Protein sequence–structure compatibility criteria in terms of statistical hypothesis testing

Shamil Sunyaev, Eugene Kuznetsov, Igor Rodchenkov, and Vladimir Tumanyan

Engelhardt Institute of Molecular Biology RAS, 32 Vavilov str., Moscow 117984 and Institute for Control Sciences RAS, 65 Profsoyuznaya str., Moscow 117806, Russia

© Oxford University Press 635

Introduction

In spite of numerous efforts for more than 30 years, the prediction of protein spatial structure from its amino acid sequence alone is not yet possible. Although progress was achieved in understanding the basic principles of protein folding (Chemeris et al., 1994; Eisenhaber et al., 1995) it is not sufficient for an a priori elucidation of tertiary structure. Meanwhile, the rapid growth of the databank of experimentally determined protein tertiary structures has encouraged the development of prediction methods based on statistical analysis of structures already known. It is clear that the number of different protein folding topologies is much smaller than the number of existing amino acid sequences. This means that the tertiary structure prediction problem can be partially solved in many cases by assigning one of the known folding types to a protein with unknown tertiary structure according to its amino acid sequence, i.e. by recognizing an appropriate fold for a particular primary structure. A variety of different approaches to this sequence–structure recognition problem have recently been developed (Eisenhaber et al., 1995).

Most existing fold recognition methods (threading methods) place an amino acid sequence into (thread it through) a tertiary structure and then evaluate their match for compatibility (Eisenhaber et al., 1995). This approach implies the application of a compatibility measure (goodness-of-fit evaluation function) and takes into account appropriate insertions and deletions. Therefore, these methods include statistically derived 3D–1D (structure–sequence) compatibility criteria such as the mean force potential energy or scoring function and the sequence–structure alignment procedure. The 3D–1D compatibility criteria are obtained from statistical preferences of amino acid residues of particular types to be located in the specific spatial environment described by a set of environmental variables such as residue accessibility, number of neighboring residues, torsion angles, hydrogen bonds, secondary structure and contacts (residue pair interactions or distances) with amino acids of various types (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Sippl and Weitckus, 1992; Bryant and Lawrence, 1993; Nishikawa and Matsuo, 1993; Abagyan et al., 1994; Kocher et al., 1994; Wilmanns and Eisenberg, 1995). Therefore, tertiary structure is represented either as a one-dimensional string of environmental variable values sometimes called an environmental profile (or a set of such strings—for a group of variables) or as a two-dimensional contact matrix (or a set of such matrices). Thus, threading is the alignment of an amino acid sequence and a structure represented by an environmental profile or a contact matrix. In this paper, we will consider only one-dimensional structure representation (profile) for the purpose of convenience in writing mathematical equations. The derivation of equations for the case of a contact matrix is straightforward.

There are a number of sequence–structure recognition methods which differ in their definitions of a residue’s spatial environment and alignment procedures. Generally, the mathematical form of 3D–1D compatibility criteria, taken as an empirical energy function, is often given by

\[ E = -kT \sum \ln \frac{f_a(x)}{g(x)} \]

where \( x \) is an environmental variable, \( f_a(x) \) the frequency of amino acid \( a \) to be located in environment \( x \) and \( g(x) \) the equivalent frequency for all amino acid types. The sum is taken over all alignment positions (Godzik et al., 1992; Jones et al., 1992; Sippl and Weitckus, 1992; Bryant and Lawrence, 1993; Nishikawa and Matsuo, 1993; Kocher et al., 1994). Following the general framework of statistical mechanics, the authors assume a Boltzmann-like distribution of an environ-
Fig. 1. Empirical distribution functions, histogram and Parzen probability density estimates of solvent accessibility for lysine residues and for all residues irrespective of their type. For variable definition and calculation technique and for protein database used to calculate distributions, see text. In histograms the left-most narrow bars are not shown.
An alignment of an amino acid sequence and an environmental variable profile to be tested for compatibility

**correct sequence - structure match**

\[
\begin{array}{cccccccc}
K & F & A & D & K & A & G & K \\
0.48 & 0.21 & 0.00 & 0.23 & 0.34 & 0.00 & 0.03 & 0.69 \\
\end{array}
\]

\{y\}_K = \{ 0.48, 0.34, 0.69 \}

\[F^*_K(x)\]

Hypothesis \( H_0 \): each sample \( \{y\}_a \) is drawn from its corresponding separate distribution \( F_a \), i.e. sequence is compatible with given structure

**incorrect sequence - structure match**

\[
\begin{array}{cccccccc}
K & F & A & D & K & A & G & K \\
0.21 & 0.00 & 0.23 & 0.34 & 0.00 & 0.03 & \\
\end{array}
\]

\{y\}_K = \{ 0.21, 0.23, 0.03 \}

\[G^*(x)\]

Hypothesis \( H_1 \): every sample \( \{y\}_a \) is drawn from one and the same common distribution \( G \), i.e. sequence is not compatible with given structure

---

**Fig. 2.** The fold recognition problem stated as a problem of statistical hypothesis testing. As an example of correct sequence–structure match, the proper alignment of sequence fragment of 8abp (in one-letter notation) with its accessibility profile (shown below by raw numbers) is presented. As an example of incorrect sequence–structure match, an inappropriate alignment of the same sequence and profile fragments is shown to the right (underlines denote gaps). In the case of a correct match the query sample of environmental variable values corresponding to lysine residues are thought to be drawn from the distribution for lysine residues (the empirical distribution of accessibility values for lysine residues is shown on the left). Otherwise, this sample is thought to be drawn from the distribution for all residues regardless of their type (the empirical distribution of accessibility values is shown on the right).

---

Statistical criteria in 1D–3D compatibility

mental variable. The concrete form of these distributions is thought to be dependent on the energy function (Wilbur and Liu, 1994; Wilbur et al., 1996). Even if the scoring function is not considered to be an energy (Bowie et al., 1991), its mathematical form does not differ significantly.

Scoring functions for threading methods give priority to those sequence–structure alignments which comply with statistical rules known for proteins with determined structures. In correct sequence–structure matches amino acids of a certain type are expected to be located mostly in the environment they prefer statistically in native proteins. In other words, it is tested how likely it is that the spatial location of amino acid residues of various types resembles the same distribution as in known protein structures. In our opinion, the fold recognition problem as formulated above can be a subject of mathematical statistics more than of statistical mechanics.

Two possible ways to improve the efficiency of statistical fold recognition methods can be pointed out. The first includes the application of more sophisticated environmental variables which properly reflect physical principles of protein folding as well as taking into account different combinations of variables, while the second implies the development of an adequate mathematical functional form of compatibility criteria. Up to now, mainly the first approach was explored. Various authors use distinct sets of heuristically chosen environmental variables (Eisenhaber et al., 1995). This work will explore the second approach.

Here we present a general formal approach to the sequence–
Structure recognition problem which is formalized in terms of statistical mathematics as a problem of hypothesis testing. It enables us to derive several mathematical forms of 3D–1D compatibility criteria and to develop techniques for the evaluation of various environmental variables. However, owing to space constraints, we focus here only on the mathematical forms of the scoring function.

These 3D–1D compatibility criteria were derived as criteria for statistical hypothesis testing without any assumption of the Boltzmann or any other distribution of environmental variables. Two less strict assumptions underlying the presented approach can be verified by widely used statistical tests. We consider here only additive criteria (namely three of them) because otherwise the alignment problem becomes computationally difficult, which makes the practical application of non-additive criteria problematic. One of the mathematical forms of a scoring function suggested here, although based on different underlying statistical concepts, happens to be Equation 1, and two others are new.

To test and compare the criteria developed, we performed searches through the library of non-homologous sequences of proteins with known 3D structure to recognize proteins which adopt a certain folding topology. Tests revealed the benefit of the combined implementation of several criteria because in some cases only one of them is able to detect correct sequence–structure matches. This can be explained by the fact that criteria are based on different statistical concepts. On the other hand, the reliability of an inference increases when all the criteria simultaneously recognize an amino acid sequence as best fitting the template fold.

Methods

Threading algorithms and statistical hypothesis testing

Threading prediction algorithms use the well known fact that amino acid residues of various types statistically prefer to be located in specific positions in protein globula. For example, hydrophobic residues more often occur in the hydrophobic core while polar residues more often are located on the surface of proteins; oppositely charged residues tend to be in close contact; glycine residues prefer loops, etc. Threading techniques place a query amino acid sequence into one of the known template folds and thus produce a model of a protein molecule. This hypothetical protein molecule is further tested for how well it complies with the statistical rules observed in native proteins. The 3D–1D compatibility criterion is thus a statistical decision rule function which evaluates the likelihood that the spatial location of residues in the hypothetical protein is determined by the same distribution as the spatial location of residues in known native protein structures. In our opinion, the approach of statistical hypothesis testing is suitable for the evaluation of correct sequence–structure fits.

Mathematical statistics operates on data just as with samples of random variable realizations. This means that data samples are considered to be drawn from a probability distribution function. In statistics problems this function is usually assumed to be unknown. Propositions concerning the type or parameters of the distribution function of data under study are called statistical hypotheses. Statistical theory aims to determine which proposition is more likely. The problem of statistical hypothesis testing is an old and well studied problem in mathematical statistics. Theoretical recipes for decision rule derivations in various cases have been developed. We will consider the fold recognition problem in the conventional framework of statistical hypothesis testing theory.

We assume that the tertiary structure of a protein can be represented by a profile of the environmental variable values of its residue sites. Let us consider a natural set (a population) of all existing proteins with their spatial structures represented by profiles (irrespective of the fact that they are already known or not). In addition, we can think of each profile not only as a general collection (not an ordered set) of values. Hence the whole protein population can be represented as an enormous set of environmental variable values.

S.Sunyaev et al.

Table I. Learning set of 167 proteins with known 3D structure

<table>
<thead>
<tr>
<th>1aaj</th>
<th>1cfb</th>
<th>1gmf</th>
<th>1lba</th>
<th>1php</th>
<th>1php</th>
<th>2gst</th>
<th>2pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1bkb</td>
<td>1chm</td>
<td>1gof</td>
<td>1len</td>
<td>1piu</td>
<td>1pkp</td>
<td>1pr</td>
<td>2hbg</td>
</tr>
<tr>
<td>1add</td>
<td>1emc</td>
<td>1gox</td>
<td>1lga</td>
<td>1mpy</td>
<td>1sr</td>
<td>2hmp</td>
<td>3aa</td>
</tr>
<tr>
<td>1ads</td>
<td>1col</td>
<td>1grp</td>
<td>1lis</td>
<td>1poa</td>
<td>1trb</td>
<td>2hp</td>
<td>3bcl</td>
</tr>
<tr>
<td>1akp</td>
<td>1epc</td>
<td>1gpl</td>
<td>1lid</td>
<td>1pox</td>
<td>1tg</td>
<td>2hp</td>
<td>3cd4</td>
</tr>
<tr>
<td>1amnp</td>
<td>1epq</td>
<td>1hbpq</td>
<td>1lpe</td>
<td>1poc</td>
<td>1vaa</td>
<td>2hl</td>
<td>3chy</td>
</tr>
<tr>
<td>1aoz</td>
<td>1crcl</td>
<td>1hec</td>
<td>1lts</td>
<td>1ppt</td>
<td>1nahb</td>
<td>2lgs</td>
<td>3cla</td>
</tr>
<tr>
<td>1ars</td>
<td>1ctm</td>
<td>1hne</td>
<td>1mat</td>
<td>1ppn</td>
<td>1xkb</td>
<td>2liv</td>
<td>3cox</td>
</tr>
<tr>
<td>1arb</td>
<td>1ddd</td>
<td>1hoo</td>
<td>1mdc</td>
<td>1rcb</td>
<td>2alp</td>
<td>2mg</td>
<td>3dfr</td>
</tr>
<tr>
<td>1aya</td>
<td>1dhr</td>
<td>1hsl</td>
<td>1min</td>
<td>1rec</td>
<td>2bop</td>
<td>2mm</td>
<td>3ebx</td>
</tr>
<tr>
<td>1ace</td>
<td>1dsb</td>
<td>1hsw</td>
<td>1mup</td>
<td>1rib</td>
<td>2edv</td>
<td>2obs</td>
<td>3gly</td>
</tr>
<tr>
<td>1bbp</td>
<td>1ecp</td>
<td>1hag</td>
<td>1nar</td>
<td>1rop</td>
<td>2epl</td>
<td>2pcd</td>
<td>3mdd</td>
</tr>
<tr>
<td>1bet</td>
<td>1ede</td>
<td>1iae</td>
<td>1nba</td>
<td>1rp</td>
<td>2cpl</td>
<td>2pia</td>
<td>3sdh</td>
</tr>
<tr>
<td>1bta</td>
<td>1enh</td>
<td>1lfc</td>
<td>1ntt</td>
<td>1so1</td>
<td>2cro</td>
<td>2por</td>
<td>4blm</td>
</tr>
<tr>
<td>1bmd</td>
<td>1gba</td>
<td>1grp</td>
<td>1nkq</td>
<td>1sac</td>
<td>2sec</td>
<td>2reb</td>
<td>4enl</td>
</tr>
<tr>
<td>1bhb</td>
<td>1fha</td>
<td>1isu</td>
<td>1oib</td>
<td>1s1t</td>
<td>2cic</td>
<td>2nm</td>
<td>4ox</td>
</tr>
<tr>
<td>1bsa</td>
<td>1fkf</td>
<td>1isc</td>
<td>1omp</td>
<td>1sr</td>
<td>2dkb</td>
<td>2sas</td>
<td>4ger</td>
</tr>
<tr>
<td>1cau</td>
<td>1gdh</td>
<td>1kab</td>
<td>1ppb</td>
<td>1tca</td>
<td>2dni</td>
<td>2scp</td>
<td>4xia</td>
</tr>
<tr>
<td>1crr</td>
<td>1gks</td>
<td>1knt</td>
<td>1pbe</td>
<td>1thv</td>
<td>2end</td>
<td>2sil</td>
<td>5p21</td>
</tr>
<tr>
<td>6fab</td>
<td>6tta</td>
<td>7ccp</td>
<td>7pti</td>
<td>8abp</td>
<td>8fab</td>
<td>8rnx</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 3. Histograms of scores for ‘structure seeks sequence’ searches with the template fold of TIM barrel type (1timA) for the three mathematical forms of 3D–1D compatibility criteria, $S_1$, $S_2$, and $S_3$. Score values are shown on the abscissa and corresponding number of sequences in the library (frequency) on the ordinate. Scores of sequences of structural homologs recognized among the top 10 hits (top 2% level) are shown together with their PDB codes (native sequence 1timA was not in the library whereas 1tpn, its closest sequence homolog, was).

We consider now each value of environmental variable $x$ of the amino acid residues of type $a$ in a protein population as a realization of a random variable $X$, which is subject to a probability distribution function $F_a(x)$. Thus, we obtain 20 random variables with distribution functions $F_a(x), \{a = 1,20\}$ for each amino acid type. The main idea of the method is to rely on their probable distinction.

An environmental variable value of any residue, regardless of its type, taken by random from a protein structure is then homologous proteins. As an example, empirical distributions of solvent accessibility for lysine residues $F_{K}(x)$ and for all residues irrespective to their type $G^*(x)$ are shown in Figure 1.
Fig. 4. Histograms of scores for ‘structure seeks sequence’ searches with the template fold of $\alpha/\beta$ doubly wound type (4fxn) for the three mathematical forms of 3D–1D compatibility criteria, $S_1$, $S_2$ and $S_3$ introduced here. Score values are shown on the abscissa and the corresponding numbers of sequences in library (frequency) on the ordinate. Scores of sequences of structural homologs recognized among the top 10 hits (top 2% level) are shown together with their PDB codes.

We denote by $\{x\}_a$ a sample of the environmental variable values for residues of type $a$ in the representative set of proteins; $\{x\} = \bigcup \{x\}_a$ is the total sample of these values for all residue types (mixture of the samples $\{x\}_a$). According to pattern recognition theory (Fukunaga, 1972) these sets $\{x\}$ and $\{x\}_a$ will sometimes be referred to as learning samples and the representative set of the sequentially non-homologous proteins, used to obtain the empirical probability distributions $F_a(x)$, $\{a = 1,20\}$ and $G^*(x)$ for the learning set of proteins.

The statistical hypotheses with respect to threading
It is assumed that two basic conditions are satisfied for the threading method to be valid:

(i) the environmental variable values of $a>$-type residues in different proteins and protein families are subject to the same distribution law function $F(x)$ and similarly the environmental variable values of residues irrespective of type conform to the single mixed distribution function $G(x)$.
(ii) distributions $F_a(x)$ for various amino acid types are not identical, and at least some of them differ significantly and are distinct from the mixed distribution function $G(x)$.

This means that the chosen environmental variable does not depend on the peculiarities of individual protein architectures but reveals general physical principles of protein folding. At the same time, it reflects the specificity of different amino acid types with respect to their spatial location preferences. As can easily be shown, if these two assumptions concerning the distribution of environmental variables are not valid, the classical implementation of statistical sequence–structure recognition methods is not efficient.

The reliability of the first assumption may be tested by the standard statistical $k$-sample tests of homogeneity on samples $\{x\}^k \{k = 1, K\}$ of environmental variable values for various $K$ proteins or protein families. Fulfillment of homogeneity is required not only for $\{x\}^k \{k = 1, K\}$ but also for each amino acid type $a$ (i.e. for samples $\{x\}_a^k \{k = 1, K\}$). Note that $\{x\}_a^k =
The second condition is satisfied when the homogeneity hypothesis of two samples \( \{ x \} \) and \( \{ x \}_a \) is rejected at least for some amino acid types (Sunyaev et al., 1995).

Given an amino acid sequence threaded through a known tertiary structure, i.e. an alignment of the amino acid sequence and the environmental profile to be tested for compatibility, we obtain 20 samples \( \{ y \}_a \) of environmental variable values located at profile sites (positions) corresponding to the amino acids of the \( a \) in the sequence. Namely, those values of the environmental variable profile are gathered into the sample \( y_a \) that corresponds to residues of amino acid type \( a \) in the aligned sequence. Hence these sets of environmental variable values \( \{ y \}_a \) \((a = 1,2,\ldots)\) will be referred to as query (testing) samples.

**Formulation of the statistical hypotheses.**

If a template structure happened to be native to query sequence and their alignment is correct, each of the sample values \( \{ y \}_a \) adheres to the corresponding probability distribution \( F_x(x) \) (hypothesis \( H_0 \)). In the case of a non-native structure or inappropriate alignment we get random superposition (threading) of the sequence residues and the environmental profile entries. Therefore, every query sample \( \{ y \}_a \) may be considered to be drawn from the single mixed distribution \( G(x) \) (hypothesis \( H_1 \)). In other words, to make the inference of the sequence–structure compatibility, hypothesis \( H_0 \) that every query sample \( \{ y \}_a \) adheres to the distribution \( F_x(x) \) for residue type \( a \) should be tested against the alternative \( H_1 \) that every sample \( \{ y \}_a \) adheres to the distribution \( G(x) \). The problem is shown schematically in Figure 2. We consider the statistical test of the hypothesis \( H_0 \) against \( H_1 \) to be the 3D–1D compatibility criterion.

**Mathematical forms of 3D–1D compatibility criteria**

In this section we present three mathematical forms of 3D–1D compatibility criteria as decision rule functions for corresponding statistical hypothesis testing (their more detailed derivation is given in an Appendix). The first two criteria are based on the estimation of the probability density of the environmental variable and the third criterion is based on a non-parametric statistic. Experimental comparison of the criteria is presented in the Results section.

According to the Neiman–Pearson lemma (Kendall and Stuart, 1977), the best decision rule for choosing between two statistical distributions (to test a simple hypothesis against a simple alternative) is the decision rule based upon the log likelihood ratio:

\[
\ln \frac{L(\{ y \}_a | H_0)}{L(\{ y \}_a | H_1)} = \ln \frac{f_x(\{ y \}_a)}{g(\{ y \}_a)},
\]

where \( f_x(\{ y \}_a) \) and \( g(\{ y \}_a) \) are probability densities of the distribution functions \( F_x(x) \) and \( G(x) \), respectively. These densities serve as likelihood functions under two hypotheses, \( H_0 \) and \( H_1 \), and in our case they are unknown. However, \( f_x(\{ y \}_a) \) and \( g(\{ y \}_a) \) can be estimated by the learning samples \( \{ x \}_a \) and \( \{ x \} \). Here we use two different probability density estimators: histogram and Parzen estimator.

**Criterion based on histogram probability density estimator**

The easiest way to estimate probability density is provided by the histogram estimator. It has the advantage of simplicity but the resulting function is not continuous. This approach requires the use of value intervals (bins), and the probability density is estimated by the set of frequencies in histogram bins. We used non-equal bins to distribute uniformly the available data. Histogram-estimated solvent accessibility probability densities for lysine and for all amino acid residues are shown in Figure 1.

As shown in Appendix A, the log likelihood ratio criterion with histogram density estimator is expressed by

\[
S_i = \sum_a \sum_i n_i \ln \left( \frac{p_i^a}{q^a} \right),
\]

where \( p_i^a \) is the frequency of values \( a \) in each histogram bin \( i \) (value interval) and analogously the probability density \( g(\{ y \}_a) \) is approximated by frequencies \( q^a \) of values \( x \) in each histogram bin, and \( n_i \) is the number of query sample values \( y_a \) in the

---

**Table II. Correct top hits in ‘structure seeks sequence’ searches**

<table>
<thead>
<tr>
<th>Template structure</th>
<th>Template fold</th>
<th>Histogram est. criterion</th>
<th>Parzen est. criterion</th>
<th>Non-parametric criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hits</td>
<td>Rank</td>
<td>Z-Score</td>
</tr>
<tr>
<td>Triose-phosphate isomerase</td>
<td>TIM barrel</td>
<td>1tpbh</td>
<td>0</td>
<td>7.17</td>
</tr>
<tr>
<td>1timA</td>
<td>Doubly wound</td>
<td>4fxn</td>
<td>0</td>
<td>8.69</td>
</tr>
<tr>
<td>Flavodoxin</td>
<td>α/β</td>
<td>0.9</td>
<td>2fx2</td>
<td>1</td>
</tr>
<tr>
<td>4fxn</td>
<td>2fcr?</td>
<td>2</td>
<td>3.06</td>
<td>2fcr?</td>
</tr>
<tr>
<td>3cly</td>
<td>3cd4</td>
<td>6</td>
<td>2.16</td>
<td>3cd4</td>
</tr>
<tr>
<td>CD4</td>
<td>Greek key</td>
<td>3cd4</td>
<td>0</td>
<td>8.44</td>
</tr>
<tr>
<td>3cd4</td>
<td>immuno-</td>
<td>3cd4</td>
<td>0</td>
<td>8.44</td>
</tr>
<tr>
<td>globulin</td>
<td>8fab?</td>
<td>5</td>
<td>2.27</td>
<td>8fab?</td>
</tr>
</tbody>
</table>

Data in the table concern correct structure–sequence matches obtained among top 10 hits (top 2%) by each of three criteria \( S_1 \), \( S_2 \) and \( S_3 \). Nine ‘structure seeks sequence’ searches with gapped alignment were performed through the library of 476 non-homologous amino acid sequences (see text). The first two columns indicate template structures with PDB code and their folding type. For each criterion the next three columns refer to (i) PDB codes of recognized alignments; (ii) Hits Rank Z-Score column indicate template structures with PDB code and their folding type. For each criterion the next three columns refer to (i) PDB codes of recognized alignments; (ii) Hits Rank Z-Score column indicate template structures with PDB code and their folding type. For each criterion the next three columns refer to (i) PDB codes of recognized alignments; (ii) Hits Rank Z-Score column indicate template structures with PDB code and their folding type. For each criterion the next three columns refer to (i) PDB codes of recognized alignments; (ii) Hits Rank Z-Score column indicate template structures with PDB code and their folding type. For each criterion the next three columns refer to (i) PDB codes of recognized alignments; (ii) Hits Rank Z-Score column indicate template structures with PDB code and their folding type.
i-th bin of the histogram, which is counted for every bin and every amino acid type $a$ (see Appendix A).

It is easily seen that this criterion can be also expressed as the sum over all alignment positions:

$$S_1 = \sum \ln \frac{p_{\hat{y} a j}^{x \cdot h a}}{q}$$  \hspace{1cm} (5)

It is worth mentioning that this form of the log likelihood ratio decision rule function is identical with the mean force potential energy (1) derived with the assumption of the Boltzmann-like distribution of an environmental variable for the set of known protein structures (Jones et al., 1992; Sippl and Weitkuss, 1992; Bryant and Lawrence, 1993; Kocher et al., 1994). Here, the 3D–1D compatibility criterion is introduced as the statistical hypothesis test decision rule without any additional physical assumptions except the two basic ones mentioned above, which could be easily tested a priori (Sunyaev et al., 1995).

**Criterion based on Parzen probability density estimator**

A more sophisticated way to estimate a probability density is provided by the Parzen estimator (Fukunaga, 1972), which gives higher accuracy and better convergence of approximated density. The Parzen estimator is a continuous function and approaches the probability density itself. The type of Parzen estimator we used here better approaches probability densities close to a normal distribution, but the asymptotic behavior of this estimator does not depend upon the actual distribution function. Parzen probability density estimates for lysine residues and for all residues irrespective of their type are shown in Figure 1.

As demonstrated in Appendix B, the log likelihood ratio estimate in this case is

$$S_2 = \sum \ln \frac{P(y_{a j} \mid N, \{x\}, h)}{P(y_{a j} \mid N, \{x\}, h)}$$

\hspace{1cm} (6)

where the sum is carried out over all alignment positions, a normal-like probability density Parzen estimator $P(y_{a j} \mid N, \{x\}, h)$ is defined in Appendix B. $N_{a}$ is the number of elements in the learning sample $\{x\}_a$ of environmental variable values of amino acid residues of $a$-type, $N$ is the number of elements in the learning sample $\{x\}$ of environmental variable values irrespective of the amino acid type, $y_{a j}$ is the $j$-th element of the testing sample $\{y\}_a$ and $h$ and $h_{a}$ are parameters (see Appendix B).

Both of these compatibility criteria $S_1$ and $S_2$ have the advantage of being additive with respect to alignment positions, which makes it possible to perform an alignment procedure by the ordinary dynamic programming approach (Needleman and Wunsch, 1970).

**Criterion based on the non-parametric statistic**

Here, we again consider the protein fold recognition problem to be a problem of testing statistical hypotheses for the distribution function of an environmental variable in the sequence–structure match or mismatch. Both distributions $F_a(x)$ ($a = 1, 20$) and $G(x)$ are unknown but they are represented by learning samples. Criteria $S_1$ and $S_2$ described above were derived by estimation of the probability density of environmental variables under both hypotheses $H_0$ and $H_1$. However, the problem of a probability density estimation may be simpler for a function of environmental variables than for the environmental variables themselves. The log likelihood ratio estimator for such a function (statistic) may serve as a criterion for a hypothesis test and thus as a protein 3D–1D compatibility criterion. Since the distribution function of an environmental variable cannot be a priori assigned to any parametric family, we use here a non-parametric statistic (function of environmental variables). Statistical criteria based on the empirical distribution function or ranks are now widely used in the problem of testing the hypothesis of two or more samples’ homogeneity (Kendall and Stuart, 1977).

Below we first introduce a non-parametric statistic that is a function of three data samples: query sample $\{y\}_a$ and two learning samples $\{x\}_1$ and $\{x\}_2$. This three-sample statistic incorporates the Wilcoxon–Mann–Whitney two-sample statistic and thus it has similar asymptotic behavior:

$$V_a = \sqrt{n_a} \sum_{j=1}^{n_a} [F_a^5(y_{a j}) + G^*(y_{a j})]$$

where all included parameters (mean values and variances estimates of statistic $V_a$ under hypotheses $H_0$ and $H_1$) are described in Appendix C, may serve as the decision rule function. This decision rule better recognizes a shift in the distribution’s median than a difference in the shape of the distribution function.

Unfortunately, this criterion, in contrast to $S_1$ and $S_2$, is not additive but, as shown in Appendix C, there exists [Equation (22)] an additive approximation $S'_1$ of $S_1$, which makes it possible to implement a powerful dynamic programming algorithm to find the best sequence–structure alignment. Note that in spite of its appearance, this expression is not computationally difficult.

**Testing 3D–1D compatibility criteria and their combination**

The formalization of the problem of protein 3D–1D compatibility as a problem of statistical hypothesis testing involved the development of several mathematical forms of compatibility criteria. Two questions arise about these mathematical forms. First, are all of the three criteria comparable or is one of them irrespective of the amino acid type, $y_{a j}$ the $j$-th element of the testing sample $\{y\}_a$ and $h$ and $h_{a}$ are parameters (see Appendix B).

Both of these compatibility criteria $S_1$ and $S_2$ have the advantage of being additive with respect to alignment positions, which makes it possible to perform an alignment procedure by the ordinary dynamic programming approach (Needleman and Wunsch, 1970).

Here, we again consider the protein fold recognition problem to be a problem of testing statistical hypotheses for the distribution function of an environmental variable in the sequence–structure match or mismatch. Both distributions $F_a(x)$ ($a = 1, 20$) and $G(x)$ are unknown but they are represented by learning samples. Criteria $S_1$ and $S_2$ described above were derived by estimation of the probability density of environmental variables under both hypotheses $H_0$ and $H_1$. However, the problem of a probability density estimation may be simpler for a function of environmental variables than for the environmental variables themselves. The log likelihood ratio estimator for such a function (statistic) may serve as a criterion for a hypothesis test and thus as a protein 3D–1D compatibility criterion. Since the distribution function of an environmental variable cannot be a priori assigned to any parametric family, we use here a non-parametric statistic (function of environmental variables). Statistical criteria based on the empirical distribution function or ranks are now widely used in the problem of testing the hypothesis of two or more samples’ homogeneity (Kendall and Stuart, 1977).

Below we first introduce a non-parametric statistic that is a function of three data samples: query sample $\{y\}_a$ and two learning samples $\{x\}_1$ and $\{x\}_2$. This three-sample statistic incorporates the Wilcoxon–Mann–Whitney two-sample statistic and thus it has similar asymptotic behavior:

$$V_a = \sqrt{n_a} \sum_{j=1}^{n_a} [F_a^5(y_{a j}) + G^*(y_{a j})]$$ \hspace{1cm} (7)

Again, the log likelihood ratio of $\{V_a\}$:

$$S_3 = \sum_a \left[ \frac{(V_a - E(V_a))^2}{2 \text{Var}(V_a)} - \frac{(V_a - E(V_a))^2}{2 \text{Var}(V_a)} \right]$$ \hspace{1cm} (8)

where all included parameters (mean values and variances estimates of statistic $V_a$ under hypotheses $H_0$ and $H_1$) are described in Appendix C, may serve as the decision rule function. This decision rule better recognizes a shift in the distribution’s median than a difference in the shape of the distribution function.

Unfortunately, this criterion, in contrast to $S_1$ and $S_2$, is not additive but, as shown in Appendix C, there exists [Equation (22)] an additive approximation $S'_1$ of $S_1$, which makes it possible to implement a powerful dynamic programming algorithm to find the best sequence–structure alignment. Note that in spite of its appearance, this expression is not computationally difficult.

Testing 3D–1D compatibility criteria and their combination

The formalization of the problem of protein 3D–1D compatibility as a problem of statistical hypothesis testing involved the development of several mathematical forms of compatibility criteria. Two questions arise about these mathematical forms. First, are all of the three criteria comparable or is one of them much better than the others? Second, what is the benefit if any in applying several criteria or all of them in being able to recognize the same cases of structural homology by threading procedure? In other words, is a combined application of several criteria more efficient? To answer these questions, the efficiencies of compatibility criteria should be tested and compared.

Practical implementation of threading methods implies either a search through a structural library with a query sequence for a compatible fold or the inverse search through a sequence library for sequences best fitting a template folding motif. Let us denote the first variant as a ‘sequence seeks structure’ search and the second as a ‘structure seeks sequence’ search. Threading methods are confined to a search for sequence–structure matches even in the absence of sequence homology where fold similarity could not be obtained by ordinary sequence–sequence alignment. Many authors have reported this kind of search result obtained by the combined use of
several criteria which differ in their environmental variables. We present below experimental results for the combined application of the three mathematical forms of the 3D–1D criterion to a 'structure seeks sequence' search with a single environmental variable.

The combined applications of statistical criteria have nothing in common with the combined application of environmental variables. Obviously, it cannot be considered as a combination of different energy terms. Simply, because statistics of criteria are random values and different criteria are more sensitive in different cases, the reliability of the statistical inference of fold assignment may be increased if several criteria are used.

There is no convention yet in the literature on fold recognition methods as to the evaluation of search results (Lemer et al., 1995). Since we work in the framework of statistical mathematics, it is natural to evaluate criteria efficiency according to their empirical power (frequency of the second kind error) while an empirical significance level (frequency of the first kind error) is fixed (Kendall and Stuart, 1977). However, since the estimator of empirical power for the case considered is difficult, we used a related value: the number of correctly recognized sequences in the top 2% of overall scores. For our sequence library size, this implies that the number of correctly recognized sequences among the top 10 hits was counted.

Results and discussion

We present here results of the sequence library search for primary structures compatible with a template fold to compare the efficiencies of newly introduced $S_3$ and $S_3^*$ and the usual $S_1$ mathematical forms of 3D–1D compatibility criteria.

Data and materials

In this section we describe the environmental variable, learning set of structures, template folds and the query sequence library necessary for our 'structure seeks sequence' search experiment.

Since various environmental variables are subject to different distributions, results of the comparison analysis of the mathematical forms of 3D–1D compatibility criteria depend strongly upon the chosen variable. Here, we used residue accessibility as a sole environmental variable. Residue accessibility is a widely used (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Kocher et al., 1994) environmental variable reflecting residue hydrophobicity and, therefore, results can be interpreted from a physical point of view. The accessibility of a residue $R$ is defined here as exposure to solvent with the surface area normalized to its average value in the tripeptide Gly–R–Gly. The accessible surface area was computed by a numerical algorithm based on the uniform allocation of points on the surface of each atom of a residue. The application of this variable makes it possible to implement gapped alignment by a dynamic programming algorithm because, unlike pair contacts, residue accessibility is a function of only one amino acid position.

Our learning set of non-homologous proteins with known three-dimensional structure used for the estimation of the accessibility distribution is shown in Table I. This set of 167 protein structures was extracted from the PDBSELECT database (Hobolm and Sander, 1994) (sequence identity threshold 25%). Only structures with resolution better than 2.5 Å were included. We used the same learning set to estimate the distribution function $G(x)$ [probability density $g(x)$] and the 20 distribution functions $F_{ij}(x)$ [probability densities $f_{ij}(x)$] in all tests.

The testing of the efficiency of the choice of template folds for the criteria $S_1$, $S_2$ and $S_3$ is based on the following idea. Really challenging tests for threading methods can be provided in those cases of structural homology where sequence homology and the common functions of query proteins are low or non-existent. There are no methods up to now except threading to detect fold similarities in such cases based only on the amino acid sequence. A significant proportion of proteins with experimentally determined spatial structure adopt one of several superfolds (Orengo et al., 1994). For the 'structure seeks sequence' search, we have chosen structures of three highly populated superfolds: flavodoxin (PDB code 4fxn) of the $\alpha/\beta$ doubly wound fold, triosephosphate isomerase (PDB code 1tim) of the TIM barrel fold and CD4 (Brookhaven code 3cd4) of the Greek key immunoglobulin fold. For each of these superfolds, several examples of proteins are known which do not share either detectable sequence homology or common function but do share a significant structural relationship.

In total, the query sequence library included 476 primary structures of proteins with known folding type from the Brookhaven Protein Data Bank (release January 19, 1995). Only sequences from the PDBSELECT database (Hobolm and Sander, 1994) non-homologous at the threshold of 35% were selected. Above this threshold, fold similarity can be readily concluded on the basis of sequence alignment alone. We added to those selected all sequences of functionally and sequentially unrelated proteins adopting the three chosen template folds mentioned by Orengo et al. (1994).

Search results

For each of the structural families, three searches were performed with compatibility criteria $S_1$, $S_2$ and $S_3$. Criteria values were normalized to the square root of an alignment length to correct for the length of each sequence. Normalized criteria values were considered as scores for sequences.

Histograms of scores (decision rule values) obtained in the 'structure seeks sequence' search for TIM barrel template fold (1tim) are presented in Figure 3. Histograms for $\alpha/\beta$ doubly wound fold (4fxn) and Greek key immunoglobulin fold (3cd4) are shown in Figures 4 and 5, respectively. Scores of protein sequences correctly recognized among the top 10 hits (top 2% level) are indicated in each histogram. PDB codes of these proteins, their ranking and Z-score values (number of standard deviations above mean score) are listed in Table II. We determined fold similarity to verify what sequence–structure matches are correct by the 3d_ali (Pascarella and Argos, 1992) and the FSSP (Holm and Sander, 1994) databases. Results of the search with the TIM barrel (1tim) template fold show distinct behavior for the three criteria. Note that the sequence library did not include the amino acid sequence of 1tim. All three criteria correctly recognized the sequence of 1pbl, which is a close homolog of 1tim. However, sequences of 1dhr, 1ald and 4enl were detected only by the non-parametric criterion $S_3$. At the same time, criterion $S_1$ was the only criterion which identified the amino acid sequence of 1nar.

In the test for $\alpha/\beta$ doubly wound fold (template structure 4fxn), all three criteria correctly identified the native sequence and sequences of 2fx2, 5p21 and 2fcr. Only the non-parametric criterion $S_3$ was able to recognize the sequence of 3akd but the sequence of 3ch was found by criterion $S_1$ alone. 3ch occupied the twelfth position in the ranking list for criterion $S_3$ but this is not shown in Figure 4 as only correct matches among the top 10 hits are shown there.
In the case of the Greek key immunoglobulin fold (3c44), the same structural homologs were detected by all three criteria. As seen, in many cases only one criterion correctly identifies a particular sequence–structure match. This may be explained by the different statistical concepts underlying the individual criteria. We note that almost all false positives obtained by criteria $S_1$, $S_2$ and $S_3$ differ. This means that if an amino acid sequence is recognized as compatible with the same template by all three criteria, one can more reliably infer the sequence–structure relationship. Therefore, the present brief analysis of search results reveals that use of several criteria can provide additional benefits.

Obviously, the results could change slightly if another environmental variable or a set of environmental variables were used. Probably for some variables, one mathematical form of the 3D–1D compatibility criterion is preferable. In other words, investigations of the mathematical forms of scoring functions can improve the results of various threading methods.

The formalization of the sequence–structure recognition problem presented here enables us to derive and test new criteria as well as to develop the methodology to evaluate structure recognition abilities of various environmental variables.

Acknowledgements
We thank M.S.Gelfand, M.A.Roytberg, D.A.Dolgikh, F.Eisenhaber and J.Selig for useful discussions. The work was supported by the Russian Fund of Fundamental Research (grant 94-07-13186).

References

Received September 20, 1996; revised December 18, 1996; accepted January 14, 1997

Appendix A

Probability densities $f_y(x)$ for each residue type $y$ can be approximated by corresponding histograms by counting frequencies $p_y^i$ of values $x_i$ in each histogram bin $i$. Analogously, the probability density $g(x)$ is approximated by frequencies $q_i$ of values $x$ in each histogram bin (division). Unequally spaced bins of the histogram were chosen to obtain equal frequencies $q_i$ for all bins. Thus, data were uniformly distributed throughout the bins. Since there is no theoretical recipe for determination of the optimal number of bins, we have chosen eight bin histograms.

Now let us describe a recognition step with our statistical approach. To test hypothesis $H_0$ against the alternative $H_1$, the number $n^k$ of query sample values $y^k$ in the $i$-th bin of the histogram is counted for every bin and every amino acid type $a$. The log likelihood ratio is then estimated for numbers $n^k$ instead of sample values $y^k$. It is clear that the $n^k$ are polynomially distributed values. The likelihood function estimator $\hat{L}(n^k|H_0)$ of $\{n^k\}$ under $H_0$ is given by

$$\hat{L}(n^k|H_0) = \prod_a \frac{n^k_a!}{n^k_a} \prod_i (p_i)^{n^k_a}$$

and if $H_1$ is true by

$$\hat{L}(n^k|H_1) = \prod_a \frac{n^k_a!}{n^k_a} \prod_i (q_i)^{n^k_a}$$

where $n_a = \sum n^k_a$ is the number of query sample values $y_a$ for amino acid $a$.

The log likelihood ratio criterion is then expressed by Equation 4.

Appendix B

The Parzen estimator (Fukunaga, 1972) is known to be a consistent estimator of the probability density function. In contrast to the histogram estimator, the Parzen estimator is a continuous function and approaches the probability density of values $y_a$ itself. The Parzen probability density estimate at a point $y_a$ is given by

$$\hat{P}(y_a | N, \{x\}, h) = \frac{1}{N} \sum_i \frac{1}{\sqrt{2\pi h}} \exp \left( \frac{-(y_a-x_i)^2}{2h^2} \right).$$

Therefore, the estimator for the likelihood functions for the query sample $\{y_a\}$ under hypotheses $H_0$ and $H_1$ can be expressed as

$$\hat{L}(\{y_a\}|H_0) = \prod_a \prod_j \hat{P}(y_j^a | N_a, \{x\}, h_a),$$

$$\hat{L}(\{y_a\}|H_1) = \prod_a \prod_j \hat{P}(y_j^a | N, \{x\}, h),$$

where $N_a$ is the number of elements in the learning sample $\{x\}$ of environmental variable values of amino acid residues of type $a$, $N$ is the number of elements in the learning sample $\{x\}$ of environmental variable values irrespective of amino acid type, $y_j^a$ is the $j$-th element of the testing sample $\{y\}$, parameters $h$ and $h_a$ were taken as $h = N^{-0.2}$ and $h_a = N_a^{-0.2}$ (Fukunaga, 1972) and summing index $l$ in Equation 11 ranges over elements of a corresponding learning sample, namely, $l = 1,N_a$ for Equation 12 and $l = 1,N$ for Equation 13.

The log likelihood ratio estimate is then given by Equation 6, where summing is carried out over all alignment positions.
Appendix C

Wilcoxon–Mann–Whitney statistics for two samples \{a\}, \(n_a = |\{a\}|\) and \{v\}, \(n_v = |\{v\}|\), with empirical distribution functions \(P^*(a)\) and \(Q^*(v)\), respectively, may be expressed as

\[
W = n_a \sum_{i=1}^{n_v} P^*(v_i).
\]

Note that in Equation 14 the empirical distribution function value \(P^*(a)\) of sample \{a\} is calculated at each point \(v_i\) of sample \{v\}. It can be shown that the distribution of statistic \(W\) is asymptotically normal (Hettmansperger, 1984).

We want to test the hypothesis \(H_0\) for the distribution function of \{y\}a coinciding with that of \{x\}a against the useful for implementing a dynamic programming algorithm to find the best sequence–structure alignment. Note that, in spite of its appearance, this expression is not computationally difficult.

All four parameters are unknown. However, their estimates can be obtained, since the empirical distributions \(F_a^*(x)\) and \(G_a^*(x)\) are computed from representative learning samples. Hence we can compute estimates of the likelihood of \(V_a\) for both hypotheses \(H_0\) and \(H_1\). Assuming the independence of \(V_a\) for different amino acid types, we obtain

\[
\hat{L}(|\{V_a\}| \ H_0) = \prod_a N(\hat{E}_0(V_a), \hat{Var}_0(V_a));
\]
\[
\hat{L}(|\{V_a\}| \ H_1) = \prod_a N(\hat{E}_1(V_a), \hat{Var}_1(V_a)),
\]

where

\[
N(\hat{E}_a(V_a), \hat{Var}_a(V_a)) = \frac{1}{\sqrt{2\pi \hat{Var}_a(V_a)}} \exp\left(-\frac{(V_a - \hat{E}_a(V_a))^2}{2\hat{Var}_a(V_a)}\right)
\]

is the normal probability density function.

The log likelihood ratio of \{\{V_a\}\} given by Equation 8 may serve as the decision rule function. We drop a negligible constant in Equation 8.