individual values for residue-pair to residue-pair replacement.

Nonrandom character of the contact substitution distributions

If one assumes that the residue-pair substitutions derived from the family contact maps occur randomly, then they should depend only on the number of substituted contacts. The probability for substitution of residues k, l with x, y , in this case, can then be modeled by:

$$p'_{k,l \rightarrow x,y} = K \theta_{k,l} \theta_{x,y} \quad (3)$$

where $\theta_{k,l}$ is the fraction of contacts observed between amino acids of type k, l , and $\theta_{x,y}$ is the fraction of residue contacts of type x, y . The proportionality coefficient, K , is subject to the condition:

$$K \sum_{x,y} p'_{k,l \rightarrow x,y} = \sum_{x,y} S_{k,l \rightarrow x,y}. \quad (4)$$

We can now examine whether a difference exists between the chance distribution modeled by Equation 3 and the experimentally derived data, $S_{k,l \rightarrow x,y}$ (Equation 2), using a χ^2 statistic:

$$\chi^2_{k,l}(\nu - 1) = \frac{\sum_{x,y} (p'_{k,l \rightarrow x,y} - S_{k,l \rightarrow x,y})^2}{p'_{k,l \rightarrow x,y}}. \quad (5)$$

The distributions are different if $\chi^2_{k,l}(\nu = 1) \approx \chi^2_{k,l}/(\nu - 1)$ is greater than 1.0 (where ν is the number of degrees of freedom for χ^2 , here 400).

If we include within the χ^2 calculations those substitutions that involve interacting pairs of the same amino acids (e.g., Tyr-Leu $\rightarrow$ Tyr-Leu), then the values of $\chi^2(1)$ all exceed 4.0. Therefore, we can reject the hypothesis that the substitutions have occurred by chance. If we do not include these identical-pair substitutions, which can dominate the χ^2 , then the picture is more complicated (Table 4).

For residues exposed to solvent, we find that for about 15% of the residue replacements, values of $\chi^2(1)$ are less than 2.0 ($1\% \leq 1.0$) and mainly include contacts with the rarer amino acids such as methionine, histidine, cysteine + half-cysteine, and tryptophan (Table 4A). For residues with less than 7% solvent-accessible surface area (buried; Table 4B), nearly 19% (75 of 400) of the values are less than 2.0 (5 of 400 or $1\% \leq 1.0$). Residues in these contacts mainly involve methionine, proline, tryptophan, cysteine + half-cysteine, and glycine. This increases to about 44% ($6\% \leq 1.0$) of the contact distributions if one of the residues in the pair is exposed to solvent and the other is buried and vice versa (Table 4C).

Where the 2 residues in contact are either both buried or both accessible to solvent, values of $\chi^2(1)$ are symmetric (Table 4A,B). In contrast, Table 4C is asymmetric. For distributions where the initial residue of the pair is exposed to solvent (rows of Table 4C) and the second is solvent inaccessible (columns of Table 4C), then the less significant $\chi^2(1)$ values are found to involve exposed histidine, proline, tryptophan, methionine, glycine, cysteine + half-cysteine, and aspartate. Where the initial residue is buried (columns) and the second residue of a contact is exposed to solvent (rows), then contacts involving buried methionine, serine, threonine, phenylalanine, isoleucine, and asparagine are less significant.

From this analysis of the substitution distributions as reflected in Table 4, we find that the less significant $\chi^2(1)$ values primarily involve the rarer amino acids tryptophan, cysteine + half-cysteine, methionine, and histidine (both the solvent-exposed and solvent-inaccessible residues of a contact pair). However, because a large proportion of the data is greater than the chance expectation, we have applied the contact substitution data to the evaluation of structure-sequence matches.

The discrimination function

One measure of the similarity of 2 contact-substitution distributions, both of which express the ability of the pairs of contacts to be substituted, is the correlation coefficient:

Table 4. Nonrandom nature of the residue-pair substitution distributions^a

	A Ala	C Cys	D Asp	E Glu	F Phe	G Gly	H His	I Ile	K Lys	L Leu	M Met	N Asn	P Pro	Q Gln	R Arg	S Ser	T Thr	V Val	W Trp	Y Tyr
A. Residues in contact, <i>k</i> and <i>l</i>, are both exposed to solvent																				
A	15.0	4.7	12.7	11.4	2.9	7.6	1.5	2.8	8.6	4.6	1.5	3.6	6.9	3.8	2.8	7.6	5.3	4.5	2.6	2.9
C	4.7	1.0	4.0	6.4	3.1	3.7	1.9	10.4	7.6	10.6	2.2	6.4	4.3	5.0	6.0	8.4	6.9	12.6	1.7	3.6
D	12.7	4.0	7.0	7.1	3.0	11.6	1.9	4.1	14.5	7.0	1.9	6.5	7.3	2.7	4.6	12.9	6.5	4.4	2.7	4.0
E	11.4	6.4	7.1	14.4	2.9	8.7	1.6	3.4	21.0	7.1	2.3	4.3	4.4	4.4	5.3	5.6	8.0	6.7	2.0	3.2
F	2.9	3.1	3.0	2.9	4.7	2.4	1.8	3.4	3.1	3.3	2.5	2.4	2.6	2.3	3.7	3.4	2.6	3.1	3.1	3.0
G	7.6	3.7	11.6	8.7	2.4	5.5	2.4	4.6	11.0	4.3	2.6	8.4	6.0	4.0	4.4	11.6	8.8	6.9	1.8	5.3
H	1.5	1.9	1.9	1.6	1.8	2.4	1.0	2.0	1.6	2.4	1.1	1.4	1.6	1.1	1.2	1.5	1.9	1.5	2.0	2.1
I	2.8	10.4	4.1	3.4	3.4	4.6	2.0	3.3	4.0	4.1	1.1	2.6	4.0	1.8	3.6	2.7	2.9	3.6	3.2	2.5
K	8.6	7.6	14.5	21.0	3.1	11.0	1.6	4.0	16.8	8.4	2.8	5.1	9.2	5.5	3.9	9.3	7.9	7.0	2.9	4.8
L	4.6	10.6	7.0	7.1	3.3	4.3	2.4	4.1	8.4	8.3	2.0	4.6	6.7	4.7	3.8	4.6	5.6	5.6	5.1	3.5
M	1.5	2.2	1.9	2.3	2.5	2.6	1.1	1.1	2.8	2.0	0.8	1.0	1.6	1.3	1.5	1.2	1.7	1.7	1.6	1.4
N	3.6	6.4	6.5	4.3	2.4	8.4	1.4	2.6	5.1	4.6	1.0	4.8	3.1	2.2	2.4	4.7	5.1	4.4	2.6	3.7
P	6.9	4.3	7.3	4.4	2.6	6.0	1.6	4.0	9.2	6.7	1.6	3.1	5.3	3.3	2.4	4.6	3.7	4.3	2.0	4.5
Q	3.8	5.0	2.7	4.4	2.3	4.0	1.1	1.8	5.5	4.7	1.3	2.2	3.3	2.0	2.6	4.2	3.7	2.6	2.4	3.0
R	2.8	6.0	4.6	5.3	3.7	4.4	1.2	3.6	3.9	3.8	1.5	2.4	2.4	2.6	1.9	3.9	2.8	4.1	2.6	3.6
S	7.6	8.4	12.9	5.6	3.4	11.6	1.5	2.7	9.3	4.6	1.2	4.7	4.6	4.2	3.9	16.6	8.6	4.0	2.3	4.1
T	5.3	6.9	6.5	8.0	2.6	8.8	1.9	2.9	7.9	5.6	1.7	5.1	3.7	3.7	2.8	8.6	7.4	6.3	3.1	2.9
V	4.5	12.6	4.4	6.7	3.1	6.9	1.5	3.6	7.0	5.6	1.7	4.4	4.3	2.6	4.1	4.0	6.3	4.4	5.0	3.6
W	2.6	1.7	2.7	2.0	3.1	1.8	2.0	3.2	2.9	5.1	1.6	2.6	2.0	2.4	2.6	2.3	3.1	5.0	1.1	3.9
Y	2.9	3.6	4.0	3.2	3.0	5.3	2.1	2.5	4.8	3.5	1.4	3.7	4.5	3.0	3.6	4.1	2.9	3.6	3.9	3.3
B. Residues in contact, <i>k</i> and <i>l</i>, are both inaccessible to solvent																				
A	4.2	2.6	3.1	2.3	3.4	4.8	3.6	4.3	3.7	4.4	1.7	3.0	4.3	2.6	5.8	2.6	3.1	4.8	3.9	2.8
C	2.6	1.2	1.4	10.6	2.7	1.1	1.4	4.5	2.3	6.8	4.5	1.8	2.0	13.8	1.1	2.4	2.5	5.7	1.1	4.1
D	3.1	1.4	4.3	2.5	5.6	1.7	1.9	7.4	2.4	8.6	3.9	2.8	1.2	2.3	2.5	3.3	4.8	8.1	0.7	4.6
E	2.3	10.6	2.5	2.0	2.5	4.1	1.4	9.0	2.0	6.1	2.1	5.2	1.2	1.7	1.1	2.3	2.4	9.2	0.8	5.9
F	3.4	2.7	5.6	2.5	3.9	2.7	7.3	4.0	3.8	3.8	1.7	3.5	5.0	1.7	4.4	3.1	3.0	3.8	4.4	3.0
G	4.8	1.1	1.7	4.1	2.7	3.4	2.0	6.8	1.2	5.7	2.4	2.0	2.1	5.3	2.6	5.3	4.0	6.6	1.9	2.1
H	3.6	1.4	1.9	1.4	7.3	2.0	1.7	8.2	2.5	4.4	3.2	2.0	1.3	2.4	3.5	4.8	2.5	6.5	2.1	2.5
I	4.3	4.5	7.4	9.0	4.0	6.8	8.2	4.3	8.5	4.9	1.5	5.8	6.6	4.1	2.1	3.5	3.8	5.0	5.1	4.7
K	3.7	2.3	2.4	2.0	3.8	1.2	2.5	8.5	13.8	7.3	4.0	4.0	1.2	3.4	3.4	3.2	6.6	7.2	1.5	2.9
L	4.4	6.8	8.6	6.1	3.8	5.7	4.4	4.9	7.3	5.2	1.9	2.9	5.5	5.0	3.6	3.5	3.1	5.4	7.2	3.8
M	1.7	4.5	3.9	2.1	1.7	2.4	3.2	1.5	4.0	1.9	1.0	2.4	2.5	3.9	3.2	2.1	1.7	1.6	2.6	1.8
N	3.0	1.8	2.8	5.2	3.5	2.0	2.0	5.8	4.0	2.9	2.4	2.1	1.1	1.2	3.7	3.6	4.0	4.2	1.6	2.2
P	4.3	2.0	1.2	1.2	5.0	2.1	1.3	6.6	1.2	5.5	2.5	1.1	1.0	1.7	1.5	4.4	3.8	8.5	1.6	4.1
Q	2.6	13.8	2.3	1.7	1.7	5.3	2.4	4.1	3.4	5.0	3.9	1.2	1.7	1.4	3.1	1.8	4.1	2.8	2.8	2.7
R	5.8	1.1	2.5	1.1	4.4	2.6	3.5	2.1	3.4	3.6	3.2	3.7	1.5	3.1	2.2	3.6	2.9	5.7	3.8	6.5
S	2.6	2.4	3.3	2.3	3.1	5.3	4.8	3.5	3.2	3.5	2.1	3.6	4.4	1.8	3.6	2.9	2.5	3.1	4.0	2.3
T	3.1	2.5	4.8	2.4	3.0	4.0	2.5	3.8	6.6	3.1	1.7	4.0	3.8	4.1	2.9	2.5	2.7	3.9	1.6	3.5
V	4.8	5.7	8.1	9.2	3.8	6.6	6.5	5.0	7.2	5.4	1.6	4.2	8.5	2.8	5.7	3.1	3.9	6.4	5.0	5.4
W	3.9	1.1	0.7	0.8	4.4	1.9	2.1	5.1	1.5	7.2	2.6	1.6	1.6	2.8	3.8	4.0	1.6	5.0	0.8	3.7
Y	2.8	4.1	4.6	5.9	3.0	2.1	2.5	4.7	2.9	3.8	1.8	2.2	4.1	2.7	6.5	2.3	3.5	5.4	3.7	2.1
C. Residue <i>k</i> (rows) is exposed to solvent and residue <i>l</i> (columns) is solvent inaccessible^b																				
A	2.2	3.0	2.9	3.4	1.6	2.7	3.6	1.6	4.3	2.0	1.1	1.6	3.0	2.8	2.0	1.2	1.4	1.6	3.1	2.1
C	2.3	0.6	1.7	1.1	3.6	1.7	0.5	2.0	2.6	2.3	1.4	0.2	0.9	0.9	1.0	4.9	3.8	3.6	0.5	2.8
D	2.4	2.3	2.1	1.7	2.6	1.8	1.4	1.8	1.9	2.5	0.9	1.5	2.5	2.4	2.1	1.4	1.6	2.7	1.7	2.4
E	2.0	2.5	2.7	4.2	2.1	2.1	4.4	2.3	1.4	3.1	1.2	1.4	2.8	1.7	2.6	1.3	1.6	2.3	2.2	2.1
F	1.4	6.4	4.6	3.4	1.4	3.8	1.2	1.7	2.6	1.8	1.4	2.7	2.0	1.5	2.6	1.3	1.7	1.7	3.9	2.6
G	2.8	1.1	1.1	1.9	2.6	1.6	1.3	2.5	1.4	3.3	1.3	0.9	1.0	1.3	1.7	3.1	2.2	3.4	1.2	1.9
H	1.6	0.9	1.3	1.4	1.3	1.6	0.9	1.3	2.9	1.2	0.8	1.4	1.9	1.8	2.2	1.4	1.3	1.3	1.8	1.9
I	2.1	4.2	2.3	2.0	2.1	2.5	2.4	1.8	3.2	2.9	0.9	2.3	3.7	2.3	2.1	1.6	1.6	1.7	4.4	2.5
K	2.6	3.1	3.3	2.1	2.9	2.8	4.1	1.9	3.2	3.7	1.4	2.7	3.6	1.9	2.3	1.4	1.6	1.7	4.4	2.5
L	2.7	5.1	2.3	2.3	2.5	2.3	4.1	2.5	4.3	2.9	1.5	2.6	4.5	1.5	4.9	1.3	1.8	3.1	4.7	2.3
M	1.0	2.0	2.4	3.5	1.0	1.6	3.0	1.0	6.6	0.8	1.1	1.6	2.2	1.2	4.4	0.8	1.3	0.9	1.5	1.3
N	1.3	2.9	2.7	2.5	1.9	2.1	2.2	1.5	3.3	1.7	1.1	2.2	2.2	1.4	3.1	1.6	1.3	1.5	2.5	1.9
P	1.8	1.4	1.4	2.2	1.9	1.9	1.8	1.9	1.6	2.0	1.3	2.1	1.8	1.7	1.2	1.7	1.6	3.2	1.9	1.6
Q	1.4	3.1	3.8	3.4	1.6	3.0	4.7	1.4	1.7	2.1	1.2	1.8	1.8	2.4	2.8	1.1	1.6	1.8	2.5	1.5
R	2.0	2.5	3.6	1.7	1.8	2.1	1.2	1.9	3.6	2.1	1.5	1.8	2.2	1.4	2.0	1.8	1.5	2.1	2.3	2.1
S	1.6	4.1	2.3	5.7	1.7	3.2	3.9	2.0	7.1	2.8	1.1	2.8	2.5	4.6	3.2	1.4	1.4	1.7	4.1	2.9
T	2.1	5.3	4.1	4.6	2.2	3.6	4.8	2.1	4.9	2.7	0.9	2.4	3.5	3.5	4.0	1.7	1.9	2.5	3.8	3.6
V	2.2	4.2	3.7	2.6	2.1	3.5	3.9	2.5	2.4	2.9	0.8	2.6	4.4	2.1	4.6	1.6	1.6	2.7	3.2	2.5
W	1.8	2.0	2.5	2.2	1.1	2.4	1.5	2.0	1.3	1.6	0.8	1.1	1.6	1.0	0.8	1.1	1.1	1.8	1.4	1.2
Y	1.8	5.2	3.5	4.3	1.8	3.3	3.0	2.1	1.4	2.6	1.2	2.4	4.1	2.8	4.9	1.8	2.1	2.3	5.3	3.8

^a $\chi^2_{k,l \rightarrow x,y}$, where, for example, the χ^2 for residue *k* (rows) = Ala and residue *l* (columns) = Asp to all residue contact pairs *x, y* is 12.7 (see Equation 5).

^b This matrix is not symmetrical. For $\chi^2_{k,l \rightarrow x,y}$ and where residue *k* is solvent accessible and *l* is solvent inaccessible, *k* refers to rows and *l* to columns. Where residue *k* is solvent inaccessible, *k* refers to the columns and *l* to rows.

$$q_{k,l \rightarrow m,n} = \frac{\sum_{k,l} (S_{k,l \rightarrow x,y} - \bar{S}_{k,l})(S_{m,n \rightarrow x,y} - \bar{S}_{m,n})}{\sigma_{k,l} \sigma_{m,n}} \quad (6)$$

Values of q will be close to 1 if the k, l and m, n types of contacts have a similar pattern of substitution to the 400 possible amino acid pairings, and 0 or negative if substitution patterns are different. For example, the correlation coefficient computed between the distribution for Gly-Leu (to all other pairs of contacts $x-y$) and the distribution for Pro-Val (to all other pairs of contacts $x-y$) can be a measure of the potential for substitution of Gly-Leu for Pro-Val or Pro-Val for Gly-Leu.

The correlation coefficient can give us a proper statistical estimate of these values, but they do not express the contact substitution probabilities directly. However, we can combine the experimental contact substitution probabilities, $p_{k,l \rightarrow x,y}$, normalized to 1.0,

$$p_{k,l \rightarrow x,y} = \frac{S_{k,l \rightarrow x,y}}{\sum_{x,y} S_{k,l \rightarrow x,y}} \quad (7)$$

with their estimates obtained from the correlation coefficients, $q_{k,l \rightarrow x,y}$, in the form of:

$$w_{k,l \rightarrow x,y} = 1/2(p_{k,l \rightarrow x,y} + q_{k,l \rightarrow x,y}) \quad (8)$$

and use $w_{k,l \rightarrow x,y}$ to estimate contact substitution probabilities for a contact of type k, l replaced by a contact of type x, y . To reduce the influence of the unequal numbers of contacts for different residue pairs, we have divided values of $w_{k,l \rightarrow x,y}$ by the number of substituted contacts:

$$w'_{k,l \rightarrow x,y} = \frac{w_{k,l \rightarrow x,y}}{\sum_{kl} S_{k,l \rightarrow x,y}} \quad (9)$$

As a means of evaluating the sequence-structure match in terms of the contacts, a discrimination function of the form:

$$C_S = \left(\sum_{i,j \in G_{k,l \rightarrow x,y}} a_{i,j} \ln w'_{k,l \rightarrow x,y} \right) / n \quad (10)$$

is obtained. $G_{k,l \rightarrow x,y}$ is the set of i, j positions in the contact matrix, $a_{i,j}$, of a protein with known structure corresponding to the contacts between residue types k and l in the protein with the known 3D structure and the implied contacts between residues x and y in the protein sequence under comparison; n is the number of contacts contributing to C_S .

Comparisons

For the comparison of structure with sequence, 1 or more (aligned) 3D structures are converted to 1D structural templates using the computer program SPROF (Johnson et al., 1993). The resulting structural template is then compared against 1 or more amino acid sequences using the program PSLAVE (Johnson et al., 1993). In a typical search, the alignment score is the result of the sum of all residue-residue match scores less

the penalties that accrue with the appearance of gaps in the alignment. The score is then normalized by the length of the alignment and the scores are ranked from highest (closer match) to lowest (more distant match).

PSLAVE has now been modified to compute the residue-residue contact matrix of the input 3D structure. Based on the optimal alignment obtained using the environmental residue substitution tables (Johnson et al., 1993) and the implied contacts specified by this alignment with the known structure, the interacting residue pairs in the sequence are compared with the corresponding pairs in the structure using the residue-residue contact substitution probabilities. These matrices include the 3 residue-accessibility states available to the pairs of residues in contact (see above). For a search employing the alignment of 2 or more 3D structures, the average family contact matrix may be used.

Our strategy for evaluating comparisons is conducted in 2 stages. Firstly, the alignment scores, A_S , are ranked from high to low. The alignment scores are defined as the raw alignment scores (sum of the scores for individual matched positions less any penalties for gaps) by the length of the alignment (ignoring terminal overhangs). Secondly, we consider a cutoff value that is set to the average of 3 scores: the alignment score (A_S), the raw alignment score normalized for the maximum obtainable score given the scores in the structural template (A'_S), and the contact score (C_S). In the following example comparisons, we show that a score that is the average of the A_S , A'_S , and the evaluation function C_S , or C_S alone, can be used to rule out matches that are likely to be unrelated proteins. One fact must be underscored: the employment of residue contact substitutions in the postevaluation of sequence-structure matches does not owe any success to the improvement of the existing alignment. Instead, success results from the elimination of false matches that fortuitously score high with the search structure, but whose sequence of amino acids is not consistent with the folded 3D structure. Indeed, a number of the higher scoring false positives (according to the alignment score) clearly score high because of unusual amino acid compositions: they are higher than average in either glycine, alanine, leucine, cysteine, lysine, or serine. When evaluated in terms of residue-pair contacts, these proteins are unequivocally eliminated as possible homologues to the search protein.

In the following examples we illustrate the potential power of this approach where we consider only a single search structure or, in 1 example, 2 aligned 3D structures, and compare them against more than 21,000 primary structures. Two of the examples are particularly difficult. In the first case, we consider a search of the 3D structure of the erythrocrucorin (a globin) of *Chironomus thummi thummi*; in the second, we show that a single bacterial serine proteinase can detect members of the distantly related mammalian group. Although it is clear that we can improve upon these results by including the multiple alignment of 3D structures and using a composite structural template and average family contact map (see below) (for example, Barton & Sternberg [1990] have shown that by using the alignment of secondary structure elements of the erythrocrucorin with those of an α -globin chain that all globin sequences but 1 in a data bank could be identified with their flexible profile procedure), we wanted to show the effectiveness of this procedure in the worst case scenario: where a structure from only 1 search protein of a family is known.

Globin searches

A search of a sequence databank with the structural template created for the erythrocrucorin from midge larvae *C. thummi thummi* (Brookhaven code: 1ecd; Steigemann & Weber, 1979) is depicted in Figure 2 and Table 5. A structural template was constructed for the erythrocrucorin as described in Johnson et al. (1993) where each position along the fold is described by environment-dependent amino acid substitution scores according to the local structural environment of the residue: the residue solvent accessibility and secondary structure. The template constructed from the single structure was searched against a databank of sequences containing 21,792 sequences >55 residues in length.

In this search, 537 globins were detected prior to the first non-globin: the chicken "B-G antigen" (Owl code: CHKBGAF). The last of 618 globin sequences, an extracellular earthworm globin (Owl code: GLBI_PHESE), is found at ranked position 901 (where the highest score obtained, A_S , is located at position 1). Between the last of the contiguous high-scoring globin matches at position 537 and the final globin at position 901 there are a total of 283 non-globins (Table 5). Note the overlap between the non-globin and the authentic globin matches in the inset of Figure 2A.

In Figure 2B,C and Table 5 are shown the consequences of reevaluating the alignments in terms of residue-residue contact substitutions. For example, the top-ranked matches in the erythrocrucorin search can be screened according to the score obtained for the implied contacts (C_S). For a cutoff value of $C_S = 9.0$, we can both increase the number of globin matches before the first observed non-globin to 576 (Table 5) and reduce the number of contaminating matches before the last globin at position 901 to only 11! Thus, only 42 globins had contact scores less than 9.0 and many of these were within the top matches on the basis of the alignment scores and would have been detected on the basis of the A_S alone. Figure 2B shows, in comparison with Figure 2A, the massive reduction of non-globin matches when matches have been eliminated according to this criterion.

The results obtained when we consider a cutoff value that is set to the average of 3 scores: the alignment score (A_S), the normalized alignment score (A'_S), and the contact score (C_S) are shown in Figure 2C and Table 5. For example, with the combined-score cutoff value of 7.5, 604 of the 618 globin matches have a score larger than this cutoff; only 14 globins were eliminated from the qualifying matches due to scores lower than the cutoff. Although the first non-globin was detected at ranked position 538, there are 35 remaining non-globins between the globin at ranked position 537 and the earthworm globin at position 901. At the higher cutoff score of 9.0, this number is reduced to 6, the first non-globin appears at ranked position 595 (Owl code: TPC_TACTR, horseshoe crab troponin C), but 46 globins score less than the cutoff and 572 globins achieve the cutoff value (Table 5).

In comparison with the results reported above, when the sequence of the erythrocrucorin was compared against all of the sequences in the databank and scored using the Dayhoff matrix (Dayhoff et al., 1978), only 239 globins were detected prior to the first non-globin (Johnson et al., 1993).

In Table 6 are shown 2 additional globin comparisons: a search of the α -chain of human hemoglobin (2hhb; Fermi et al., 1984) and a search of a template constructed from a structural align-

Fig. 2. Searches of a single 3D structure against each of the 21,792 sequences (>55 residues) in a sequence databank (Owl version 9.0; Bleasby & Wooton, 1990). The erythrocrucorin of *C. thummi thummi* (1ecd; Steigemann & Weber, 1979) was encoded as a structural template with the program SPROF (Johnson et al., 1993). The searches were made with the program PSLAVE (Johnson et al., 1993) adapted to evaluate each alignment in terms of the match of residue-pair substitution with those collected from 96 families of aligned 3D structures as described in the text; no globins were present among the 96 families contributing to the residue-pair substitution data. **A–C:** The frequency of globin alignment scores, normalized for the length of the alignment (A_S), are depicted with solid vertical lines; the non-globins are indicated with dotted lines. Frequencies of scores greater than 500 (7,566 scores in A; 1047 in B; 0 in C) have been removed for the sake of clarity. In (A) all scores are depicted (less those with frequencies >500); the inset shows the overlap between the authentic globin matches and the non-globin matches. In (B) only alignments where a minimum contact score (C_S) of 9.0 was obtained are shown; 69% of the non-globins are eliminated. In (C) a cutoff score of 7.5 was required (average of the combination of A_S , C_S , and A'_S , where the latter is calculated as the raw alignment score relative to the maximum possible score obtained from the largest values at each position along the structural template); 84% of the non-globin sequences are eliminated by this screen. A dotted vertical line indicates the position of the last globin match at position 901 of the original ranked alignment scores (A_S): the extracellular earthworm globin (GLBI_PHESE).

Table 5. Search of the structural template of the erythrocrucorin of *Chironomus thummi thummi* against a sequence data bank containing 21,792 sequences ≥ 55 residues in length

Cutoff score	Globins < cutoff	Globins ≥ cutoff	Contaminants ^a (prior to last globin)	First non-globin		Identity
				Rank	Code	
1. Alignment score, A_S						
^b	81	537	283	538	CHKBGAF	B-G antigen (AA at 1)— <i>Gallus gallus</i>
2. Contact score, C_S						
6.0	8	596	41	538	CHKBGAF	B-G antigen (AA at 1)— <i>G. gallus</i>
7.0	27	591	26	563	TPC2_ASTPO	Troponin C, isotype gamma—crayfish
8.0	32	586	19	565	R7SP12	70S ribosomal protein L12—spinach
9.0	42	576	11	595	TPC_TACTR	Troponin C—Japanese horseshoe crab
10.0	50	568	9	595	TPC_TACTR	Troponin C—Japanese horseshoe crab
3. $(A_S + A'_S + C_S)/3^c$						
7.0	8	610	84	538	CHKBGAF	B-G antigen (AA at 1)— <i>G. gallus</i>
7.5	14	604	35	538	CHKBGAF	B-G antigen (AA at 1)— <i>G. gallus</i>
8.0	25	593	19	538	CHKBGAF	B-G antigen (AA at 1)— <i>G. gallus</i>
8.5	33	585	10	565	R7SP12	70S ribosomal protein L12—spinach
9.0	46	572	6	595	TPC_TACTR	Troponin C—Japanese horseshoe crab

^a Non-globins that exceed the cutoff value but are ranked ahead of the last globin match.

^b Only the ranked alignment scores, A_S , normalized by the number of aligned positions, were considered.

^c C_S = contact probability score; A_S = alignment score divided by the number of aligned positions; A'_S = normalized alignment score, which takes into account the maximum possible score given the residue types in each pair of compared proteins.

ment between the erythrocrucorin (1ecd) and the α -chain of human hemoglobin (2hhb). For the latter search, the residue-residue contact map was derived from the human α -chain only. On the basis of the template alignment score alone (Table 6), these comparisons net 571 and 606 of the total 618 globins, respectively.

With a contact score cutoff, $C_S = 9.0$, the human α -chain finds all of the globins prior to the first non-globin at ranked position 673 (the last globin appears at ranked position 668); 11 of the 618 globins score less than the cutoff score (6 of the 11 are incomplete sequences). From the average combined

score (cutoff = 7.5), all but 3 globins are found before the first non-globin that exceeds the cutoff value; 2 contaminating non-globins are found before the final globin.

For the joint 2hhb/1ecd search (Table 6) ($C_S = 9.0$), all 618 globins are located before the first non-globin at ranked position 825 (last globin at 653); 12 of the 618 globins attained a score less than the cutoff (6 of the 12 are incomplete sequences). With the combined cutoff score every globin is found before the first non-globin, but there are no contaminating non-globins before the appearance of the last globin, and there are no globins that score less than the cutoff score (Table 6).

Table 6. Search of the structural template based on (2hhb α) the structure of the α -chain of hemoglobin and (2hhb α /1ecd) the structures of the α -chain of hemoglobin and the erythrocrucorin

	Cutoff score	Globins < cutoff	Globins ≥ cutoff	Contaminants (prior to last globin)	First non-globin		Identity
					Rank	Code	
1. Alignment score, A_S							
2hhbα	a	47	571	50	572	RLA4_YEAST	Ribosomal protein YPA1 – Baker's yeast
2hhbα/1ecd	a	12	606	35	607	R5EC7	50S ribosomal proteins L7/L12 – <i>E. coli</i>
2. Contact score, C_S							
2hhbα	9.0	11	607	0	673	YKP5_KLULA	Hypothetical killer plasmid protein 5 – yeast
2hhbα/1ecd	9.0	12	606	0	825	ATPD_RHOBL	ATP synthase delta chain – <i>Rhodopseudomonas</i>
3. $(A_S + A'_S + C_S)/3$							
2hhbα	7.5	3	615	2	626	R5EC6	50S ribosomal protein L6 – <i>E. coli</i>
2hhbα/1ecd	7.5	0	618	0	702	ACP1_SPIOL	Acyl carrier protein – spinach

^a All matches prior to first non-globin.

Serine proteinases

A structural template for the α -lytic proteinase from *Lysobacter enzymogenes* (2alp; Fujinaga et al., 1985) was searched against the sequence databank. The first 11 sequences identified in this

approach are all members of the serine proteinase family: proteinases A and B from *Streptomyces griseus* and 4 matches with α -lytic proteinase sequences themselves. In Table 7 are shown some of the results of this search: the first 41 top-scoring hits (ranked by values of A_S), where arrows indicate members of

Table 7. Top 41 hits in a search of the α -lytic proteinase structure (Brookhaven code: 2alp; Fujinaga et al., 1985) against the 21,792 sequences in a sequence data bank^a

Rank	Code	A_S	A'_S	C_S	$(A_S + A'_S + C_S)/3$	Title
1	← MXBPALP	44.69	44.69	100.00	63.13	MXBPALP prepro-alpha-lytic protease— <i>Lysobacter enzymogenes</i>
2	← N_2ALP1	44.69	44.69	100.00	63.13	Alpha-lytic protease (EC number not assigned)— <i>L. enzymogenes</i>
3	← MXBPALPA	44.69	44.69	100.00	63.13	MXBPALPA prepro-alpha-lytic protease— <i>L. enzymogenes</i>
4	← PRLA_LYSEN	44.69	44.69	100.00	63.13	Alpha-lytic protease precursor (EC 3.4.21.12)— <i>L. enzymogenes</i>
5	← N_ISGC1	20.01	19.62	33.77	24.47	Proteinase A complex with chymostatin— <i>S. griseus</i>
6	← PRSMBG	19.98	18.84	24.15	20.99	Protease B (EC 3.4.21.—)— <i>S. griseus</i>
7	← N_3SGB1	19.96	18.81	25.80	21.52	Proteinase B (EC number not assigned) (SGPB)— <i>S. griseus</i>
8	← PRSMAG	19.89	19.51	38.25	25.88	Protease A (EC 3.4.21.—)— <i>S. griseus</i>
9	← PRTA_STRGR	19.85	19.38	30.42	23.21	Protease A precursor (EC 3.4.21.—)— <i>S. griseus</i>
10	← STMSPRB	18.50	16.14	19.41	18.02	STMSPRB protease B precursor (GTG start codon)— <i>S. griseus</i>
11	← STMSPPRA	15.72	12.15	19.11	15.66	STMSPPRA protease A precursor (GTG start codon)— <i>S. griseus</i>
12	CHB1_BOMMO	14.18	13.60	0.58	9.45	Chorion class B protein L11 precursor—silk moth (<i>Bombyx mori</i>)
13	CHB4_BOMMO	14.10	13.59	0.60	9.43	Chorion class B protein B.L1 (410) (fragment)—silk moth (<i>B. mori</i>)
14	CHKIGLAM	13.88	13.75	0.50	9.38	CHKIGLAM Ig lambda-chain V-J region (AA at 2)— <i>Gallus gallus</i>
15	GRP1_PHAVU	13.88	12.32	1.96	9.39	Glycine-rich cell wall structural protein GRP 1.0 precursor— <i>Phaseolus vulgaris</i> (kidney bean)
16	→ TRPGTR	13.77	10.98	10.46	11.74	Trypsinogen (EC 3.4.21.4)—pig
17	→ TRFF	13.45	10.70	8.57	10.91	Trypsinogen-like (EC 3.4.21.—) proenzyme precursor—fruit fly
18	→ KCUF	13.38	10.75	8.17	10.76	Collagenolytic protease (EC 3.4.21.32)—Atlantic sand fiddler crab
19	CHKIGLAN	13.24	12.94	0.34	8.84	CHKIGLAN Ig lambda-chain V-J region (AA at 2)— <i>G. gallus</i>
20	KRBO23	13.22	12.28	1.03	8.84	Keratin, 60K type II cytoskeletal, component III—bovine (fragment)
21	→ N_INTP1	13.22	10.31	6.66	10.06	Modified beta trypsin (monoisopropylphosphoryl inhibited)—bovine pancreas
22	→ N_ITGB1	13.21	10.30	6.94	10.15	Trypsinogen-CA from PEG—bovine (<i>Bos taurus</i>) pancreas
23	BASE77	13.20	12.30	0.55	8.68	Chorion (egg shell) C protein 10A—silk moth (<i>Antheraea polyphemus</i>)
24	→ N_ITGS1	13.16	10.23	7.46	10.28	Trypsinogen complex with porcine pancreatic secretory trypsin inhibitor
25	→ TRBOTR	13.16	10.23	7.54	10.31	Trypsinogen (EC 3.4.21.4)—bovine
26	→ EZEC540	13.11	10.20	7.65	10.32	Trypsinogen (EC 3.4.21.4)—bovine
27	→ TRDGC	13.10	10.59	6.92	10.20	Trypsinogen (EC 3.4.21.4), cationic, precursor—dog
28	→ EL1_PIG	13.10	10.34	7.79	10.41	Elastase I precursor (EC 3.4.21.36)—pig (<i>Sus scrofa</i>)
29	→ ELPG	13.07	10.34	7.32	10.24	Elastase (EC 3.4.21.11)—pig
30	→ N_1EST1	13.07	10.34	7.32	10.24	Tosyl-elastase (EC 3.4.21.11)—porcine (<i>S. scrofa</i>) pancreas
31	→ N_ITGN1	13.07	10.17	6.17	9.80	Trypsinogen—bovine (<i>B. taurus</i>) pancreas
32	B25297	13.07	12.02	0.91	8.67	Class B protein m2G12 precursor—silkworm (fragment)
33	→ TRY1_HUMAN	13.03	10.59	6.94	10.19	Trypsinogen I precursor (EC 3.4.21.4)—human (<i>Homo sapiens</i>)
34	→ TRDFS	13.01	10.96	8.88	10.95	Trypsinogen (EC 3.4.21.4)—spiny dogfish
35	→ TRY2_HUMAN	12.94	10.51	6.89	10.11	Trypsinogen II precursor (EC 3.4.21.4)—human (<i>H. sapiens</i>)
36	A24713	12.90	11.72	1.26	8.63	Sericin
37	→ N_2TPI1	12.87	10.06	5.85	9.59	Trypsinogen pancreatic trypsin inhibitor
38	→ TRY3_RAT	12.84	9.74	7.29	9.96	Trypsinogen III cationic precursor (EC 3.4.21.4)—rat (<i>Rattus norvegicus</i>)
39	SODC_XENLA	12.83	12.36	0.55	8.58	Superoxide dismutase I (Cu-Zn) (EC 1.15.1.1)—African clawed frog (<i>Xenopus laevis</i>)
40	→ EZEC562	12.83	10.19	4.41	9.14	Kallikrein (EC 3.4.21.34)—pig pancreas
41	CHHA_BOMMO	12.80	12.69	0.29	8.51	Chorion class high-cysteine HCA protein 12 precursor—silk moth (<i>B. mori</i>)

^a All 21,792 matches were ranked according to their alignment scores A_S and A'_S ; in addition, the contact probability score C_S and the average contact alignment score are shown for each match. Bacterial serine proteinases closely related to the search structure are indicated (←), as are those that belong to the mammalian group (→). Access codes are noted along with their titles in the data bank.

the bacterial family (left pointing) and the mammalian family (right pointing); the remaining matches are to proteins unrelated to the serine proteinases and are all characterized by unusual sequences with compositions exceptionally high (17–63%) in a few amino acids such as glycine, serine, and cysteine. Notice that on the basis of the contact score alone, C_S , the values for the unrelated protein matches range from 0.29 to 1.96; for the mammalian serine proteinase matches, the contact scores range between 4.41 and 10.46. From the combined average score, the largest value seen in Table 7 for a non-serine proteinase corresponds to a match with the sequence of the silk moth chorion protein (9.45). All of the authentic serine proteinases shown in Table 7 exceed this value except one: the sequence of pig porcine kallikrein (9.14).

Discussion

In the structural template method (Johnson et al., 1993) each 3D structure is recoded at each residue position as 21 scores relating the probability of a match between the fold and a sequence at that location. The scores that appear at each position are determined by the structural environment about the residue. These scores are themselves derived from the amino acid substitutions observed in nearly 100 families of aligned 3D structures where the substitutions were segregated according to these same structural features. But these scores consider only information local to the sequence, for individual positions, and without regard to any other residues in the fold.

In an effort to introduce data on contacts within proteins into our procedure, a study has been made of the effect of side-chain main-chain interactions on amino acid substitutions (Edwards et al., 1994; Johnson et al., 1994a).

In this study we have shown how residue-contact substitution tables can be used to evaluate a comparison after the alignment between sequence and structure has taken place. These substitutions take into account residue contacts and the solvent accessibility of the residues. It was our aim to use these data, which are local in space but may be distant in sequence, to evaluate each sequence–structure match.

The linchpin of this procedure is the alignment. Before contacts can be assessed, an alignment must be constructed that provides the mapping of contacts that occur in the known 3D structure to the sequence whose structure is not known. To evaluate all possible folds is clearly a task that belongs to a set of computationally insoluble problems known as “NP-complete” and where approximations are the only reasonable solution. Relationships, such as contacts between residues in different portions of a chain, are not easily accommodated directly within dynamic programming approaches, although Jones et al. (1992) have used the procedure of Taylor and Orengo (1989) to incorporate residue–residue potentials into their evaluation of structure and sequence. One major disadvantage of that approach is the length of time required for comparison before the final alignment is achieved because a large number of solutions of a Needleman and Wunsch-type (1970) matrix are required for a single comparison of 2 proteins. In an alternative approach, Godzik et al. (1992) have used Monte Carlo simulations to generate alternative alignments between structure and sequence.

In this work, the alignment uses a 3D → 1D template and a single dynamic programming comparison is made using the fast

Fredman (1984) procedure. The major computational overhead is the subsequent evaluation of the contact matrix for the search structure and this occurs 1 time only, whether 1 or thousands of sequences are compared with the structure. Once the contact matrix is obtained, the alignment between the structural template and the sequence is used to assign the putative contacts within the sequence of unknown 3D structure. Relatively little time is involved in the evaluation step. For a search of 1 or more sequences against all the known 3D structures, the computational time requirements for calculating a new contact map for each structure can be offset by the use of precalculated contact maps for individual structures and for aligned sets of families.

The postevaluation of the 1D to 3D matches does not introduce changes into any of the alignments. What does occur is the elimination of matches that have scored high based on the sequence–structure template match but whose residue–residue contacts, implied by the match to the known structure, are not consistent with the patterns of pair-contact amino acid substitutions observed across 96 families of 3D structures.

The globins are a particularly difficult set of proteins in which to detect all members of the family. There are no peaks of sequence identity across an alignment that reflect localized conserved motifs that could act as a signature for the globins as a whole. The sequence similarity between the 618 globins in the sequence databank and one of the search structures, the erythrocruorin, ranges from nearly complete identity with the 22 midge globins in the sequence data bank to between about 23% (myoglobins) and to less than the 15% identity recorded for a match with a leghemoglobin. Indeed, a structural alignment between the midge protein and the hemoglobin of *Scapharca inaequivalvis* (Brookhaven code: 1sdh) indicates a sequence identity of only 13% (Table 1), although their contact maps, given their aligned 3D structures, are conserved over 75% of the contacts.

The postalignment evaluation of the matches and the use of contact scores or the combined scores permits the detection of a majority of the globins in the data bank before the first non-globin (Table 3). This includes the unequivocal detection of such distantly related matches as the bacterial globin from *Vitreoscilla* and the globin from the alga *Parasponia*. By including the structure of the human hemoglobin α -chain in the structural template, we have found that 100% of the globins can be detected prior to the first non-globin.

A second set of difficult comparisons involves alignments between a bacterial serine proteinase structure and sequences of the mammalian serine proteinase family. Although they are clearly a homologous group, there are large differences mainly reflected in a number of large gaps and the low levels of sequence identity that makes their comparison difficult. For example, the search of a sequence of a bacterial serine proteinase against a sequence databank will fail to place any of the mammalian counterparts within the top scoring matches (Johnson et al., 1994a). Here we have shown that the structural template procedure can by itself identify proteinases of the mammalian group and that the resolution of these matches is increased when evaluated according to substitutions involving residue–residue contacts.

Barton and Sternberg (1990) have very successfully used a “flexible profile,” constructed from elements of secondary structure of the 3D structure from the search protein, to locate related sequences in sequence databank searches. Given the alignment of two 3D structures, the α -chain of human hemo-

globin and the erythrocrucorin, all but 1 globin were detected in a search (Barton & Sternberg, 1990). In the procedure of Barton and Sternberg (1990), it is information from the elements of secondary structure, which largely comprise the conserved core of the protein structure, that is used in their comparisons; loop regions are not considered.

In the present study, we include all residue positions including loops, but it is also clear that more contacts will be tabulated for buried positions than for those residues exposed to solvent at the surface of a protein. We have repeated the globin comparison made by Barton and Sternberg (1990) using our own procedure (Table 6). From a template constructed from these 2 aligned structures and the postevaluation of the highest scoring matches on the basis of the contact map for the α -chain of human hemoglobin (average combined score), all 618 globins were successfully detected prior to the appearance of the first non-globin match that achieved the cutoff value. One useful advantage of our own approach is that the mapping of the residue-residue contact map of the known 3D structure to the sequence whose structure is not available lends itself directly to the reconstruction of a 3D model for the sequence on the basis of these matched contacts.

Methods

Computer programs were written in C and Fortran. Calculations were made on a personal Iris (Silicon Graphics) and a personal computer operating under the Unix operating system and employing Intel i860 processors. Release 9 of the Owl sequence databank (Bleasby & Wooton, 1990) was used. Values (~ 9 megabytes) of $w'_{k,l \rightarrow x,y}$ (see Equation 9) used in the calculation of the discrimination function, along with the computer programs described here, will be available from March 1, 1995 onward on our World Wide Web home page: <http://serv.lfy.utu.fi/molmol> or by contacting the authors at the Centre for Biotechnology, P.O. Box 123, FIN-20521, Turku, Finland.

Acknowledgments

We gratefully acknowledge the support provided by the Royal Society and the European Molecular Biology Organization. The Imperial Cancer Research Fund has also contributed toward the costs of this research. We thank John P. Overington for the current release of the alignment databank.

References

- Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications*. Data Commission of the International Union of Crystallography. pp 107–132.
- Barton GJ, Sternberg MJE. 1990. Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. *J Mol Biol* 212: 389–402.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, 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.
- Bleasby AJ, Wooton JC. 1990. Construction of validated, non-redundant composite protein sequence databases. *Protein Eng* 3:153–159.
- Blundell TL, Johnson MS. 1993. Catching a common fold. *Protein Sci* 2:877–883.
- Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into known three-dimensional structures. *Science* 253: 164–170.
- Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model for evolutionary change. In: Dayhoff MO, ed. *Atlas of protein sequence and structure*, vol 5, suppl 3. National Biomedical Research Foundation. pp 345–358.
- Edwards YJ, Johnson MS, Moss DS, Blundell TL. 1994. The effects of local environments on the pattern of amino-acid substitution in homologous protein structures. The role of side-chain to main-chain van der Waals interactions. In: Crabb JW, ed. *Techniques in protein chemistry*. San Diego, California: Academic Press pp. 405–412.
- Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
- Fermi G, Perutz M, Shaanan B, Fourme R. 1984. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol* 175:159–174.
- Fredman ML. 1984. Computing evolutionary similarity measures with length independent gap penalties. *Bull Math Biol* 46:553–566.
- Fujinaga M, Delbaere LTJ, Brayer GD, James MNG. 1985. Refined structure of α -lytic protease at 1.7 Å resolution: Analysis of hydrogen bonding and solvent structure. *J Mol Biol* 184:479–502.
- Galaktionov SG, Rodionov MA. 1980. Calculation of the tertiary structure of proteins on the basis of analysis of matrices of contacts between amino acid residues. *Biofizika* 25:395–403. [Translated from *Biofizika* 25:385–392.]
- Galaktionov SG, Rodionov MA, Golubovich VP. 1975. Investigation of residue-residue contact systems in globular proteins. *Abst Soviet-French Symp Phys Chem Proteins, Pushino on Oka*:66.
- Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse folding problem. *J Mol Biol* 226:227–238.
- Hubbard TJP, Blundell T. 1987. Comparison of solvent inaccessible cores of homologous proteins: Definitions useful in protein modelling. *Protein Eng* 1:155–171.
- Johnson MS, Overington JP, Blundell TL. 1993. Alignment and searching for common protein folds using a data bank of structural templates. *J Mol Biol* 231:735–752.
- Johnson MS, Overington JP, Edwards Y, May ACW, Rodionov MA. 1994a. The comparison of structures and sequences: Alignment, searching and the detection of common folds. *Proc 27th Hawaii Int Conf Sys Sci* 5:296–305.
- Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. 1994b. Knowledge-based protein modelling. *Crit Rev Biochem Mol Biol* 29:1–68.
- Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature* 358:86–89.
- Kuntz ID. 1975. An approach to the tertiary structure of globular proteins. *J Am Chem Soc* 97:4362–4366.
- Lüthy R, McLachlan AD, Eisenberg D. 1991. Secondary structure-based profiles: Use of the structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins Struct Funct Genet* 10:229–239.
- Needleman SB, Wunsch C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:444–453.
- Ooi T, Nishikawa K. 1974. Comparison of homologous tertiary structures. *J Theor Biol* 43:351–374.
- Ouzounis C, Sander C, Scharf M, Schneider R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness. *J Mol Biol* 232:805–825.
- Overington JP, Donnelly D, Sali A, Johnson MS, Blundell TL. 1992. Environmental-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci* 1:216–226.
- Overington JP, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structural prediction. *Proc R Soc Lond B* 241:146–152.
- Overington JP, Zhu ZY, Johnson MS, Sowdhamini R, Louie GV, Blundell TL. 1993. Molecular recognition in protein families: A database of aligned 3D structures of related proteins. *Biochem Soc Symp* 21:597–604.
- Phillips DS. 1970. Development of crystallographic enzymology. *Biochem Soc Symp* N31:11–28.
- Rodionov MA, Galaktionov SG. 1992a. The analysis of protein three-dimensional structures in terms of residue-residue contact matrices. I. Contact criterion. *Mol Biol* 26:773–776. [Translated from *Mol Biol (Russia)* 26:1160–1166.]
- Rodionov MA, Galaktionov SG. 1992b. The analysis of protein three-dimensional structures in terms of residue-residue contact matrices. II. Coordination numbers. *Mol Biol* 26:777–783. [Translated from *Mol Biol (Russia)* 26:1167–1180.]
- Rodionov MA, Galaktionov SG, Aklreem AA. 1981. Prediction of amino acid residue accessibilities in globular proteins. *Dokl Akad Nauk SSSR* 261:756–759.

- Rodionov MA, Guzevich AV, Galaktionov SG. 1993. The analysis of protein three-dimensional structures in terms of residue-residue contact matrices. III. Residue affinity. *Mol Biol* 27:220-224. [Translated from *Mol Biol (Russia)* 27:363-370.]
- Sali A, Overington JP, Johnson MS, Blundell TL. 1990. From comparison of protein sequences and structures to protein modelling and design. *Trends Biochem Sci* 15:235-240.
- Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. *J Mol Biol* 213:859-883.
- Steigemann W, Weber E. 1979. Structure of erythrocyruorin in different ligand states refined at 1.4 Å resolution. *J Mol Biol* 127:309-338.
- Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol* 208:1-22.