Protein Sequence-Structure Alignment Based on Site-Alignment Probabilities

Sanzo Miyazawa[1]
miyazawa@smlab.sci.gunma-u.ac.jp
[1] Faculty of Technology, Gunma University, Kiryu, Gunma 376, Japan

to be presented at
GIW 2000 The Eleventh Workshop on Genome Informatics
held in Decemberer 18 to 19 of 2000
at Yebisu Garden Place in Tokyo, Japan.

1 Introduction

Purpose:

- A method of **pairwise sequence-structure alignment** is **developed** and **examined** on how effectively compatibilities between protein sequences and structures can be identified,
  - towards a structure/function prediction from sequence and
  - in order to better understand what kind of interactions are essential for protein structures to fold.

  In the sequence-structure alignments, **only structural information** from one of a protein pair and sequence information from the other are used.

Methods:

- What kind of **scoring function** is used?

  A scoring function consists of **structure-dependent gap penalties** and **statistical energy potentials** which were estimated from statistical preferences observed in known protein structures and modified to measure approximately the stabilities of structures.

- How are **two-body interactions** handled to obtain an optimum alignment?

  Pairwise interactions are evaluated **in a mean field approximation** on the basis of site-alignment probabilities, whose self-consistent values are calculated by an iteration method.

- What kind of **alignment method** is used?

  In addition to **minimum energy alignments**, we use probability alignments which are made by successively aligning site pairs in order of their alignment probabilities.

Analyses:

- To examine the qualities of sequence-structure alignments, their overall characteristics such as **r.m.s.d.** are compared with those of conventional sequence alignments.

- Capabilities of both methods to **identify homologous proteins** are compared with each other.

2 Methods

2.1 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

"–"	means a deletion,
s_p	is the conformational state of the pth residue in the structure,
i_q	means the qth residue of amino acid type i_q in the sequence.

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

$$-\log\{\mathcal{P}(\{s_p\}|\{i_q\}, A)\} = \beta E^{\rm conf}(\{s_p\}|\{i_q\}, A) + \log[\sum_{\{s_p\}} \exp(-\beta E^{\rm conf}(\{s_p\}|\{i_q\}, A))] \quad (2)$$

$$(3)$$

$$\approx \beta \Delta E^{\rm conf}(\{s_p\}|\{i_q\}, A) + n_r^{\rm aligned}\sigma \quad (4)$$

where

β	$\equiv 1/(kT)$,
$n_r^{\rm aligned}$	is the number of aligned site pairs,
σ	is a conformational entropy per residue in k units for native-like structures,

$$\Delta E_p^{\mathbf{conf}}(\{s_p\}|\{i_q\}, A) \equiv E_p^{\mathbf{conf}}(\{s_p\}|\{i_q\}, A) - < E_p^{\mathbf{conf}}(\{s_p\}|\{i_q\}, A) >_{\mathbf{native\ structures}}$$

is an alignment energy of $\{s_p\}$, which is a conformational energy modified to measure approximately the stabilities of structures (Miyazawa S. and Jernigan R. L., *Proteins* 36:357-369, 1999);

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

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \frac{\mathcal{P}(\{s_p\}|\{i_q\},A)\mathcal{P}(A)}{\Sigma_A\,\mathcal{P}(\{s_p\}|\{i_q\},A)\mathcal{P}(A)} \tag{5}$$

where

$\mathcal{P}(A)$	is the *a priori* probability for an alignment A, $-\log\{\mathcal{P}(A)\} \equiv n_r^{\text{aligned}}(\beta\mathcal{E}_0 - \sigma) + \beta\,[\Sigma_{\text{all gaps in } A}\,\mathcal{W}\,] + \text{constant}$
$\mathcal{W}$	is a positive quantity to represent gap penalties,
$\mathcal{E}_0$	is a negative constant as a scaling parameter.

Thus,

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \frac{1}{\mathcal{Z}}\exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{6}$$

$$\mathcal{Z} = \sum_A \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{7}$$

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \Delta E^{\text{conf}}(\{s_p\}|\{i_q\},A) + n_r^{\text{aligned}}\mathcal{E}_0 + \sum_{\text{all gaps in } A}\mathcal{W} \tag{8}$$

where

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

2.2 Energy Potentials and Gap Penalties

Statistical energy potentials are used;

$\Delta E_p^{\mathrm{conf}}(\{s_p\}|\{i_q\}, A) \equiv$ pairwise contact + repulsive packing + secondary structure energies all of which were estimated from statistical preferences observed in known protein structures by Miyazawa S. and Jernigan R. L., *Proteins* 34:49-68, 1999.

Gap penalties are structure-dependent;

A deletion penalty of a residue is assumed to be proportional to the number of residue-residue contacts at each residue position in a protein structure, in order to take account of the dependence of residue mutability on residue position.

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

An energy scoring function used includes a two-body potential $\mathcal{E}_2$ between residues in addition to an intrinsic energy $\mathcal{E}_0$ and a one-body potential $\mathcal{E}_1$.

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \sum_{(p,q)\in A} \mathcal{E}(\{s_p\}|i_q,A) + \sum_{\text{all gaps in } A} \mathcal{W} \tag{9}$$

$$\mathcal{E}(\{s_p\}|i_q,A) \equiv \mathcal{E}_0 + \mathcal{E}_1(s_p|i_q) + \frac{1}{2}\sum_{(p',q')\in A} \mathcal{E}_2(s_p,s_{p'}|i_q,i_{q'}) \tag{10}$$

Here, the pairwise interaction energies for alignment A that significantly contributes to the partition function in Eq. 7 are approximated as:

$$\sum_{(p',q')\in A} \mathcal{E}_2(s_p,s_{p'}|i_q,i_{q'}) \approx \sum_{p'}\sum_{q'} \mathcal{E}_2(s_p,s_{p'}|i_q,i_{q'})\mathcal{P}(p',q') \tag{11}$$

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

$$\mathcal{P}(p,q) = \frac{1}{\mathcal{Z}} \sum_{A \text{ with } (p,q)} \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \tag{12}$$

$$\simeq \frac{1}{\mathcal{Z}}\mathcal{Z}_{p-1,q-1}\ \exp[-\beta\mathcal{E}(\{s_p\}|i_q,\mathcal{P}(p',q'))]\ \mathcal{Z}'_{p+1,q+1} \tag{13}$$

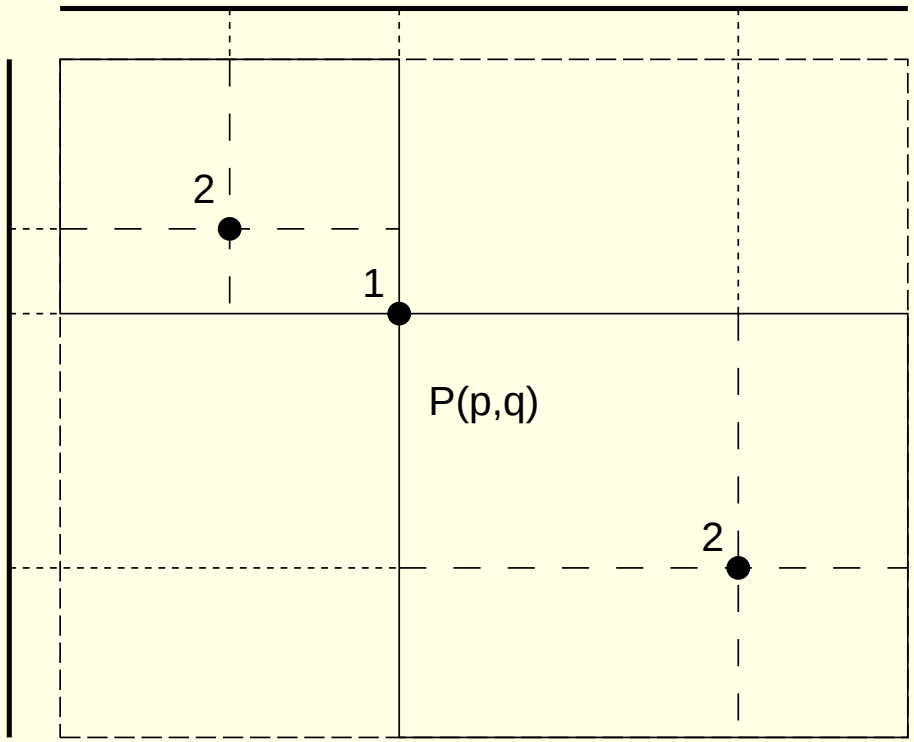
$$\mathcal{P}(p,-) = 1-\sum_q \mathcal{P}(p,q) \quad , \quad \mathcal{P}(-,q) = 1-\sum_p \mathcal{P}(p,q) \tag{14}$$

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

2.4 Alignment Methods

(i) **Minimum energy alignment**, $A^{\min}$.
$\mathcal{E}(\{s_p\}|\{i_q\}, A^{\min}) \equiv \min_A \mathcal{E}(\{s_p\}|\{i_q\}, A) \sim \min_A \mathcal{E}(\{s_p\}|\{i_q\}, \mathcal{P}(p', q'))$

(ii) **Probability alignment**, which is made by successively aligning site pairs in order of their alignment probabilities $\mathcal{P}(p, q)$ (Miyazawa S., *Protein Engineering* 8:999-1009, 1995).



Aligning a site pair in order of 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 $\mathcal{P}(p', q') \ge \mathcal{P}(-, q')$)

2.5 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.

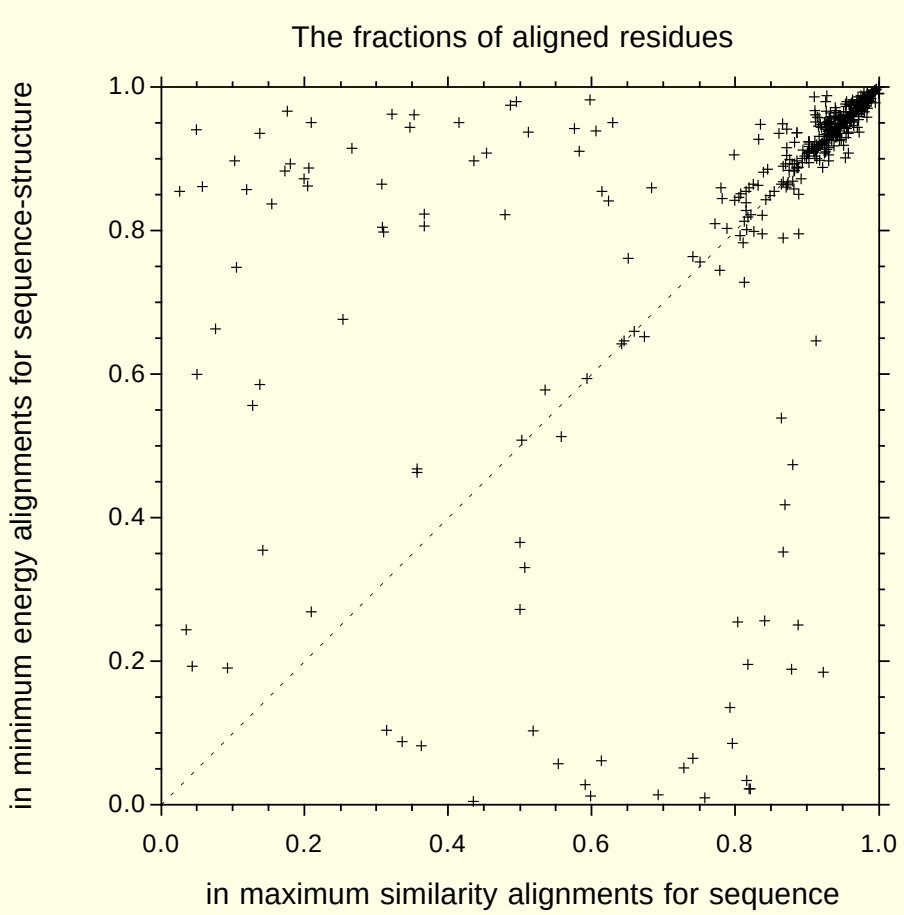
(i) **A dataset of 548 homologous protein pairs:** by pairing the protein representatives of families with those of different species within the families.

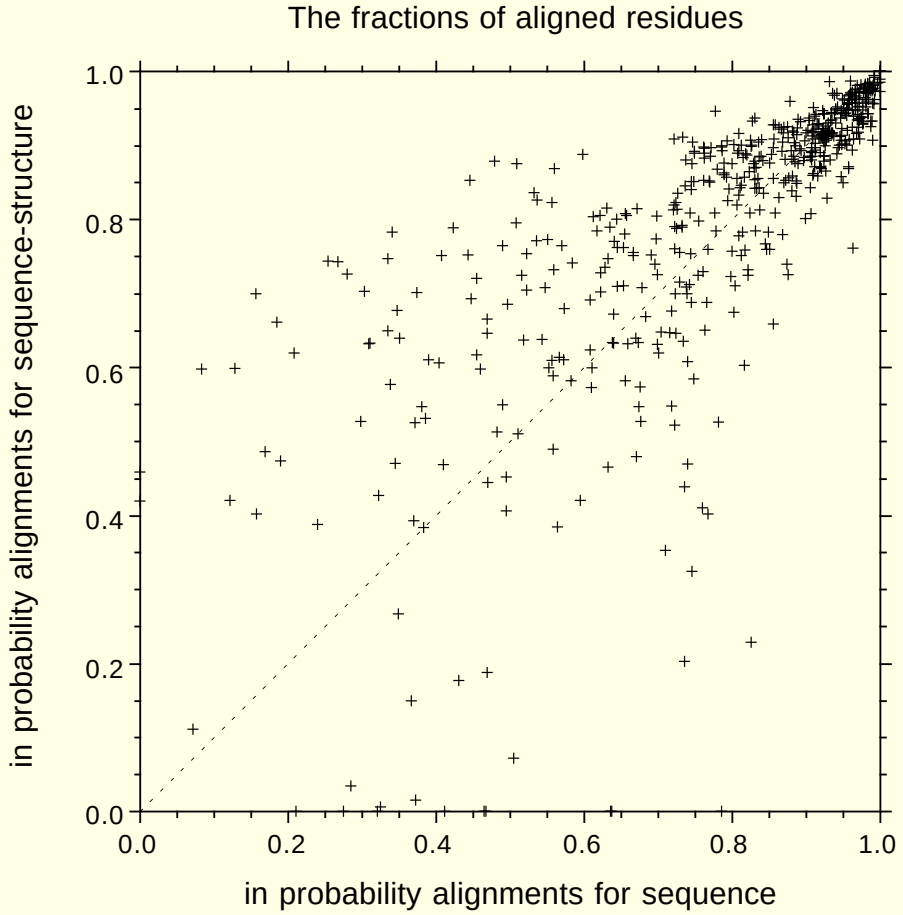
(ii) **A dataset of 505 or 5041 dissimilar protein pairs:** by arbitrarily choosing protein pairs from all possible pairs of superfamily representatives.

3 Results

3.1 How were the values of gap penalties determined?

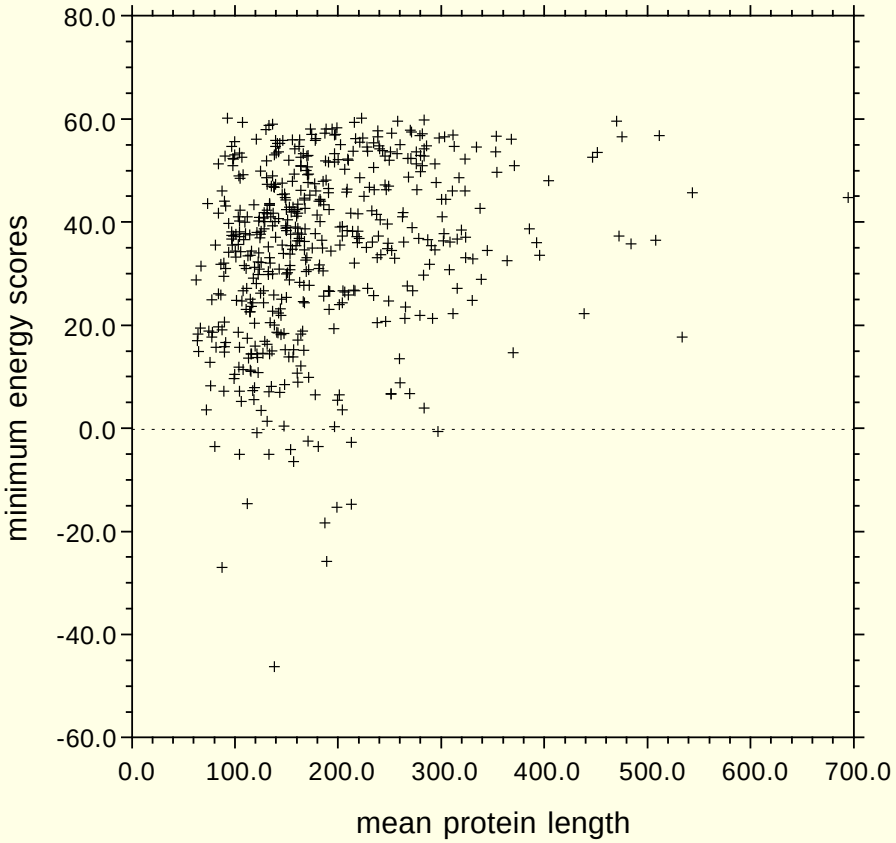
The present values of gap parameters are adjusted to yield similar fractions of aligned residues in minimum energy alignments for homologous protein pairs to those in sequence alignments, and β is also adjusted to yield similar fractions of aligned residues in probability alignments for sequence-structure compared with those in probability sequence alignments.



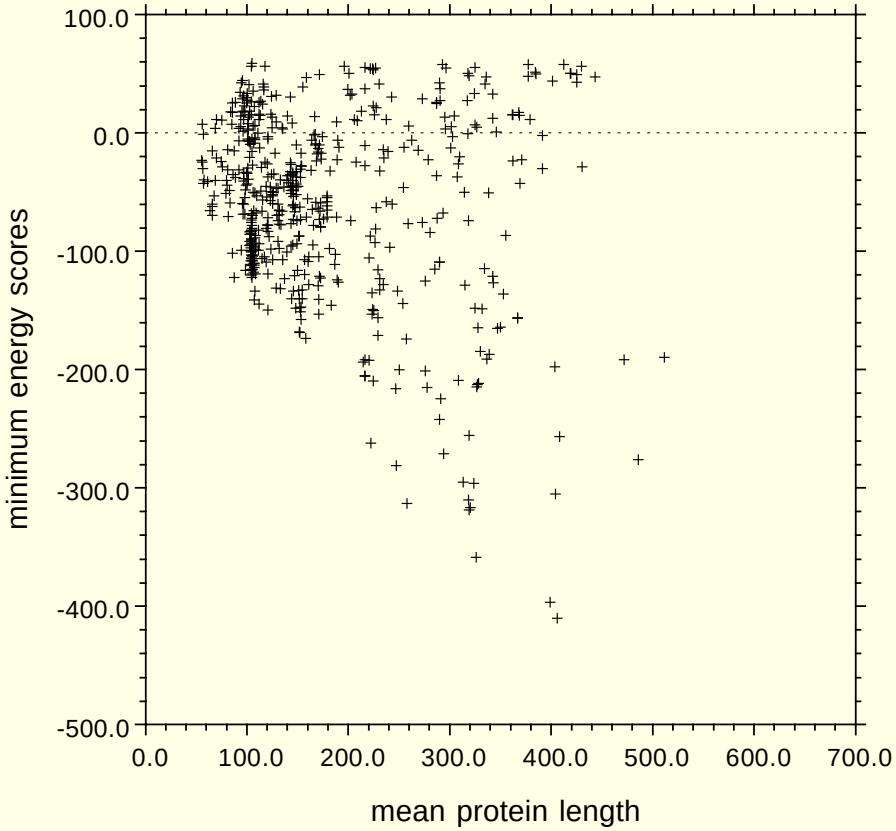


Homologous protein pairs are plotted in both figures.

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.



Dissimilar protein pairs are plotted.



Homologous protein pairs are plotted.

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.

3.2 Characteristics of Sequence-Structure Alignments

3.2.1 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.

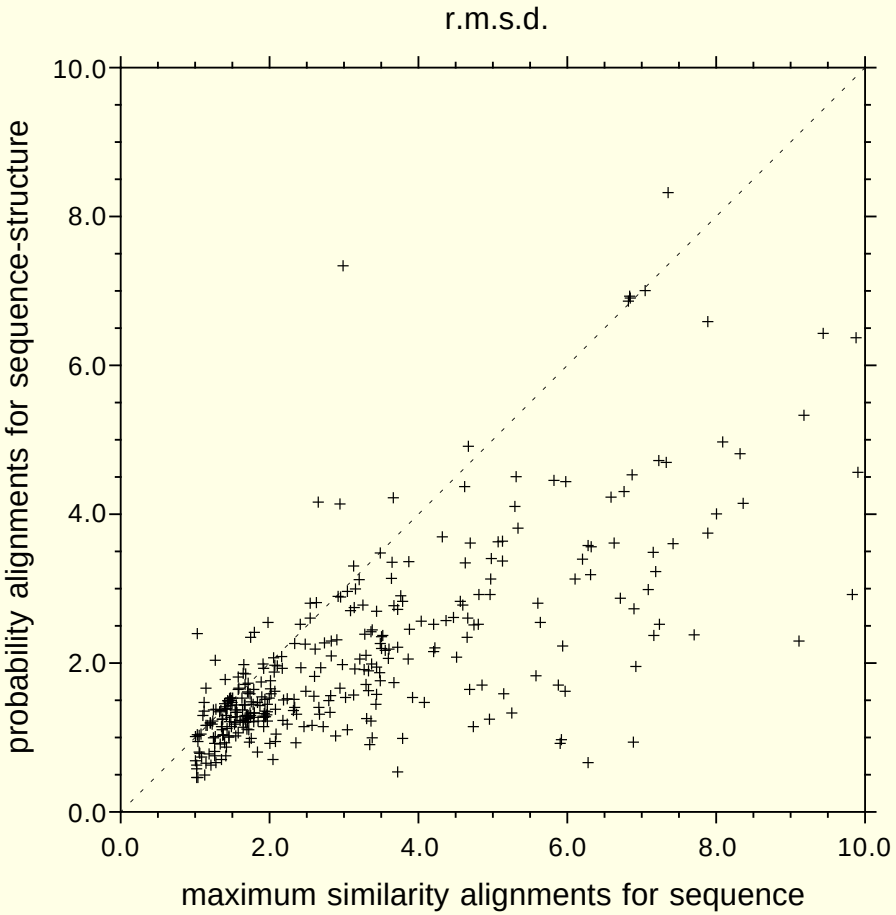


Fig. 1A 357 homologous protein pairs, which have negative minimum energy scores and positive maximum similarity scores and also whose alignments have aligned residue pairs ≥ 50, are plotted.

3.2.2 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. in superpositions of aligned residue pairs.

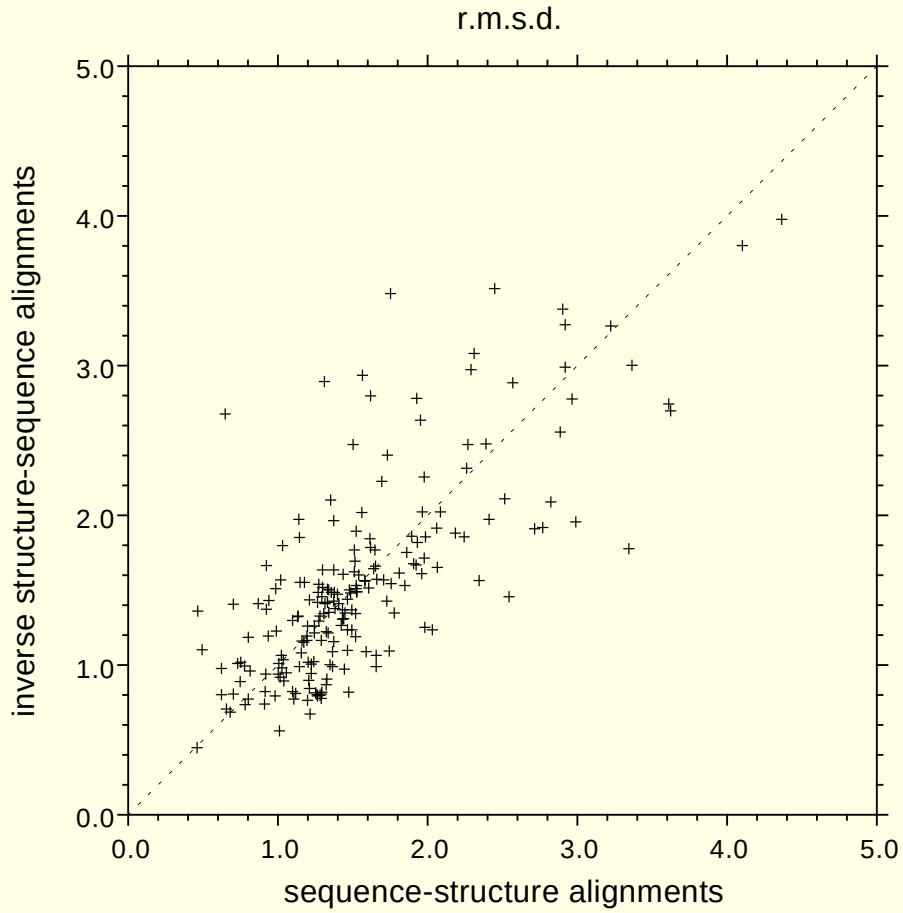


Fig. 1B The r.m.s.d. for 216 homologous protein pairs with negative energy scores and with ≥ 50 residues aligned with probabilities ≥ 0.5 are shown.

3.2.3 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.

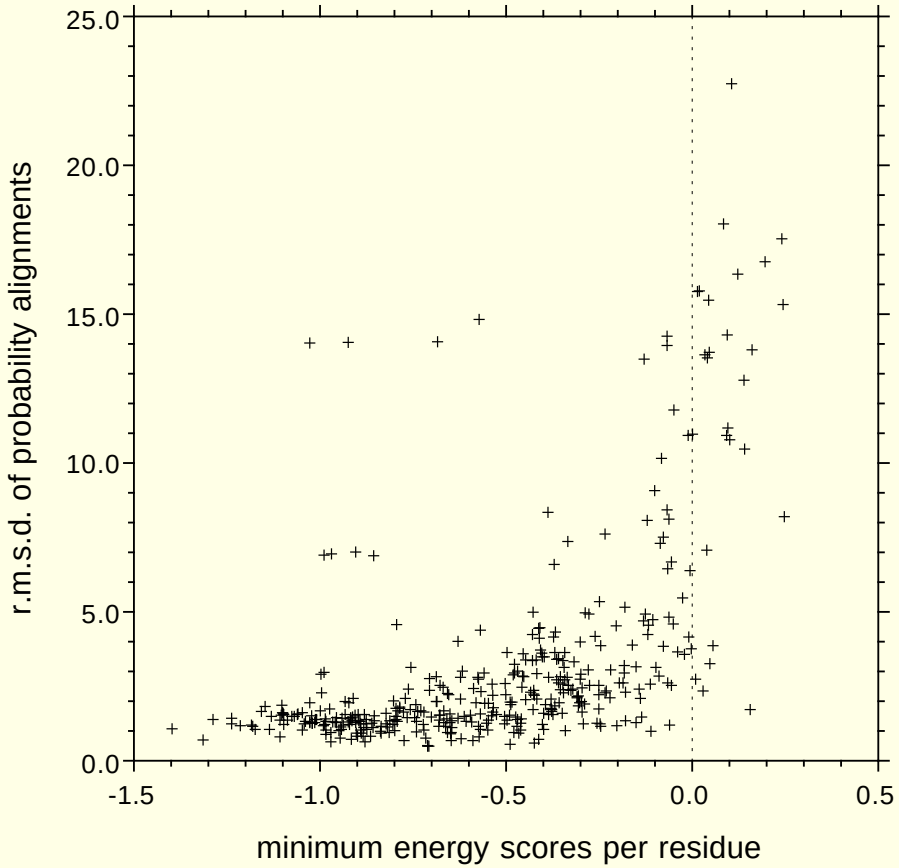


Fig. 1C 398 protein pairs whose aligned residue pairs with probability ≥ 0.5 are more than 50 are plotted.

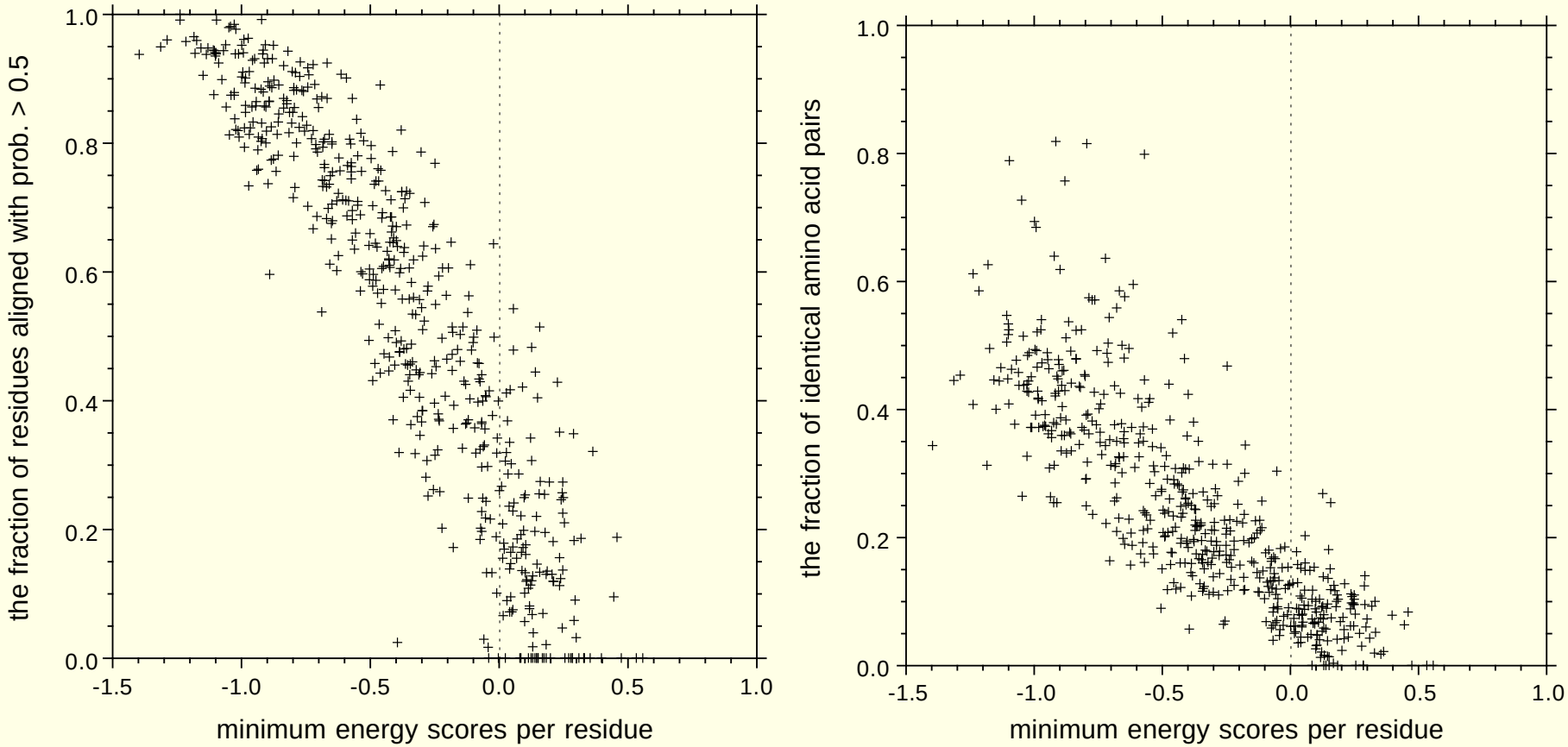


Fig. 1D and 1E Characteristics of probability sequence-structure alignments for 548 homologous protein pairs are shown.

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.

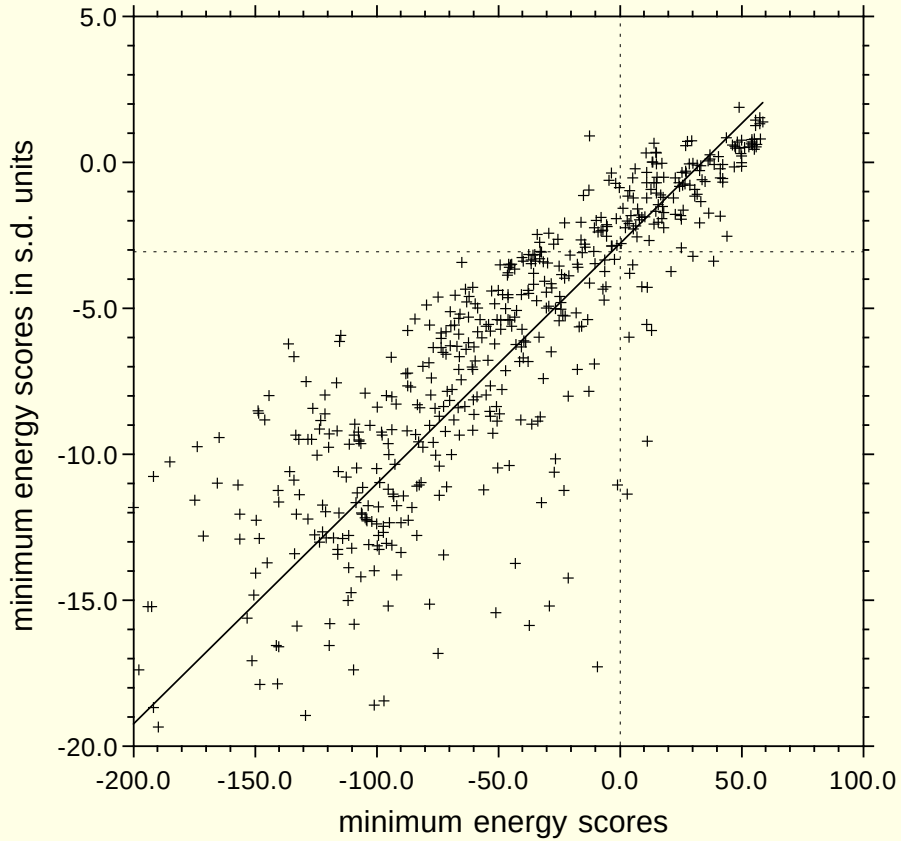
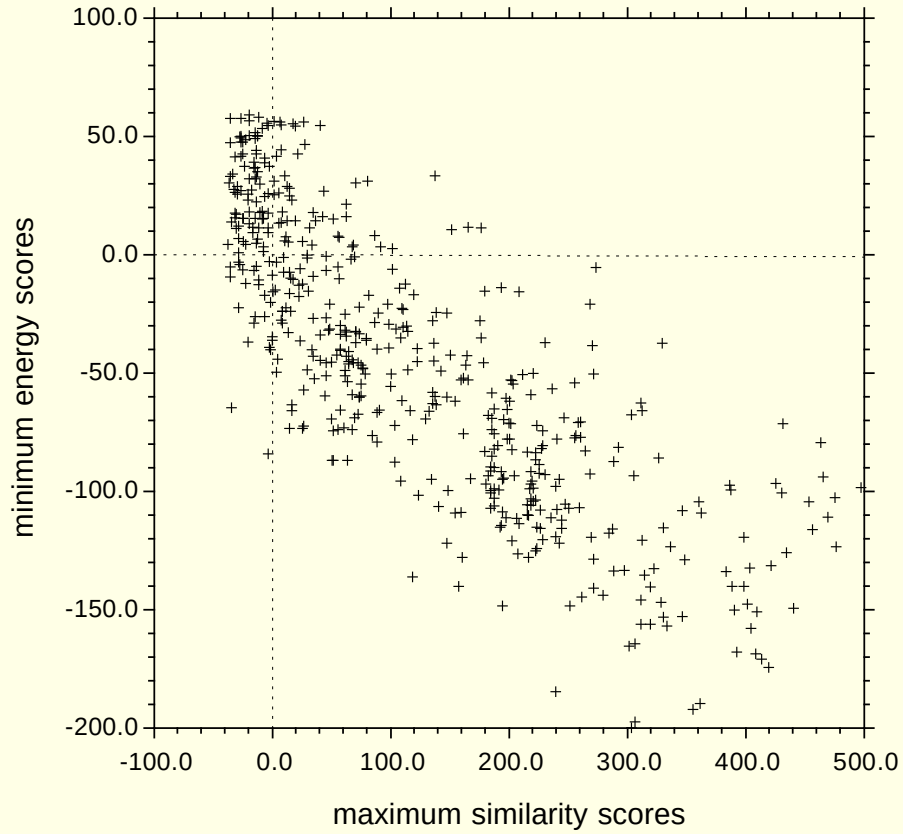


Fig. 1F Homologous protein pairs are plotted.

3.3 Detection of Homologous Proteins from Dissimilar Proteins



Homologous protein pairs are plotted.

The overall capability to identify homologous protein pairs is slightly better for the conventional sequence alignment 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

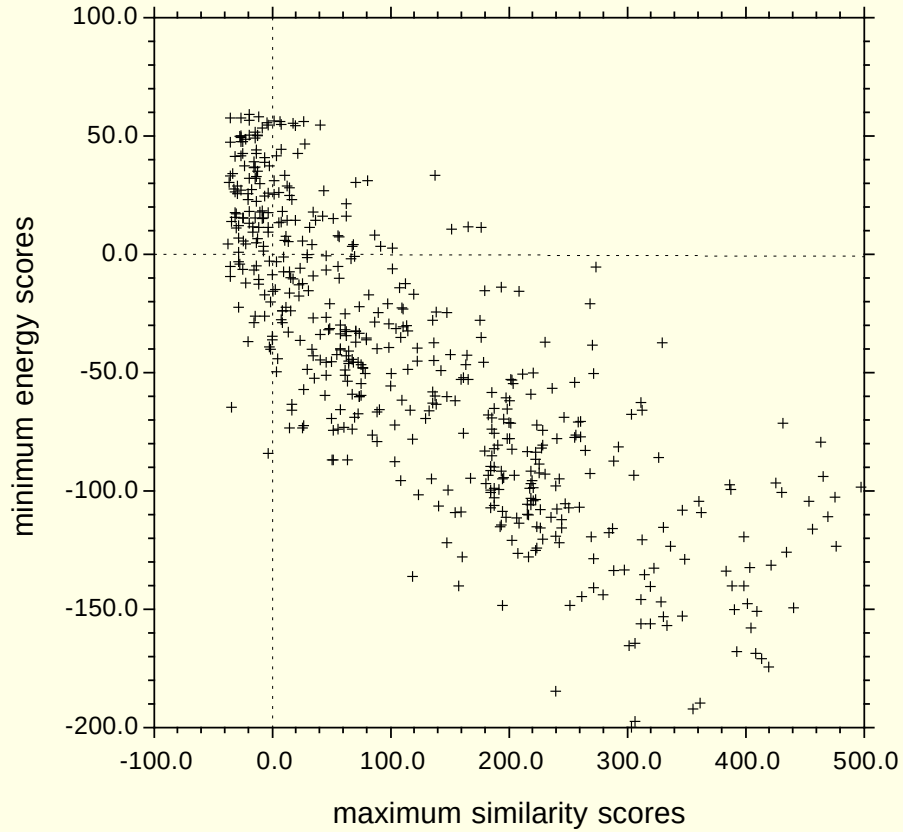
†Homologous protein pairs whose maximum similarity alignments include less than 30% identity.

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 alignment includes less than 30% identity.

Both methods can complement each other to recognize some different homologous protein pairs.



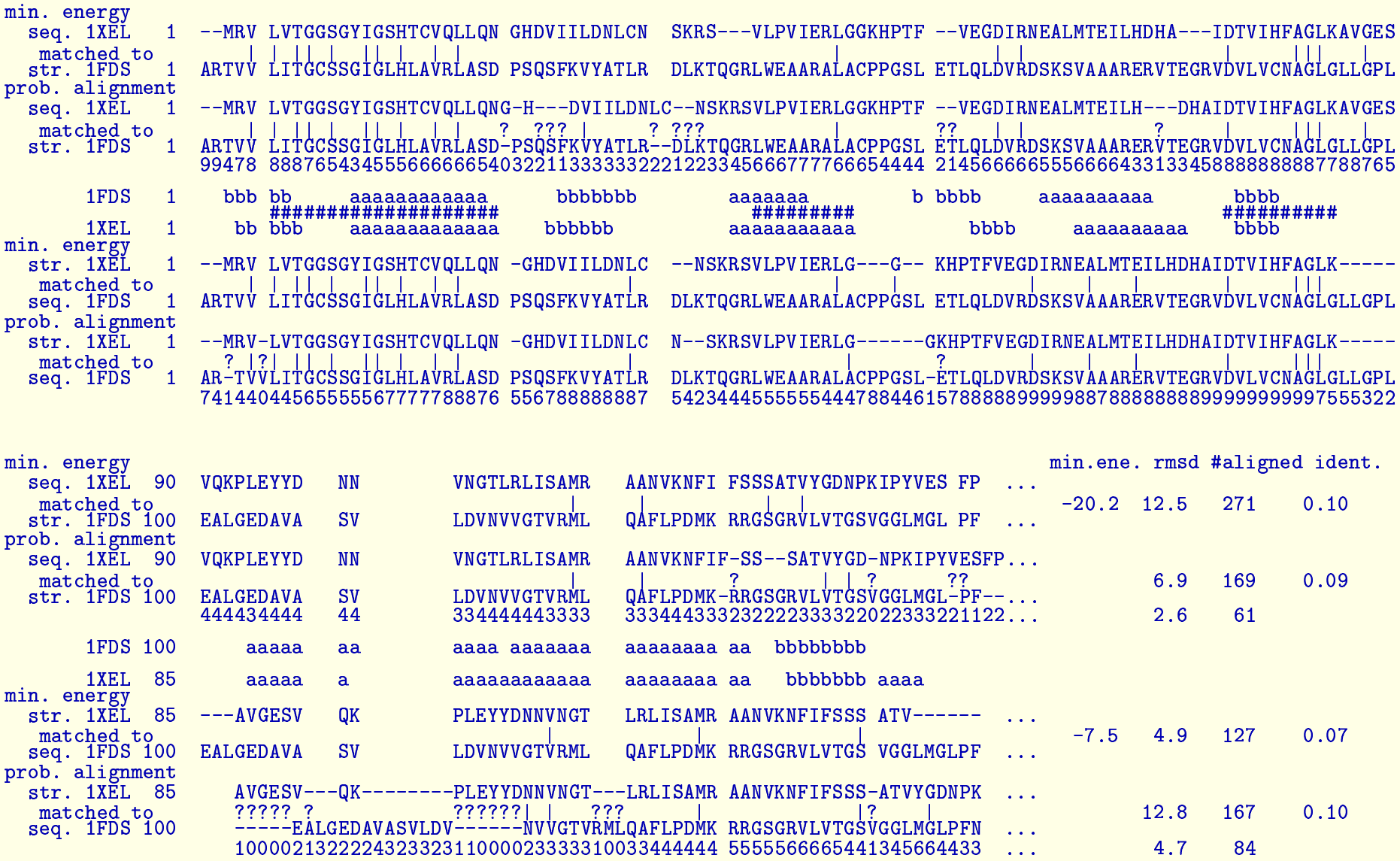
Homologous protein pairs are plotted.

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.

3.4 An Example of Sequence-Structure Alignments



4 Conclusion

- 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.
- The probability alignment method provides information about how reliable each aligned site pair is. 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.
- Remarkably, by this method some individual sequence-structure pairs are detected having only 5-20 % sequence identity.
- The present energy function and alignment method for sequence and structure can complement the conventional sequence alignment method to detect some different homologous proteins.