An Energy Potential and Alignment Method for Identifying Protein Sequence - Structure Compatibilities

Sanzo Miyazawa[1]
miyazawa@smlab.sci.gunma-u.ac.jp

[1] Faculty of Technology, Gunma University, Kiryu, Gunma 376, Japan

presented at
International Conference on Structural Genomics 2000
held in November 2nd to 5th of 2000
at Pacifico Yokohama in Yokohama, Japan.

Abstract

We develop a method for sequence - structure alignments and examine how effectively simple potential functions previously developed can identify compatibilities between sequences and structures of proteins for database searches. The stabilities of structures are assumed here as a primary requirement for compatibilities between sequences and structures. The stabilities of conformations depend on not only their conformational energies but the whole ensemble of conformations. The amino acid composition dependencies of the latter are taken into account. The potential function consists of pairwise contact energies, repulsive packing potentials of residues for overly dense arrangement and short-range potentials for secondary structures, all of which were estimated from statistical preferences observed in known protein structures (Proteins, 34:49-68, 1999). In the preceding paper (Proteins, 36:357-369, 1999), it was shown that this simple potential function can distinguish native structures from alternate folds and also recognize native sequences from non-native sequences by threading sequences into other structures in all possible ways without gaps. Here, it is more thoroughly examined by allowing deletions and insertions in sequence - structure alignments (Protein Eng. 13:459-475, 2000).

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. To obtain the self-consistent values of alignment probabilities of site pairs, an iterative method is employed. 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 (Protein Eng. 8:999-1009, 1995) that are made by successively aligning site pairs in order by pairwise alignment probabilities and provide information of how reliable each aligned site pair is.

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, and yield mostly

similar alignments for these two types of sequence and structure pairs. Probability alignments consisting of most reliable site pairs only can yield extremely small root mean square deviations, and including less reliable pairs increases the deviations. Also it is observed that secondary structure potentials are usefully complementary to yield improved alignments with this method. Remarkably, by this method some individual sequence-structure pairs are detected having only 5-20 % sequence identity.

1 Methods

1.1 A Statistical Ensemble of Sequence-Structure Alignments

An example of a specific **sequence–structure alignment A:**

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

where

s_p	is the conformational state of the pth residue,
i_q	means the qth residue of type i_q.

A conditional probability $\mathcal{P}(\{s_p\}|\{i_q\}, A)$ for alignment A to take a specific conformation $\{s_p\}$:

$$-\log\{\mathcal{P}(\{s_p\}|\{i_q\}, A)\} \approx \beta\Delta E^{\text{conf}}(\{s_p\}|\{i_q\}, A) + n_r^{\text{aligned}}\sigma \quad (2)$$

where

β	$\equiv 1/(kT)$,
n_r^{aligned}	is the number of aligned site pairs,
σ	is a conformational entropy per residue for native-like structures,
$\Delta E_p^{\text{conf}}(\{s_p\}|\{i_q\}, A)$	
	s an alignment energy of $\{s_p\}$, which is a sum of pairwise contact energies, repulsive packing potentials, and secondary structure potentials, and is modified approximately to represent the stabilities of structures, [3] $< \Delta E_p^{\text{conf}}(\{s_p\}|\{i_q\}, A) >_{\text{native structures}} = 0$

Then, **the conditional probability** $\mathcal{P}(A|\{s_p\},\{i_q\})$ of an alignment A for a given structure $\{s_p\}$:

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \mathcal{P}(\{s_p\}|\{i_q\},A)\mathcal{P}(A) \;/\; [\, \sum_A \mathcal{P}(\{s_p\}|\{i_q\},A)\mathcal{P}(A)\,] \tag{3}$$

$$-\log\{\mathcal{P}(A)\} \equiv n_r^{\text{aligned}}(\beta\mathcal{E}_0 - \sigma) + \beta\,[\sum_{\text{all gaps in } A} \mathcal{W}\,] + \text{constant} \tag{4}$$

where

$\mathcal{P}(A)$	is *a priori* probability for an alignment A,
$\mathcal{W}$	is a positive quantity to represent a gap penalty,
$\mathcal{E}_0$	is a negative constant as a scaling parameter.

Thus,

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \frac{1}{\mathcal{Z}} \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{5}$$

$$\mathcal{Z} = \sum_A \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{6}$$

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \Delta E^{\text{conf}}(\{s_p\}|\{i_q\},A) + n_r^{\text{aligned}}\mathcal{E}_0 + \sum_{\text{all gaps in } A} \mathcal{W} \tag{7}$$

where

$\mathcal{Z}$	is a partition function for alignments,
$\mathcal{E}(\{s_p\}|\{i_q\},A)$	is the energy score of an alignment A.

1.2 Pairwise Interactions Approximated on the Basis of Site-Alignment Probabilities

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.

$$\mathcal{E}(\{s_p\}|\{i_q\}, A) \equiv \sum_{(p,q)\in A} \mathcal{E}(\{s_p\}|i_q, A) + \sum_{\text{all gaps in } A} \mathcal{W} \tag{8}$$

$$\mathcal{E}(\{s_p\}|i_q, A) \equiv \mathcal{E}_0 + \mathcal{E}_1(s_p|i_q) + \frac{1}{2}\sum_{(p',q')\in A} \mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'}) + \cdots \tag{9}$$

Here, the pairwise interaction energies for alignment A that significantly contributes to the partition function in Eq. 6 are approximated as:

$$\sum_{(p',q')\in A} \mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'}) \approx \sum_{p'}\sum_{q'} \mathcal{E}_2(s_p, s_{p'}|i_q, i_{q'})\mathcal{P}(p', q') \tag{10}$$

The alignment probabilities $\mathcal{P}(p, q)$ for structure-sequence site pairs (p, q):

$$\begin{aligned} \mathcal{P}(p, q) &= \frac{1}{\mathcal{Z}} \sum_{A \text{ with } (p,q)} \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\}, A)] && (11) \\ &\simeq \frac{1}{\mathcal{Z}} \mathcal{Z}_{p-1,q-1} \exp[-\beta\mathcal{E}(\{s_p\}|i_q, \mathcal{P}(p', q'))]\, \mathcal{Z}'_{p+1,q+1} && (12) \\ \mathcal{P}(p, -) &= 1 - \sum_q \mathcal{P}(p, q) \quad , \quad \mathcal{P}(-, q) = 1 - \sum_p \mathcal{P}(p, q) && (13) \end{aligned}$$

A self-consistent solution for alignment probabilities $\mathcal{P}(p, q)$ is calculated by an iteration method.

1.3 Alignment Based on Site-Alignment Probabilities

Two types of alignment methods are used;

(i) **Minimum energy alignment**, $A^{\min}$; $\mathcal{E}(\{s_p\}|\{i_q\}, A^{\min}) \equiv \min_A \mathcal{E}(\{s_p\}|\{i_q\}, A)$.

(ii) **Probability alignment**, [1] consisting of the most probable site pairs by successively aligning a site pair in order of pairwise alignment probabilities $\mathcal{P}(p, q)$.

1.4 Structure-Dependent Gap Penalties

The dependence of residue mutability on residue position 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 present values of gap parameters are adjusted to yield similar fractions of aligned residues in sequence-structure alignments for homologous protein pairs to those in sequence alignments.

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.

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	$5.5 + \sum_{p=q}^{q1}(1.05 + 0.43 n_p^c)$ in the middle
	$3.25 + \sum_{p=q}^{q1}(0.53 + 0.22 n_p^c)$ 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 $6.5\AA$ from the side chain center of the pth residue, excluding neighboring residues along a sequence.

1.5 Datasets of Protein Structures

Two datasets of protein pairs were prepared from SCOP 1.35; structures with high resolution from α, β, α/β, $\alpha + \beta$, and multi-domain proteins are used.

(i) **A dataset of 548 homologous protein pairs:** by pairing the protein representatives of families with those of different species within the families.

(ii) **A dataset of 505 or 5041 dissimilar protein pairs:** by arbitrarily choosing protein pairs from all possible pairs of superfamily representatives.

2 Results

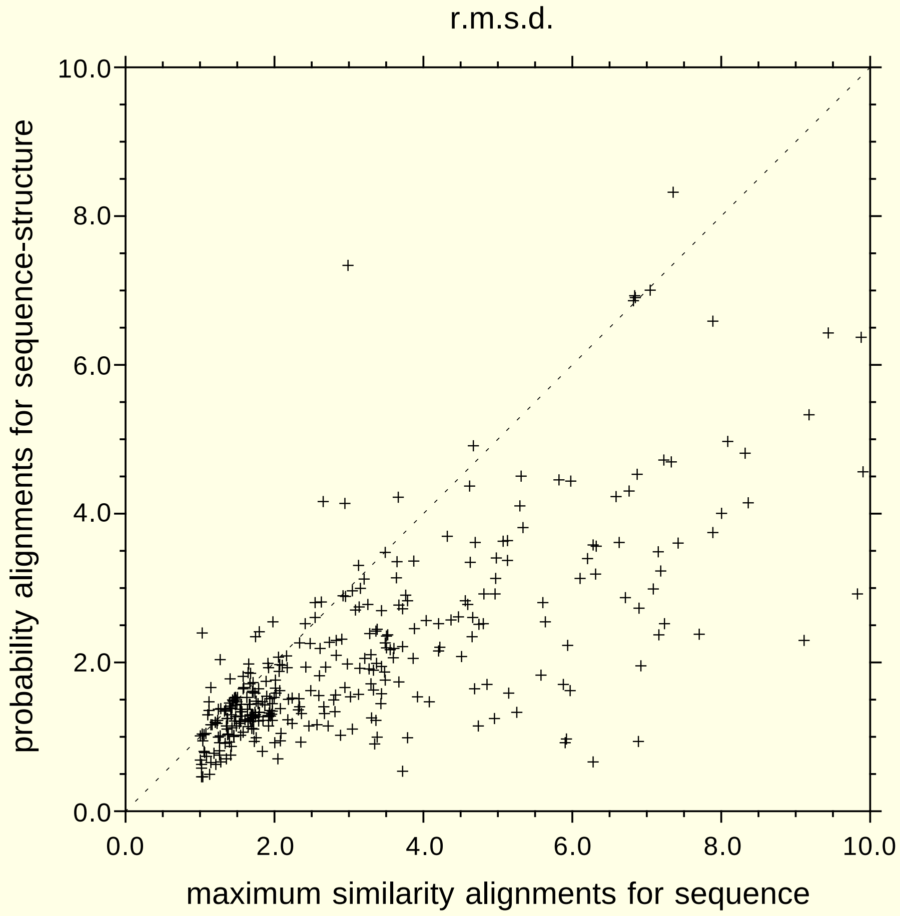
2.1 Characteristics of Sequence-Structure Alignments

2.1.1 Comparison of probability sequence-structure alignments with maximum similarity sequence alignments

Significant improvements in the values of r.m.s.d. are shown, although these improvements are made partially by choosing only residue pairs most reliably aligned.

2.1.2 Comparison between sequence-structure and inverse structure-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.



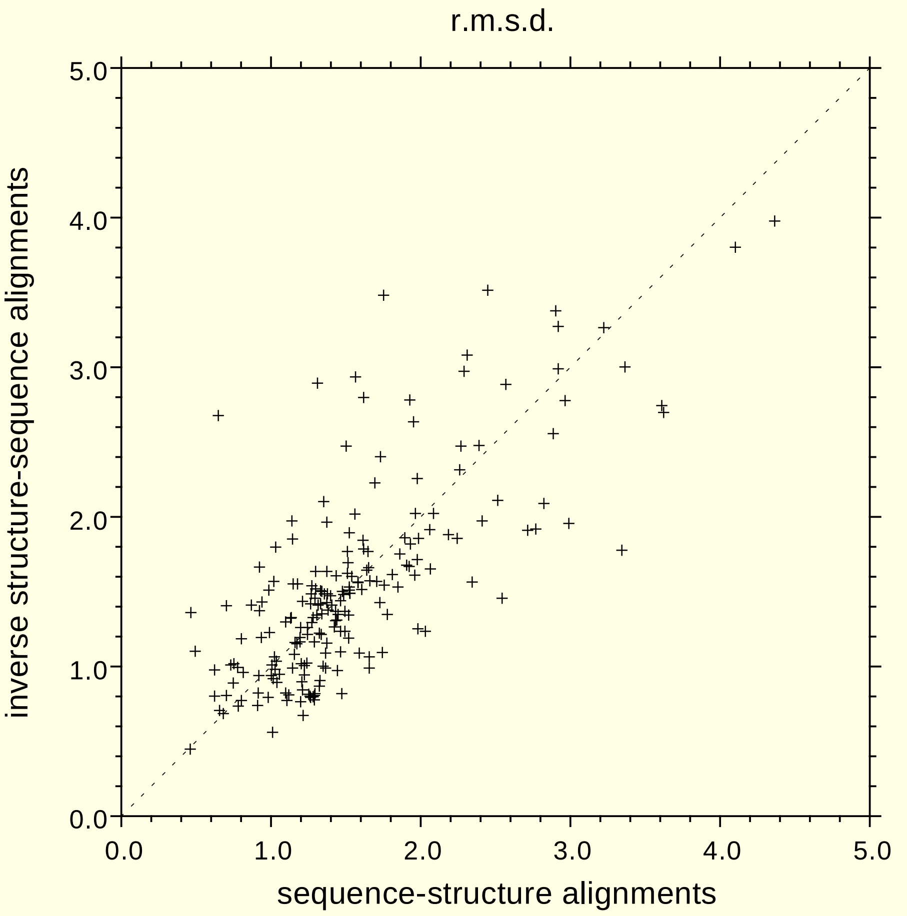


Fig. 1A 357 homologous protein pairs, which have negative minimum energy scores and positive maximum similarity scores and also whose alignments have aligned residue pairs ≥ 50, are plotted.

Fig. 1B 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.

2.1.3 Relationships between minimum energy scores and characteristics of alignments

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.

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.

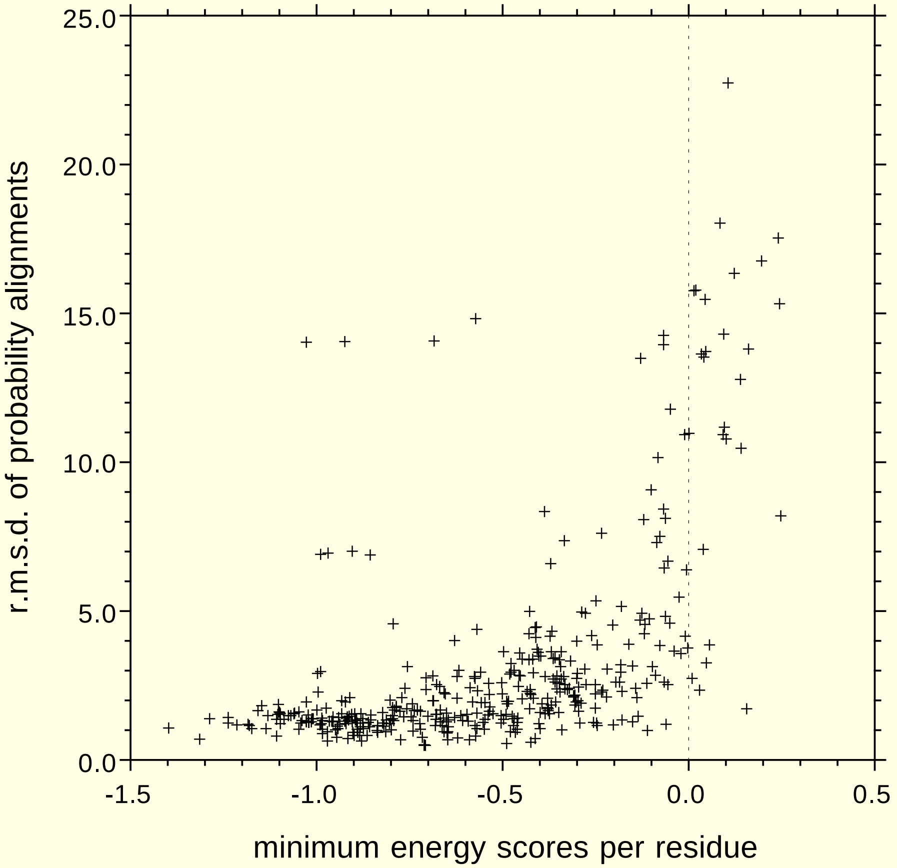

25.0
20.0
15.0
10.0
5.0
0.0
-1.5
-1.0
-0.5
0.0
0.5
r.m.s.d. of probability alignments
minimum energy scores per residue

Fig. 2A

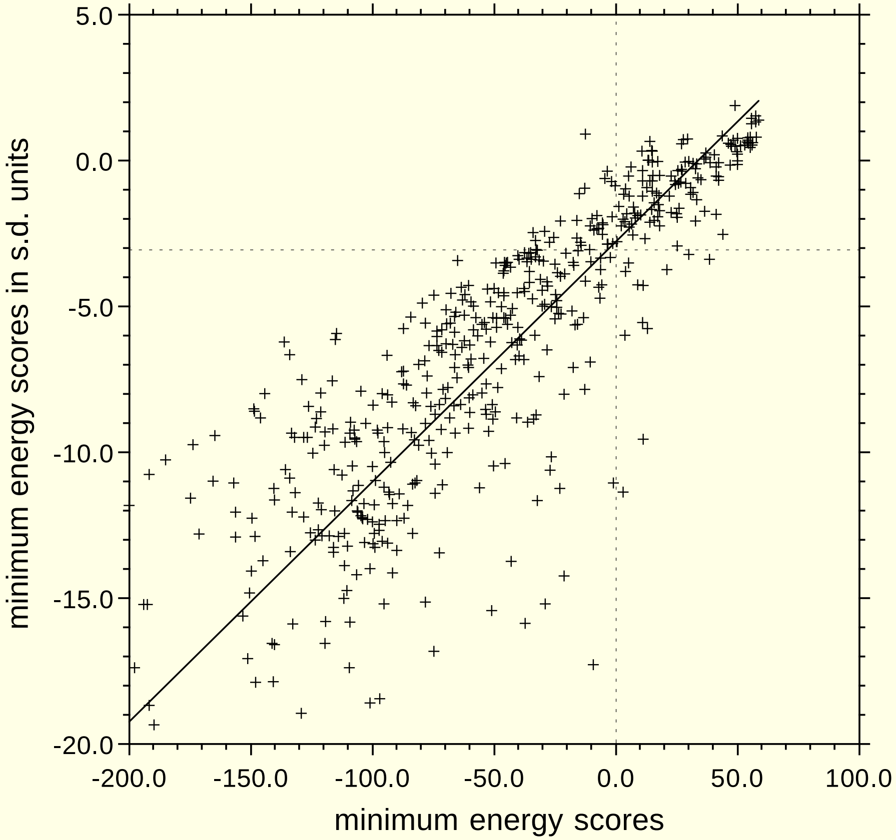
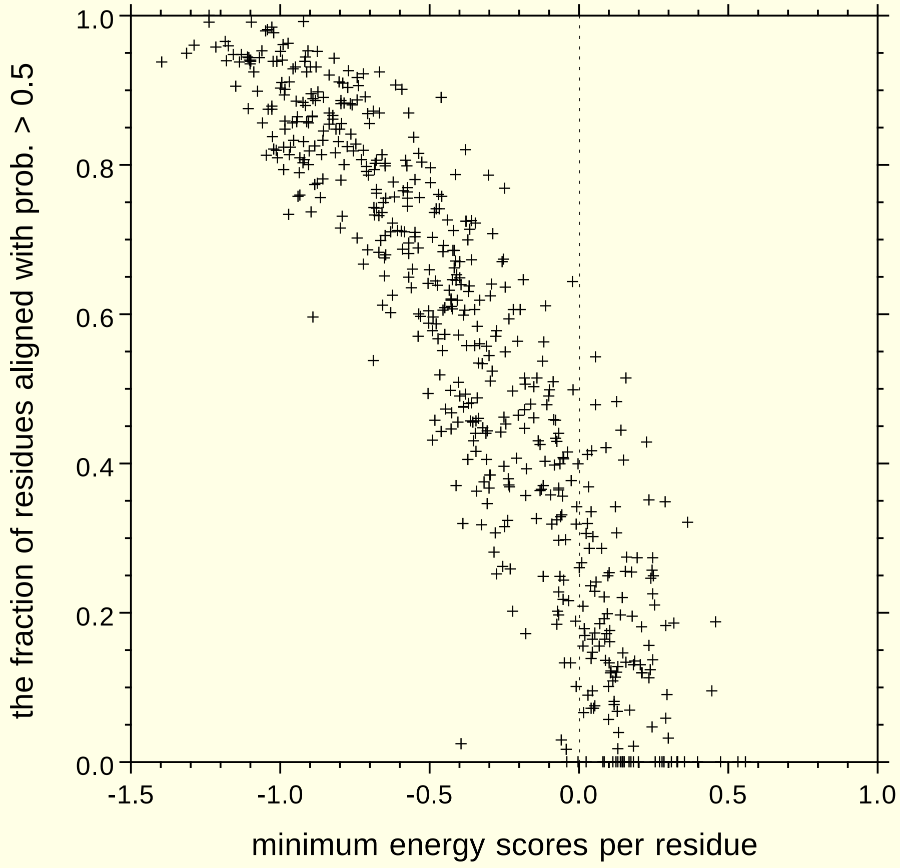
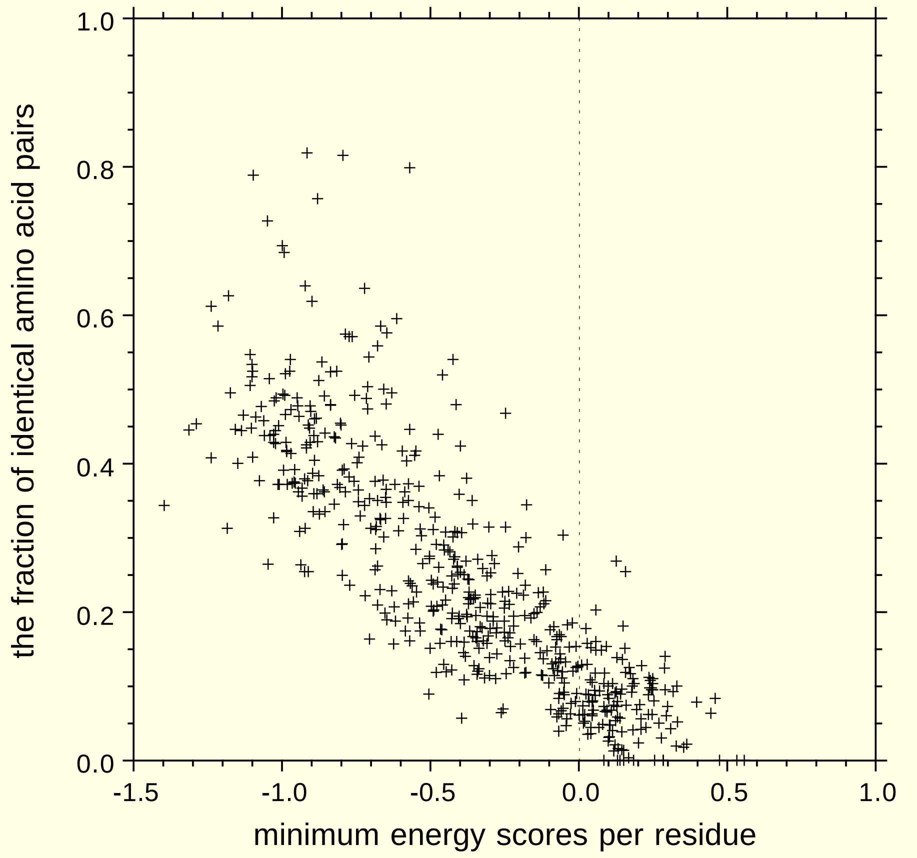

5.0
0.0
-5.0
-10.0
-15.0
-20.0
-200.0
-150.0
-100.0
-50.0
0.0
50.0
100.0
minimum energy scores in s.d. units
minimum energy scores

Fig. 2B



the fraction of residues aligned with prob. > 0.5
minimum energy scores per residue

Fig. 3A



the fraction of identical amino acid pairs
minimum energy scores per residue

Fig. 3B

2.2 Detection of Homologous Proteins from Dissimilar Proteins

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† with score	False negatives in homologous protein pairs† with z-score	False positives in dissimilar protein pairs with score		False positives in dissimilar protein pairs with z-score	Alignment method
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. similarity score	seq.-str. energy score $<$	seq.-str. energy score $\geq$	inverse energy score $<$	inverse energy score ≥ 0	seq.-seq. similarity z-score	seq.-str. energy z-score $<$	seq.-str. energy z-score $\geq$	inverse energy z-score $<$	inverse energy z-score ≥ -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 alignment includes less than 30% identity.

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 (Å)	maximum similarity score	z-score	identities	# aligned residue pairs	rmsd (Å)
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.

2.3 An Example of Sequence-Structure Alignments

```
min. energy
  seq. 1XEL   1  --MRV LVTGGSGYIGSHTCVQLLQN GHDVIILDNLCN  SKRS---VLPVIERLGGKHPTF --VEG
   matched to        | | || |  || |  | |                                |
  str. 1FDS   1  ARTVV LITGCSSGIGLHLAVRLASD PSQSFKVYATLR  DLKTQGRLWEAARALACPPGSL ETLQL
prob. alignment
  seq. 1XEL   1  --MRV LVTGGSGYIGSHTCVQLLQNG-H---DVIILDNLC--NSKRSVLPVIERLGGKHPTF --VEG
   matched to        | | || |  || |  | |   ?  ??? |     ? ???           |        ??
  str. 1FDS   1  ARTVV LITGCSSGIGLHLAVRLASD-PSQSFKVYATLR--DLKTQGRLWEAARALACPPGSL ETLQL
                 99478 88876543455566666540322113333332221223345666777766654444 21456

       1FDS   1    bbb bb     aaaaaaaaaaaa      bbbbbbb        aaaaaaa         b bbbb
                       ####################                      #########
       1XEL   1     bb bbb    aaaaaaaaaaaaa    bbbbbb          aaaaaaaaaaa          bb
min. energy
  str. 1XEL   1  --MRV LVTGGSGYIGSHTCVQLLQN -GHDVIILDNLC  --NSKRSVLPVIERLG---G-- KHPTF
   matched to        | | || |  || |  | |              |                 |    |
  seq. 1FDS   1  ARTVV LITGCSSGIGLHLAVRLASD PSQSFKVYATLR  DLKTQGRLWEAARALACPPGSL ETLQL
prob. alignment
  str. 1XEL   1  --MRV-LVTGGSGYIGSHTCVQLLQN -GHDVIILDNLC  N--SKRSVLPVIERLG------GKHPTF
   matched to      ? |?| || |  || |  | |              |                  |       ?
  seq. 1FDS   1  AR-TVVLITGCSSGIGLHLAVRLASD PSQSFKVYATLR  DLKTQGRLWEAARALACPPGSL-ETLQL
                 74144044565555567777788876 556788888887  5423444555555444788446157888


min. energy
  seq. 1XEL  58  DIRNEALMTEILHDHA---IDTVIHFAGLKAVGESVQKPLEYYD   NN        VNGTLRLISAMR
   matched to    | |                 |     |||   |                                  |
  str. 1FDS  65  DVRDSKSVAAARERVTEGRVDVLVCNAGLGLLGPLEALGEDAVA   SV        LDVNVVGTVRML
prob. alignment
  seq. 1XEL  58  DIRNEALMTEILH---DHAIDTVIHFAGLKAVGESVQKPLEYYD   NN        VNGTLRLISAMR
   matched to    | |           ?     |     |||   |                                  |
  str. 1FDS  65  DVRDSKSVAAARERVTEGRVDVLVCNAGLGLLGPLEALGEDAVA   SV        LDVNVVGTVRML
                 66665556666433133458888888887788765444434444   44        334444443333

       1FDS  65      aaaaaaaaaa       bbbb              aaaaa   aa        aaaa aaaaaaa
                                     ##########
       1XEL  55  bb     aaaaaaaaaa    bbbb              aaaaa   a         aaaaaaaaaaaa
min. energy
  str. 1XEL  55  VEGDIRNEALMTEILHDHAIDTVIHFAGLK--------AVGESV   QK        PLEYYDNNVNGT
   matched to       |    |   |       |     |||                                    |
  seq. 1FDS  65  DVRDSKSVAAARERVTEGRVDVLVCNAGLGLLGPLEALGEDAVA   SV        LDVNVVGTVRML
prob. alignment
  str. 1XEL  55  VEGDIRNEALMTEILHDHAIDTVIHFAGLK-----   AVGESV---QK--------PLEYYDNNVNGT
   matched to       |    |   |       |     |||         ????? ?            ??????| |
  seq. 1FDS  65  DVRDSKSVAAARERVTEGRVDVLVCNAGLGLLGPL   -----EALGEDAVASVLDV------NVVGTV
                 88999999887888888889999999997555322   1000021322224323323110000233333


min. energy                                              min.ene. rmsd #aligned ident.
  seq. 1XEL 113     AANVKNFI FSSSATVYGDNPKIPYVES FP  ...
   matched to        |          |  |                      -20.2  12.5   271    0.10
  str. 1FDS 123     QAFLPDMK RRGSGRVLVTGSVGGLMGL PF  ...
prob. alignment
  seq. 1XEL 113     AANVKNFIF-SS--SATVYGD-NPKIPYVESFP...
```

```
   matched to          |       ?       | | ?      ??              6.9   169    0.09
  str. 1FDS 123     QAFLPDMK-RRGSGRVLVTGSVGGLMGL-PF--...
                    3334443332322223333220223332211 22...           2.6    61

       1FDS 123     aaaaaaaa aa  bbbbbbbb

       1XEL 105     aaaaaaaa aa   bbbbbbb aaaa
min. energy
  str. 1XEL 105     LRLISAMR AANVKNFIFSSS ATV------  ...
   matched to            |             |                  -7.5   4.9   127    0.07
  seq. 1FDS 123     QAFLPDMK RRGSGRVLVTGS VGGLMGLPF  ...
prob. alignment
  str. 1XEL 105  ---LRLISAMR AANVKNFIFSSS-ATVYGDNPK  ...
   matched to    ???     |             |?     |                12.8   167    0.10
  seq. 1FDS 120  RMLQAFLPDMK RRGSGRVLVTGSVGGLMGLPFN  ...
                 10033444444 5555566665441345664433  ...           4.7    84
```

References

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

[2] 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.

[3] 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.

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