Protein Sequence-Structure Alignment Based on Site-Alignment Probabilities

Sanzo Miyazawa
miyazawa@smlab.sci.gunma-u.ac.jp
Faculty of Technology, Gunma University, Kiryu, Gunma 376, Japan

Abstract

A protein sequence-structure alignment method for database searches is examined on how effectively this method together with a simple scoring function previously developed can identify compatibilities between sequences and structures of proteins. The scoring function consists of pairwise contact energies, repulsive packing potentials of residues for overly dense arrangement and short-range potentials for secondary structures. Pairwise contact interactions in a sequence-structure alignment are evaluated in a mean field approximation on the basis of probabilities of site pairs to be aligned. Gap penalties are assumed to be proportional to the number of contacts at each residue position, and as a result gaps will be more frequently placed on protein surfaces than in cores. In addition to minimum energy alignments, we use probability alignments made by successively aligning site pairs in order by pairwise alignment probabilities. Results show that the present energy function and alignment method can detect well both folds compatible with a given sequence and, inversely, sequences compatible with a given fold. Probability alignments consisting of most reliable site pairs only can yield small root mean square deviations, and including less reliable pairs increases the deviations. Remarkably, by this method some individual sequence-structure pairs are detected having only 5–20% sequence identity.

Keywords: empirical potentials, inverse protein folding, protein fold recognition, sequence-structure alignment, threading and inverse threading with gaps and insertions

1 Introduction

A number of works [1, 2, 4, 6, 7, 9, 12, 13, 21, 26, 27, 30] indicate that simple empirical potentials [16, 18, 19, 20, 23, 25, 28, 29] without atomic details may be sufficient to determine overall folds, although some limitation to pairwise potentials is indicated [15]. Many types of empirical energy functions were tested for their abilities to distinguish correct from incorrect folds, which were generated by threading sequences into the structures of other proteins at all possible positions without gaps [2, 7, 30] or by relaxing native structures with molecular dynamics or other methods [26, 27]. Such a method to generate alternative folds is appropriate, because a simple comparison of conformational energy values between different sequences is meaningless. However, measuring compatibilities between sequences and structures is neither simple nor easy.

In order to allow gaps in sequence-structure alignments, two types of problems must be overcome. One must take into account not only the conformational energies of folds but also the sequence dependencies of the whole ensemble of protein conformations in order to evaluate the relative stabilities of sequences or alignments [21]. Here, the stabilities of structures are assumed as a primary requirement for compatibilities between sequences and structures.

The second problem is how to evaluate multi-body interactions among residues. The frozen approximation, in which the residue's environment is evaluated for the native sequence rather than the trial sequence, was used [1, 6, 14]. However, in principle, the assumption of the native structure environment is inappropriate for evaluating interactions among residues for extremely divergent proteins.

A double dynamic programming method was used [10] as an approximate method to take account of pairwise potentials. A search algorithm for finding exact global optimum threadings into protein core segments connected by variable loops, was devised [11] for pairwise interaction potentials; gaps are allowed only into the variable loops.

Here, we propose a method in which pairwise contact interactions between residues are evaluated in a mean field approximation on the basis of the probabilities of site pairs being aligned, and examine how effectively this method together with a simple energy potential can identify compatibilities between sequences and structures of proteins; also see [22]. Gaps are allowed anywhere in a protein, with structure-dependent gap penalties. To obtain the self-consistent values of alignment probabilities of site pairs, an iterative method is employed. In addition to the minimum energy alignment, an alignment termed a probability alignment [17] is also made by successively assigning aligned site pairs by their alignment probabilities. A scoring function used is one previously developed and shown successfully to identify native structures for sequences and inversely native sequences for structures in threadings without gaps [21].

2 Methods

2.1 A Statistical Ensemble of Sequence-Structure Alignments

An example of a specific sequence−structure alignment A is

$$A \equiv \begin{bmatrix} \cdots & - & i_3 & i_4 & i_5 & i_6 & \cdots \\ \cdots & s_2 & s_3 & - & - & s_4 & \cdots \end{bmatrix} \tag{1}$$

where "−" means a deletion, and s_p is the conformational state of the pth residue in a given structure, and i_q means the qth residue of type i_q in a sequence that is threaded into the structure.

The conditional probability of an alignment A for a given structure $\{s_p\}$ is represented [22] as

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \frac{1}{\mathcal{Z}} \exp[-\beta \mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{2}$$

$$\mathcal{Z} = \sum_{A} \exp[-\beta \mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{3}$$

where β is equal to $1/(kT)$ and $\mathcal{Z}$ is a partition function for alignments. The energy score $\mathcal{E}(\{s_p\}|\{i_q\},A)$ of an alignment A for a given structure $\{s_p\}$ is defined as

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \Delta E^{\mathrm{conf}}(\{s_p\}|\{i_q\},A) + n_r^{\mathrm{aligned}} \mathcal{E}_0 + \sum_{\text{all gaps in } A} \mathcal{W} \tag{4}$$

n_r^{aligned} is the number of aligned site pairs in the alignment A. ΔE^{conf} is the alignment energy [21] of a structure $\{s_p\}$ for the alignment A whose zero energy state is adjusted to make its unweighted average over typical native structures equal to zero. Here, it consists of pairwise contact energies, [16, 18, 19] repulsive packing potentials for residues, [18] and short-range potentials for secondary structures; [20] the contact energies [19] divided by $\alpha' \simeq 0.263$ are used as the values of contact energies in the present calculations. $\mathcal{E}_0$ is a favorable energy for a site match and $\mathcal{W}$ is gap penalties.

2.2 Pairwise Interactions Evaluated in a Mean Field Approximation

In general, an energy scoring function can be represented in a sum of an intrinsic energy $\mathcal{E}_0$, a one-body $\mathcal{E}_1$, two-body $\mathcal{E}_2$, and higher orders of interaction.

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \sum_{(p,q)\in A} \mathcal{E}(\{s_p\}|i_q,A) + \sum_{\text{all gaps in } A} \mathcal{W} \tag{5}$$

$$\mathcal{E}(\{s_p\}|i_q, A) \equiv \mathcal{E}_0 + \mathcal{E}_1(s_p|i_q) + \frac{1}{2}\sum_{(p',q')\in A}\mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'}) + \cdots \tag{6}$$

Therefore, it is difficult to calculate the most probable alignment and the partition function of Eq. 3. Here, the pairwise interaction energies for alignment A that significantly contributes to the partition function in Eq. 3 are approximated with pairwise energies for amino acid pairs $(i_q, i_{q'})$ located at neighboring sites (p, p') in structure with alignment probabilities $\mathcal{P}(p', q')$ of structure-sequence site pairs (p', q').

$$\sum_{(p',q')\in A}\mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'}) \approx \sum_{p'}\sum_{q'}\mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'})\mathcal{P}(p', q') \tag{7}$$

$\mathcal{P}(p, q)$ and the probabilities for deletions $(p, -)$ and $(-, q)$ are calculated from

$$\mathcal{P}(p,q) = \frac{1}{\mathcal{Z}}\sum_{A \text{ with } (p,q)}\exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\}, A)] \tag{8}$$

$$\simeq \frac{1}{\mathcal{Z}}\mathcal{Z}_{p-1,q-1}\ \exp[-\beta\mathcal{E}(\{s_p\}|i_q, \mathcal{P}(p', q'))]\ \mathcal{Z}'_{p+1,q+1} \tag{9}$$

$$\mathcal{P}(p,-) = 1 - \sum_q \mathcal{P}(p,q) \quad , \quad \mathcal{P}(-,q) = 1 - \sum_p \mathcal{P}(p,q) \tag{10}$$

where $\mathcal{Z}_{p-1,q-1}$ is also a partition function but for aligning the N-terminal, partial sequence from 1 to $(q-1)th$ residues with the N-terminal, partial structure from 1 to $(p-1)th$ residues in the whole structure. $\mathcal{Z}'_{p+1,q+1}$ is a partition function for aligning the C-terminal sequence starting from $(q+1)$th residue with the C-terminal part from $p+1$ to the terminal end in the whole structure. Therefore, the following relation is satisfied; $\mathcal{Z} = \mathcal{Z}_{n_r^{str}, n_r^{seq}} = \mathcal{Z}'_{1,1}$ Such partition functions can be calculated by a transfer matrix method; see Miyazawa [17] for a specific description of this method for alignments. A self-consistent solution for $\mathcal{P}(p, q)$ in Eq. 9 is calculated by an iteration method.

2.3 Alignment Based on Site-Alignment Probabilities

By evaluating the energy score of alignments with the self-consistent alignment probabilities of site pairs (Eq. 7), we can approximately calculate the minimum energy score alignment $A^{\min}$ with a conventional dynamic programming method; $\mathcal{E}(\{s_p\}|\{i_q\}, A^{\min}) \equiv \min_A \mathcal{E}(\{s_p\}|\{i_q\}, A)$.

In addition, we also employ here probability alignments [17] consisting of the most probable site pairs by successively aligning a site pair in order of pairwise alignment probabilities $\mathcal{P}(p, q)$ of Eq. 9. (i) Set p_1 and p_2 to the N-terminal and C-terminal site position of a partial structure to align, and q_1 and q_2 to the N-terminal and C-terminal site position of a sequence segment to align. (ii) If there is a site pair (p, q) such that $\mathcal{P}(p,q) = \max_{p_1\le p'\le p_2, q_1\le q'\le q_2}$ ($\mathcal{P}(p', q') \mid \mathcal{P}(p', q') \ge \mathcal{P}(p', -)$ and $\mathcal{P}(p', q') \ge \mathcal{P}(-, q')$), align them. Otherwise, assign deletions to all sites of $p_1 \le p \le p_2$ and of $q_1 \le q \le q_2$. Then, repeat steps (i) and (ii) to align the remaining segments until all the sites are aligned.

A whole ensemble of sequence-structure alignments can be characterized by such quantities as the minimum energy score, free energy score, and internal energy score. A preliminary test indicates that the capability of recognition of sequence-structure compatibilities seems to be about the same among these three energy scales. In the following, minimum energy scores are employed to judge sequence-structure compatibilities.

2.4 Structure-Dependent Gap Penalties

Here the dependence of residue mutability on residue position [5] is taken into account by setting the gap penalty to be proportional to the number of contacts at each residue position in a protein structure. The number of contacts is utilized here as a simple measure of burial and packing density

of residues. In other words, gaps will tend to be inserted in alignments more often on protein surfaces than in protein cores.

Table 1: Gap parameters used in sequence-structure alignments.

Gap penalty	Value in kT units
$\mathcal{E}_0$	-1.2
Structure deletions from q to q_1	$5.5 + \sum_{p=q}^{q1}(1.05 + 0.43n_p^c)$ in the middle
	$3.25 + \sum_{p=q}^{q1}(0.53 + 0.22n_p^c)$ at termini
n sequence insertions between q and $q+1$	$5.5 + n(1.05 + 0.43(1 + (n_q^c + n_{q+1}^c)/2))$ in the middle
	$3.25 + n(0.53 + 0.22(1 + n_{terminal}^c))$ at termini
The upper limits for gap penalty	60.9 for gaps in the middle
	30.45 for terminal gaps
Relative temperature, $1/\beta$	2.6

n_p^c is the number of residues whose side chain centers are within $6.5\mathring{A}$ from the side chain center of the pth residue, excluding neighboring residues along a sequence.

The values of gap parameters are listed in Table 1. The present values of gap parameters are adjusted to yield similar fractions of aligned residues in minimum energy alignments for homologous protein pairs to those in conventional sequence alignments. The relative temperature ($1/\beta$) is also adjusted to yield similar fractions of aligned residues in probability alignments for the homologous protein pairs compared to those in probability sequence alignments [17]. The parameter $\mathcal{E}_0$ is chosen in such a way that minimum energy scores for most of the dissimilar protein pairs fall above zero; also there is no clear indication that the minimum energy scores depend linearly on the sequence length.

2.5 Datasets of Protein Structures

Two datasets of protein pairs were prepared; one is a set of homologous protein pairs, and the other is a set of dissimilar protein pairs. Release 1.35 of the SCOP database [24] is used for the classification of protein folds. Only protein classes 1 to 5 corresponding to all α, all β, α/β, $\alpha+\beta$, and multi-domain proteins are used. Proteins whose structures were determined by NMR or with resolution worse than 2.5 Å, lack many atoms or which are shorter than 50 residues are removed. By using the first entries in the protein lists of each superfamily, family or species as protein representatives from each protein fold, the set of 548 homologous protein pairs is made by pairing the protein representatives of families with those of different species within the families. The set of dissimilar protein pairs is made by arbitrarily choosing only every $100th$ or $10th$ pair from the ordered list of all possible pairs of superfamily representatives; 505 or 5041 protein pairs are chosen.

3 Results

3.1 Characteristics of Sequence-Structure Alignments

First, the adequacy of sequence-structure alignments with the present method has been examined by comparing the overall characteristics of sequence - structure alignments to those of conventional sequence alignments (global alignments). Folds of multimeric proteins and domains are evaluated in the multimeric state or within a whole protein even for sequences of monomeric proteins. Dayhoff 250 PAM matrix [3] is used as a scoring matrix for the sequence alignment, but alternatively BLOSUM matrices [8] could have been used. Both the sequence-structure alignments and the conventional sequence alignments give similar aligned fractions of residues for most proteins, indicating the values of $\mathcal{E}_0$ and gap parameters to be appropriate [22].

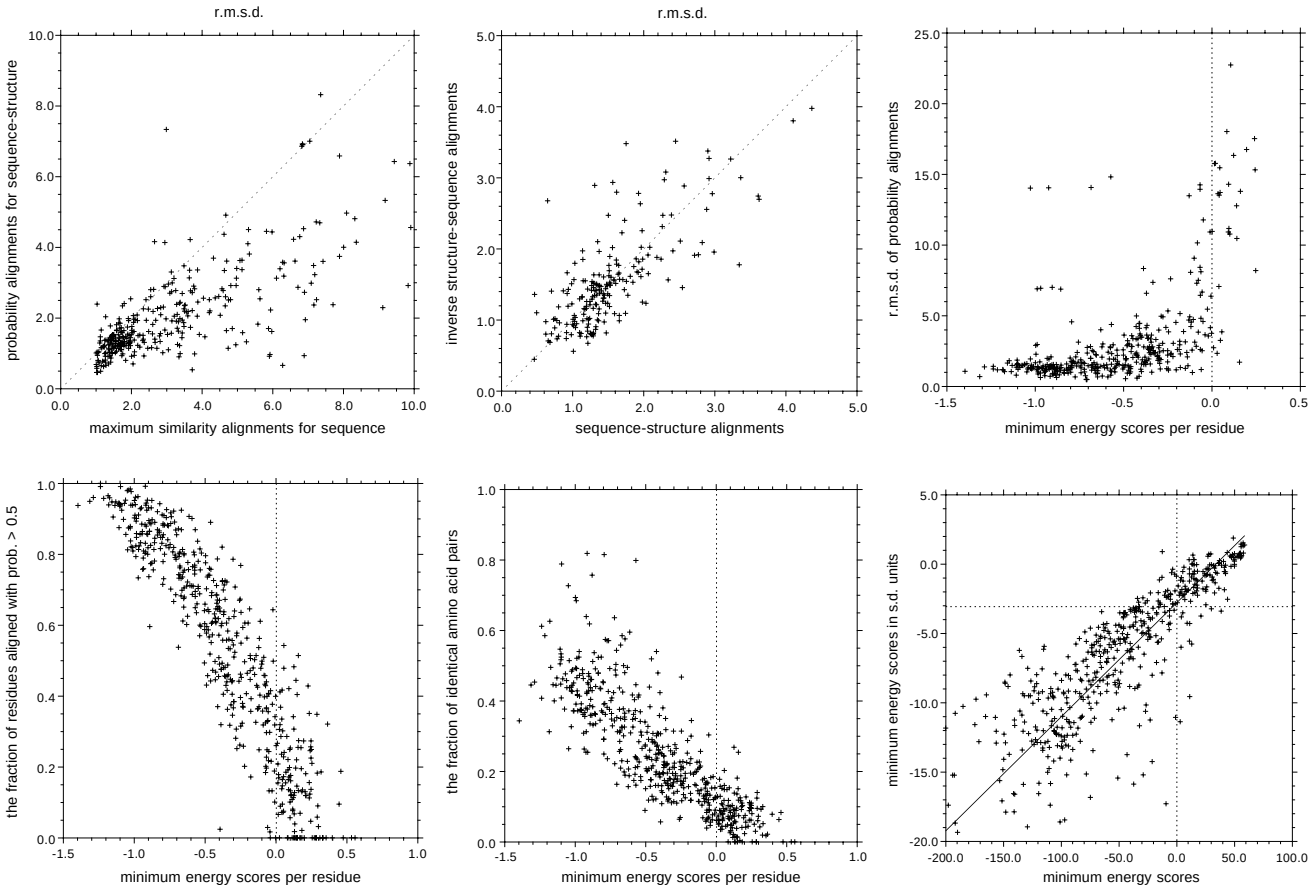

Figure 1: Characteristics of sequence-structure alignments; with the top 3 panels being, from left to right, a - c, and the bottom 3 being, from left to right, d - f.

To further examine the quality of the present sequence-structure alignments, the root mean square deviations (r.m.s.d.) in superpositions of C_α atoms of aligned residue pairs in the sequence-structure alignments are compared in Figure 1a to those from the maximum similarity alignments of sequences. For this purpose, in this figure, 357 homologous protein pairs, which have negative minimum energy scores and positive maximum similarity scores and also whose alignments have aligned residue pairs ≥ 50, are plotted; note that r.m.s.d. for small numbers of superposed C_α pairs may take small values even for dissimilar structures. Significant improvements in the values of r.m.s.d. are shown in this figure. Although these improvements are made partially by choosing only residue pairs most reliably aligned, they also indicate that the quality of the probability alignments of sequence-structure are usually better than those for the corresponding conventional sequence alignments.

As expected, both types of sequence-structure and inverse structure-sequence alignments take similar values for the fraction of aligned residues, for the fraction of identical amino acid pairs, and for the r.m.s.d. of aligned residues; the r.m.s.d. for 216 homologous protein pairs with negative energy scores and with ≥ 50 residues aligned with probabilities ≥ 0.5 are shown in Figure 1b.

It is also useful to know the relationships between minimum energy scores and characteristics of alignments. In Figure 1c, minimum energy scores per residue are plotted against r.m.s.d. in superposition of residues aligned with probabilities ≥ 0.5; it shows only 398 homologous protein pairs with ≥ 50 residues aligned with probabilities ≥ 0.5. Most of the probability alignments whose minimum energy scores fall below zero energy score have r.m.s.d. less than 5 Å. Interesting cases appear if one looks closely at the exceptional protein pairs; they are 1NCX sequence compared with 1TCO-B, 1WDC-C, 1WDC-B, 1LIN, 1CLL, 3CLN, 1OSA, and 4CLN structures in the calmodulin-like

family. There is a helix in the middle of the sequences whose lengths vary among these proteins.

Figure 1d shows dependences of the fractions of residues aligned with probabilities ≥ 0.5 on minimum energy scores per residue, and Figure 1e shows the relationship between minimum energy scores per residue and the fractions of identical amino acid pairs in the probability alignments, for the 548 homologous protein pairs.

3.2 Detection of Homologous Proteins from Dissimilar Proteins

One of the most important questions is how well this energy scale can recognize a compatible pair of structure and sequence, particularly those not found from sequence comparisons. The parameter $\mathcal{E}_0$ is chosen so that there is no clear indication that the minimum energy scores of the dissimilar structure pairs depend linearly on the lengths of proteins and also so that compatible sequence and structure pairs tend to take negative energy scores and incompatible ones positive energy scores. Thus, judgements for compatible sequence-structures may be made on the basis of the values of scores. Alternatively, to judge whether such alignment scores are statistically significant, one may use a z-score that is defined as an alignment score expressed in standard deviation (s.d.) units from the average score for randomized sequences. Figure 1f shows that the present energy scores roughly correlate with the z-scores evaluated from 100 randomized sequences, and that a zero energy score corresponds to about -3 standard deviation units; the correlation coefficient is 0.81. In this manuscript, protein sequence-structure pairs have been judged to be compatible if their energy scores are negative or z-scores are more negative than -3.

As shown in Figure 1e, the present set of homologous protein pairs includes many distantly related protein pairs whose alignments have fractions of identical amino acid pairs below 10 % and therefore which are not identified as compatible sequence-structure pairs. The conventional sequence alignment method cannot detect similarities for all of those homologous protein pairs, either. Table 2 lists the numbers of false positives and false negatives for the present sequence-structure alignment method and for the conventional sequence alignment method on the basis of score and also of z-score. The overall capability to identify homologous protein pairs is slightly better for the conventional sequence method than for the present sequence-structure alignment method, but Table 3 shows that both methods can complement each other to recognize some different homologous protein pairs.

Table 2: Discrimination of homologous protein pairs from dissimilar protein pairs.

False negatives in homologous protein pairs†		False positives in dissimilar protein pairs			Alignment method
with score	with z-score	with score		with z-score	
106/322	108/322	5/505	83/5041	4/505	Sequence-sequence
129/322	147/322	17/505	173/5041	4/505	Sequence-structure
123/322	152/322	24/505	236/5041	7/505	Inverse structure-sequence

†Homologous protein pairs whose maximum similarity alignments include less than 30% identity.

Table 3: Recognition of homologous protein pairs†.

seq.-seq.	seq.-str.		inverse		seq.-seq.	seq.-str.		inverse	
similarity	energy score				similarity	energy z-score			
score	$<$	$\geq$	$<$	≥ 0	z-score	$<$	$\geq$	$<$	≥ -3
> 0	168	48	172	44	> 3	158	56	152	62
≤ 0	25	81	27	79	≤ 3	17	91	18	90

†Homologous protein pairs whose maximum similarity alignments include less than 30% identity.

To establish that those alignments are reasonable, the root mean square deviations of the sequence-structure alignments are examined. To assure that the r.m.s.d. are reliable, only protein pairs having

Table 4: Protein pairs† whose compatibilities are not identified by sequence alignments but by sequence-structure or inverse structure-sequence alignments.

sequence	length	structure	length	sequence-structure probability alignment					sequence-sequence maximum similarity alignment				
				minimum energy score	z-score	identities	# residues† with prob. ≥ 0.5	rmsd ($\AA$)	maximum similarity score	z-score	identities	# aligned residue pairs	rmsd ($\AA$)
1ARB	263	1SGT	223	30.1	−3.2	0.09	83	16.3	−36	−1.3	0.04	44	11.7
1ECF-A:250-469	220	1HMP-A	214	−10.7	−3.1	0.09	88	4.6	−11	1.0	0.14	193	15.3
1NCX	162	2SAS	185	−17.3	−7.1	0.10	85	9.1	−6	1.6	0.14	161	14.5
1PBN	289	1ECP-A	237	−6.5	−4.7	0.08	99	5.4	−25	−0.1	0.02	27	8.0
1PII:1-254	254	1TTQ-A	256	−12.3	0.9	0.09	62	11.8	−22	−0.3	0.03	36	9.2
1PTV-A	297	1YTS	278	−36.2	−9.0	0.11	105	4.9	0	3.3	0.19	260	9.5
1XEL	338	1ENY	268	−3.1	−2.9	0.08	57	10.9	−2	2.6	0.12	189	18.2
1XEL	338	1FDS	282	−20.2	−3.2	0.09	61	2.6	−1	4.0	0.05	54	13.7
2DRI	271	2LBP	346	−26.4	−10.2	0.12	157	7.3	−14	0.2	0.15	211	23.1
2DRI	271	2LIV	344	−37.1	−15.9	0.11	165	8.1	−20	−0.8	0.04	63	17.2
2HVM	273	1NAR	289	−84.2	−5.4	0.11	103	4.0	−3	2.7	0.17	266	6.1
2HVM	273	2EBN	285	−22.7	−2.1	0.11	111	10.1	−28	−0.3	0.04	59	8.3
2OHX-A:175-324	150	1QOR-A:136-265	130	−40.2	−6.3	0.19	99	4.9	−1	3.5	0.22	127	6.0
3GRS:364-478	115	1NPX:322-447	126	−26.4	−5.0	0.12	73	3.0	−6	2.5	0.13	115	17.1
8FAB-A:3-105	103	1HNF:4-104	101	−39.3	−6.1	0.11	61	2.8	−2	2.5	0.12	98	3.9
2RSP-A	115	1DIF-A	99	−19.1	−4.7	0.18	51	5.4	0	2.1	0.22	90	10.5
1OPR	213	1ECF-A:250-469	220	−14.5	−2.9	0.12	86	7.2	−2	1.9	0.14	209	18.8
1ORO-A	213	1ECF-A:250-469	220	−8.9	−2.4	0.12	85	8.9	−4	1.7	0.13	150	18.4
1ECE-A	358	1EDG	380	−14.3	−1.3	0.09	68	4.2	−8	1.0	0.06	119	17.5
1NDH:3-125	123	1FNB:19-154	136	3.3	−5.3	0.15	64	4.5	−16	1.9	0.22	118	5.9
2AK3-A	226	1GKY	186	−18.6	−3.1	0.11	80	13.3	−16	0.8	0.16	164	21.7
1SVB:304-395	92	1GOF:538-639	102	−5.1	−3.4	0.16	68	9.8	−11	1.6	0.19	84	9.8
1ECP-A	237	1PBN	289	−14.7	−4.5	0.10	107	2.6	−25	−0.1	0.14	231	15.4
1PII:255-452	198	1PII:1-254	254	−37.4	−2.5	0.08	83	3.8	−31	−0.6	0.09	139	8.4
1FDS	282	1XEL	338	−7.5	−2.4	0.10	84	4.7	−1	2.4	0.05	54	13.7
2LBP	346	2DRI	271	−2.8	−7.2	0.10	133	6.7	−14	−0.2	0.15	211	23.1
2LIV	344	2DRI	271	9.1	−5.7	0.10	132	7.1	−20	−1.0	0.04	63	17.2
3INK-C	121	2GMF-A	121	−45.7	−2.6	0.08	51	4.8	−28	−0.4	0.11	67	12.7
2EBN	285	2HVM	273	−17.6	−4.1	0.13	79	8.7	−28	−0.1	0.04	59	8.3
1QOR-A:136-265	130	2OHX-A:175-324	150	−19.1	−6.7	0.16	87	4.3	−1	3.7	0.22	127	6.0
1GAL:3-324	322	3COX:5-318	314	30.7	−3.5	0.14	129	9.8	−12	0.9	0.05	107	18.5

† Only protein pairs with 50 or more aligned residue pairs are listed in this table.

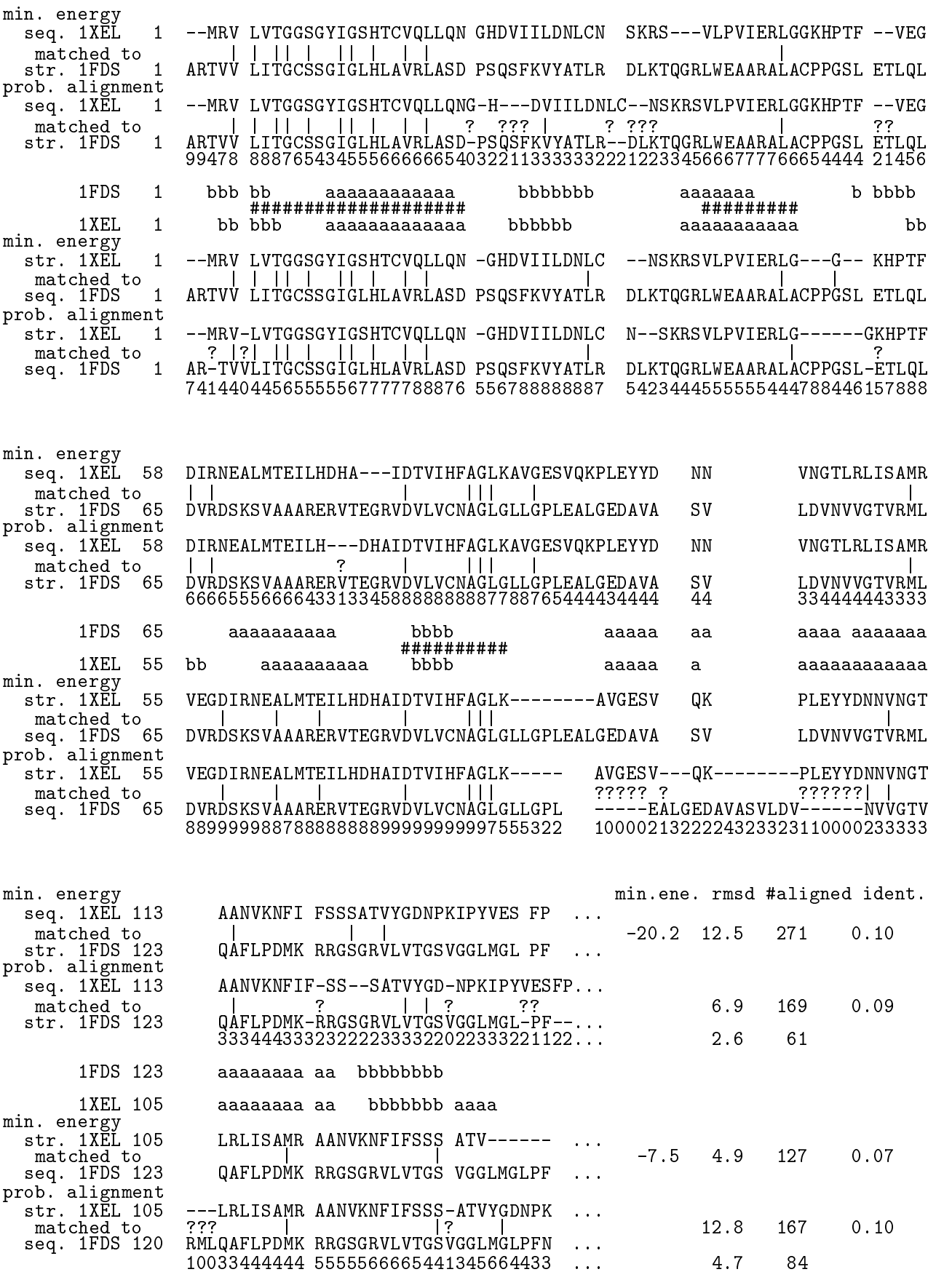

Figure 2: An example of sequence-structure alignments; only N-terminal fragments are shown.

≥ 50 residue pairs aligned with probabilities ≥ 0.5 are listed in Table 4. The relatively small values of r.m.s.d. for these protein pairs in sequence-structure alignments indicate that reasonable alignments for most of the protein pairs are obtained.

3.3 An Example of Sequence-Structure Alignments

Figure 2 shows sequence-structure alignments between UDP-galactose-4-epimerase from *E. coli* (1XEL) and human estrogenic 17β-hydroxysteroid dehydrogenase (1FDS) in the family of tyrosine-dependent oxidoreductases; only aligned N-terminal fragments are shown in this figure. Both types of alignment, that is, the sequence of 1XEL versus the structure of 1FDS, and inversely the structure of 1XEL versus the sequence of 1FDS, are shown. Also, for each type of sequence-structure alignment, two kinds of alignment are shown in this figure; the minimum energy score alignment and the probability alignment that is made by successively aligning site pairs in order of their alignment probabilities. The numbers below the sequences in these alignments represent probabilities with which those residue pairs are aligned; "5" for example means that the probability is greater than or equal to 0.5 and less than 0.6. The question marks between sequences indicate that those site pairs do not correspond to site pairs with maximum alignment probabilities over all other sites and thus those alignments of residues are very uncertain. This protein pair is one of the protein pairs whose compatibility was not detected by the conventional sequence alignment, but by the present sequence-structure alignment; see Table 4.

Probability alignments consisting of most reliable site pairs only can yield small root mean square deviations, and including less reliable pairs increases the deviations. The minimum energy alignments and probability alignments tend to align the same residue pairs but not always, when alignment probabilities are greater than 0.5. Also, it should be noticed that both types of sequence-structure and inverse structure-sequence alignments tend to be identical especially at sites aligned with probabilities greater than 0.5; sites commonly aligned in all alignments are marked by "#" between the alignments.

4 Discussion

Here, pairwise interaction energies have been evaluated in a mean field approximation on the basis of probabilities of site pairs being aligned. Alignments have also been made by successively aligning site pairs in order of their alignment probabilities. This probability alignment method provides information about how reliable each aligned site pair is. This feature is particularly desirable for aligning distantly related sequences and structures. The present energy function and alignment method can complement the conventional sequence alignment method in detecting homologous proteins.

References

[1] Bowie, J.U., Lüthy, R., and Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure, *Science* 253:164–170, 1991.

[2] Bryant, S.H., and Lawrence, C.E., An empirical energy function for threading protein sequence through the folding motif, *Proteins* 16:92–112, 1993.

[3] Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C., A model of evolutionary change in proteins, *Atlas of protein sequence and structure* 1978, Vol. 5, Suppl. 3, ed. Dayhoff, M.O., National Biomedical Research Foundation, Washington D.C., 1978.

[4] Finkelstein, A.V., and Reva, B.A., A search for the most stable folds of protein chains, *Nature* 351:497–499, 1991.

[5] Go, M., and Miyazawa, S., Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution, *Int. J. Peptide Protein Res.* 15:211–224, 1980.

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

[7] Hendlich, M. *et al.*, Identification of native protein folds amongst a large number of incorrect models; the calculation of low energy conformations from potentials of mean force, *J. Mol. Biol.* 216:167–180, 1990.

[8] Henikoff, S., and Henikoff, J.G., Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA* 89:10915–10919, 1992.

[9] Huang, E.S., Subbiah, S., and Levitt, M., Recognizing native folds by the arrangement of hydrophobic and polar residues, *J. Mol. Biol.* 252:709–720, 1995.

[10] Jones, D.T., Taylor, W.R., and Thornton, J.M., A new approach to protein fold recognition, *Nature* 358:86–89, 1992.

[11] Lathrop, R.H., and Smith, T.F., Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions, *J. Mol. Biol.* 255:641–665, 1996.

[12] Kocher, J.-P.A. *et al.*, Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches, *J. Mol. Biol.* 235:1598–1613, 1994.

[13] Maiorov, V.N., and Crippen, G.M., Contact potential that recognizes the correct folding of globular proteins, *J. Mol. Biol.* 227:876–888, 1992.

[14] Matsuo, Y., Nakamura, H., and Nishikawa, K., Detection of protein 3D-1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions, *J. Biochem.* 118:137–148, 1995.

[15] Mirny, L.A., and Shakhnovich, E.I., How to derive a protein folding potential? A new approach to an old problem, *J. Mol. Biol.* 264:1164–1179, 1996.

[16] Miyazawa, S., and Jernigan, R.L., Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules* 18:534–552, 1985.

[17] Miyazawa, S., A reliable sequence alignment method based on probabilities of residue correspondences, *Protein Eng.* 8:999–1009, 1995.

[18] Miyazawa, S., and Jernigan, R.L., Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading, *J. Mol. Biol.* 256:623–644, 1996.

[19] Miyazawa, S., and Jernigan, R.L., Self-consistent Estimation of Inter-residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues, *Proteins* 34:49–68, 1999.

[20] Miyazawa, S., and Jernigan, R.L., Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition, *Proteins* 36:347–356, 1999.

[21] Miyazawa, S., and Jernigan, R.L., An empirical energy potential with a reference state for protein fold and sequence recognition, *Proteins* 36:357–369, 1999.

[22] Miyazawa, S., and Jernigan, R.L., Identifying sequence-structure pairs undetected by sequence alignments, *Protein Eng.* 13:459–475, 2000.

[23] Munson, P.J., and Singh, R.K. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment, *Protein Sci.* 6:1467–1481, 1997.

[24] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., Scop: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247:536–540, 1995.

[25] Nishikawa, K., and Matsuo, Y., Development of pseudoenergy potentials for assessing protein 3-D - 1-D compatibility and detecting weak homologies, *Protein Eng.* 6:811–820, 1993.

[26] Park, B.H., Huang, E.S., and Levitt, M., Factors affecting the ability of energy functions to discriminate correct from incorrect folds, *J. Mol. Biol.* 266:831–846, 1997.

[27] Park, B., and Levitt, M., Energy functions that discriminate X-ray and near-native folds from well-constructed decoys, *J. Mol. Biol.* 258:367–392, 1996.

[28] Samudrala, R., and Moult, J., An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, *J. Mol. Biol.* 275:895–916, 1998.

[29] Sippl, M.J., Calculation of conformational ensembles from potentials of mean force, *J. Mol. Biol.* 213:859–883, 1990.

[30] Sippl, M.J., and Weitckus, S., Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations, *Proteins* 13:258–271, 1992.