第 2 回「タンパク質立体構造の分類・予測・デザイン」研究会

1999 年 12 月 18 日

蛋白質の配列・構造間の適合性判定; 配列・構造アライメントと評価関数

宮澤 三造 (群馬大学工学部)

適合性評価関数としてのエネルギーポテンシャル

(Miyazawa, S. and Jernigan, R. L. (1999) *Proteins*, **36**, 357-369)

配列 i と構造 s が互いに適合するか否かは、構造 s の安定性、つまり配列 i が構造 s をとる確率に依る。

$$-\log(\text{probability of a specific conformation } s \text{ in a sequence } i)$$
$$= \beta E^{conf}(s|i) + \log(\sum_{s} \exp(-\beta E^{conf}(s|i)))$$

1. 適合性評価関数としては第 1 項 (エネルギー項) のみを考慮すればよい場合; 第 2 項 (分配関数) は 一定
   - 残基の欠失 / 挿入を許さない threading による配列 i に適合する構造の検索
2. 第 2 項 (分配関数) が無視できない場合
   - 構造 s に適合する配列の検索
   - 残基の欠失 / 挿入を許す配列・構造アライメントによる配列 i に適合する構造の検索

How to estimate the sequence dependences of partition function

In studies of the optimum sequence design of proteins

Effective Hamiltonian: consists of inter-residue pairwise contact interactions

$$H_{\text{effect}} = \sum_{p<q} e_{i_p j_q} \Delta^c_{pq}$$

1. Random Energy Model, $P(E_1, E_2) = P(E_1)P(E_2)$, (Pande, Grosberg & Tanaka, 1997; Shaknovich, 1998): The Z depends only on the amino acid composition but details of a sequence in $T > T_g$.

$$-kT \ln Z \simeq -kT \ln Z(\beta = 0) + n^c[\overline{e} - \delta e^2/(2kT)] \qquad \text{for } T > T_g$$

2. Deutsch & Kurosky, 1996: the Z is estimated by taking account of the first cumulant in a high temperature approximation.

$$-kT \ln Z \simeq -kT \ln Z(\beta = 0) + \sum_{p<q} e_{i_p j_q} < \Delta^c_{pq} >$$

<>: unbiassed averaging over all conformations.

3. Mirny, Abkevich & Shaknovich, 1996: Z score is used as a scoring function.

$$\text{Z score} = (H - n^c\overline{e})/(n^c \delta e^2)^{1/2}$$

4. Seno, Vendruscolo, Maritan & Banavar, 1996: the Z is estimated by a dual Monte-Carlo simulation.

5. Morrissey & Shaknovich, 1996: the Z is estimated in a cumulant expansion approximation.

   - The n-th moment of the single contact energy is calculated in the mean field.
   - Each cumulant is calculated from moments of the single contact energies.

The following approximation is used here.

In the sum of Boltzmann factors over all conformations, only dominant terms, i.e., native-line conformations are taken into accounts, and then the log of the partition function is estimated in a high temperature expansion.

$$\log(\sum_s \exp(-\beta E^{conf}(s|i)))$$

$$\simeq \log(\sum_{s \in \text{native-like conformations}} \exp(-\beta E^{conf}(s|i)))$$

$$= \log(\sum_{s \in \text{native-like conformations}} 1) + \frac{\sum_{s \in \text{native-like conformations}} (-\beta E^{conf}(s|i))}{\sum_{s \in \text{native-like conformations}} 1} + \cdots$$

$$= n_r \sigma - \beta < E^{conf}(s|i)) >_{\beta=0, \text{native-like}} + \cdots$$

where

n_r**:** chain length

$k \cdot \sigma$**:** conformational entropy per residue in native-like structures

Then,

$$-\log(\text{probability of a specific conformation } s \text{ in a sequence } i)$$

$$\simeq \beta E^{conf}(s|i)$$

$$-\beta(E^{conf}{}_{\text{of a typical native structure with the same amino acid composition}}) + n_r \sigma$$

$$= \Delta E^{conf}(s|i) + n_r \sigma$$

$$\Delta E_p^{conf} \equiv (E_p^{conf} - < E_{i_p}^{conf} >_{\text{native structures}})$$

One of typical comparisons in sequence - structure compatibility ranking:

- Which is more compatible for myoglobin sequence, myoglobin native structure versus α hemoglobin monomer in a tetrameric state ?

$$
\begin{aligned}
& E^{\text{intra}}(\text{myoglobin threaded into myoglobin}) \\
& > E^{\text{intra}}(\text{myoglobin threaded into } \alpha \text{ hemoglobin}) \\
& \quad + E^{\text{between subunit}}(\text{myoglobin } \alpha\beta^2)
\end{aligned}
$$

Further modifications on energy potentials are needed for a scoring function.

- Subtracting a collapse energy from contact energy

  Only interaction energies depending on side chains are included.

$$E_p^c(e_{ij} - err) \equiv \frac{1}{2}\sum_q (e_{i_p j_q} - e_{rr})\Delta_{pq}^c$$

$$\Delta E_p^c(e_{ij} - err) = \frac{1}{2}\sum_q (e_{i_p j_q} - e_{rr})\Delta_{pq}^c - \frac{1}{2}\sum_j (e_{i_p j} - e_{rr}) N_{i_p j}/N_i$$

- Excluding intrinsic and backbone-backbone secondary structure energies

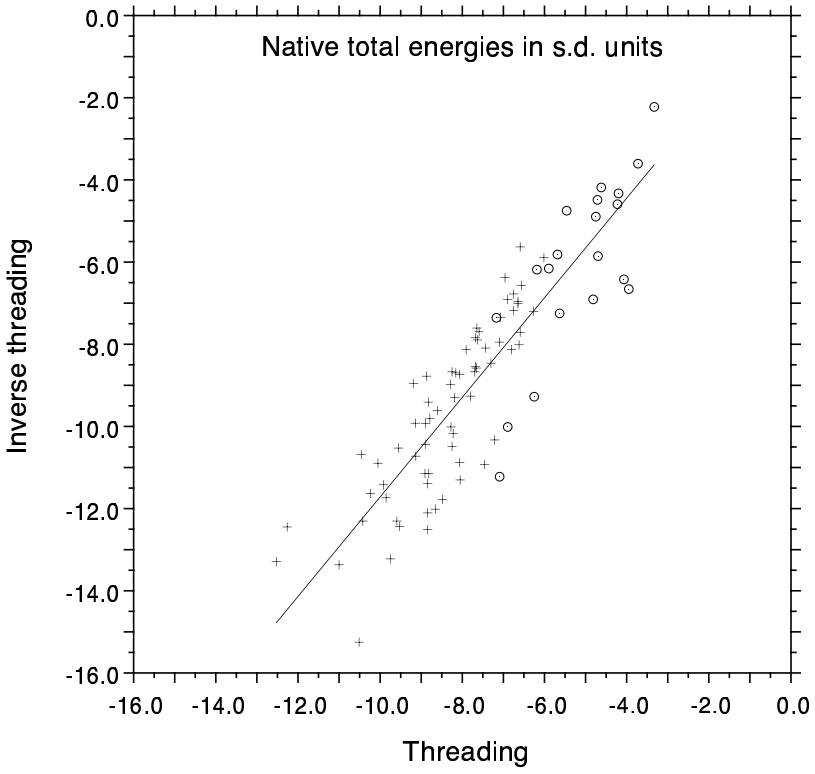
  Only interaction energies depending on side chains are included.

There modifications do not change ranking scores for sequences compatible to a protein structure.

Comparison between threading and inverse threading

Threading: recognition of the native structure in structure space with a given sequence.

Inverse threading: recognition of the native sequence in sequence space with a given structure.



Sequence-structure alignments

A specific sequence−structure alignment A:

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

A probability $\mathcal{P}(\{s_p\}|\{i_q\}, A)$ with which the aligned sequence A takes a specific conformation $\{s_p\}$: as follows.

$$-\log\{\mathcal{P}(\{s_p\}|\{i_q\}, A)\} \approx \beta \sum_{(p,q)\in A} \Delta E_p^{\rm conf}(\{s_p\}|i_q, A) + n_r^{\rm aligned}\sigma$$

According to Bayesian statistics, the conditional probability of an alignment A for a given structure $\{s_p\}$:

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

where the *a priori* probability of an alignment A, $\mathcal{P}(A)$, is represented as gaps.

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

where $\mathcal{W}$ is a positive quantity to represent a gap penalty, and $\mathcal{E}_0$ is a negative constant as a scaling parameter.

Thus, the probability of an alignment A for a given structure s_p is

$$\begin{aligned}\mathcal{P}(A|\{s_p\},\{i_q\}) &= \frac{1}{\mathcal{Z}}\exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \\ \mathcal{Z} &= \sum_A \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)]\end{aligned}$$

The energy score $\mathcal{E}(\{s_p\}|\{i_q\},A)$ of an alignment A for a given structure $\{s_p\}$:

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

where

$$\mathcal{E}(\{s_p\}|i_q,A) \equiv \Delta E_p^{conf}(\{s_p\}|i_q,A) + \mathcal{E}_0$$

How to evaluate pairwise interactions.

$$\mathcal{E}(\{s_p\}|i_q,A) \approx \mathcal{E}(\{s_p\}|i_q,\mathcal{P}(p',q'))$$

The probability for a site pair (p,q) to be aligned:

(Miyazawa, S. *Prot. Eng.*, (1995) **8**, 999-1009)

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

where

$$\mathcal{Z} = \mathcal{Z}_{n_r^{str},n_r^{seq}} = \mathcal{Z}'_{1,1}$$

An iteration method is used to obtain a self-consistent solution for $\mathcal{P}(p,q)$.

How to calculate optimum alignments.

Minimum energy score alignment:

$$\begin{aligned}\mathcal{E}(\{s_p\}|\{i_q\}, A^{\min}) &= \min_A \mathcal{E}(\{s_p\}|\{i_q\}, A) \\ &\approx \min_A [\sum_{(p,q)\in A} \mathcal{E}(\{s_p\}|i_q, \mathcal{P}(p', q')) + \sum_{\text{all gaps in } A} \mathcal{W}]\end{aligned}$$

Probability alignment: by successively aligning a site pair in order of pairwise alignment probabilities, $\mathcal{P}(p, q)$.

(Miyazawa, S. *Prot. Eng.*, **8**, 999-1009. 1995)

How to measure sequence-structure compatibilities.

The following energies may be used.

$$\begin{aligned}\mathcal{E}_{\min} &\equiv \min_A \mathcal{E}(\{s_p\}|\{i_q\}, A) \\ \mathcal{F} &\equiv -\frac{1}{\beta}\log \mathcal{Z} \\ < \mathcal{E}(\{s_p\}|\{i_q\}, A) >_A &= -\frac{1}{\mathcal{Z}}\frac{\partial \mathcal{Z}}{\partial \beta}\end{aligned}$$

Here, $\mathcal{E}_{\min}$ is used.

Energy function:

$$\Delta E_p^{conf}(\{s_p\}|i_q,\mathcal{P}(p',q')) \equiv \Delta E_p^{sec}(\{s_p\}|i_q,\mathcal{P}(p',q')) + \Delta E_p^{tert}(\{s_p\}|i_q,\mathcal{P}(p',q'))$$

where

$$\begin{aligned}
\Delta E_p^{sec}(\{s_p\}|i_q,\mathcal{P}(p',q')) &\simeq \Delta e^s(\ldots,s_{p-1},i_q,s_p,s_{p+1},\ldots) \\
&\equiv \sum_{p-3\le r\le p+3} \delta e^s(i_q,s_{r-1},s_r,s_{r+1}) \\
\Delta E_p^{tert}(\{s_p\}|i_q,\mathcal{P}(p',q')) &= \Delta E_p^c(\{s_p\}|i_q,\mathcal{P}(p',q')) + \Delta E_p^r(\{s_p\}|i_q,\mathcal{P}(p',q')) \\
&\qquad \text{with } (e_{i_q j} - e_{rr})
\end{aligned}$$

Secondary structure potentials:

Miyazawa, S. and Jernigan, R. L. (1999) *Proteins*, **36**, 347-356.

Contact energies:

Miyazawa, S. and Jernigan, R. L. (1999) *Proteins*, **34**, 49-68.

Repulsive packing energies:

Miyazawa, S. and Jernigan, R. L. (1996) *J. Mol. Biol.*, **256**, 623-644.

Gap penalty for sequence-structure alignments

In order to place gaps more frequently on protein surfaces than in cores, gap penalties are assumed to be proportional to the number of contacts at each residue position.

Heuristic knowledge about gap penalties in conventional sequence alignments is used in sequence-structure alignments.

Table II. Gap parameters used in sequence-structure alignments.†

gap penalty	value in RT 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
k sequence insertions between q and $q+1$	$5.5 + k(1.05 + 0.43(1 + (n_q^c + n_{q+1}^c)/2))$ in the middle
	$3.25 + k(0.53 + 0.22(1 + n_{terminal}^c))$ 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

† These parameter values are for a case in which secondary structure energies, contact energies, and repulsive energies are all included.

n_p^c is the number of residues in contact with the pth residue.

Identifying sequence-structure pairs undetected by sequence alignments

Datasets of protein structures

Release 1.35 of SCOP database is used for the classification of protein folds.

Protein structures are used:

- proteins which belong to classes 1 to 5, all α, all β, α/β, $\alpha+\beta$, and multi-domain proteins,
- and which were analyzed by X-ray and whose resolutions are better than 2.5 Å,
- excluding protein structures which lack many atoms or residues
- and which are shorter than 50 residues.

Two kinds of dataset of protein structures are prepared:

- Dataset of 548 homologous protein pairs:
  made by pairing the protein representatives of families with those of different domains within the families.
- Dataset of 505 dissimilar protein pairs:
  made by arbitrarily choosing only every $100th$ pair from the ordered list of all possible pairs of superfamily representatives.

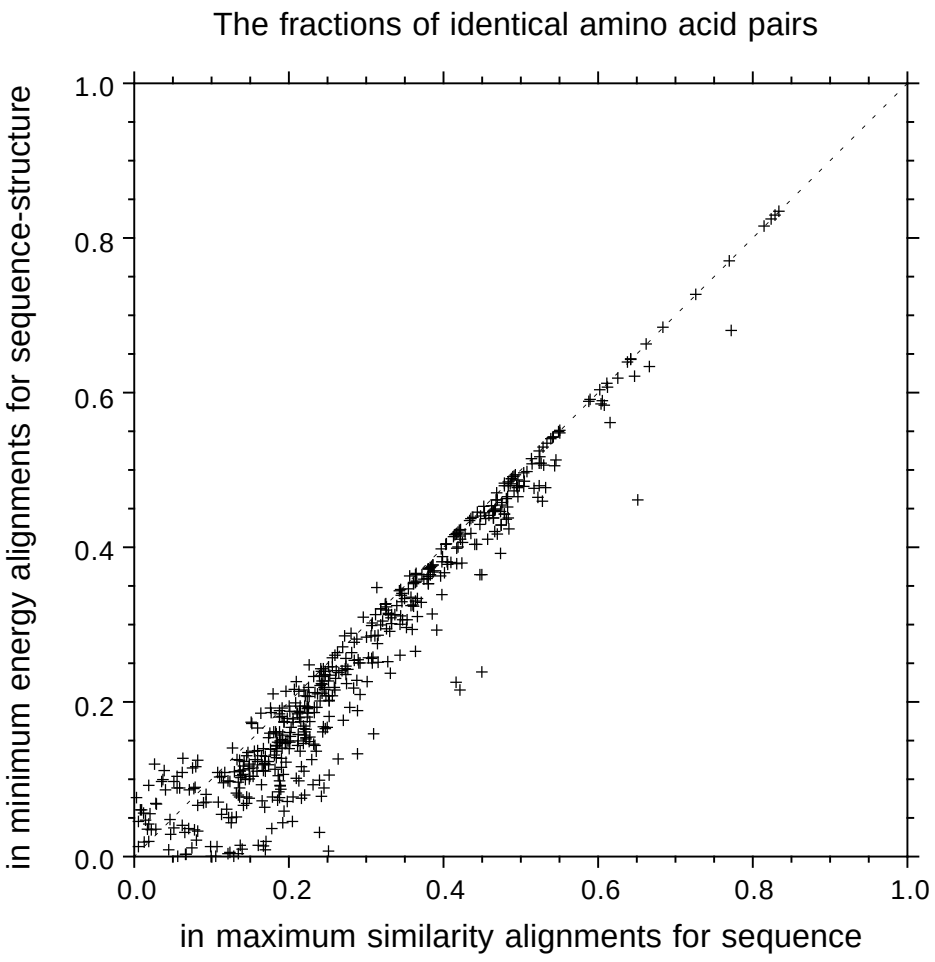
Adequacy of sequence - structure alignments

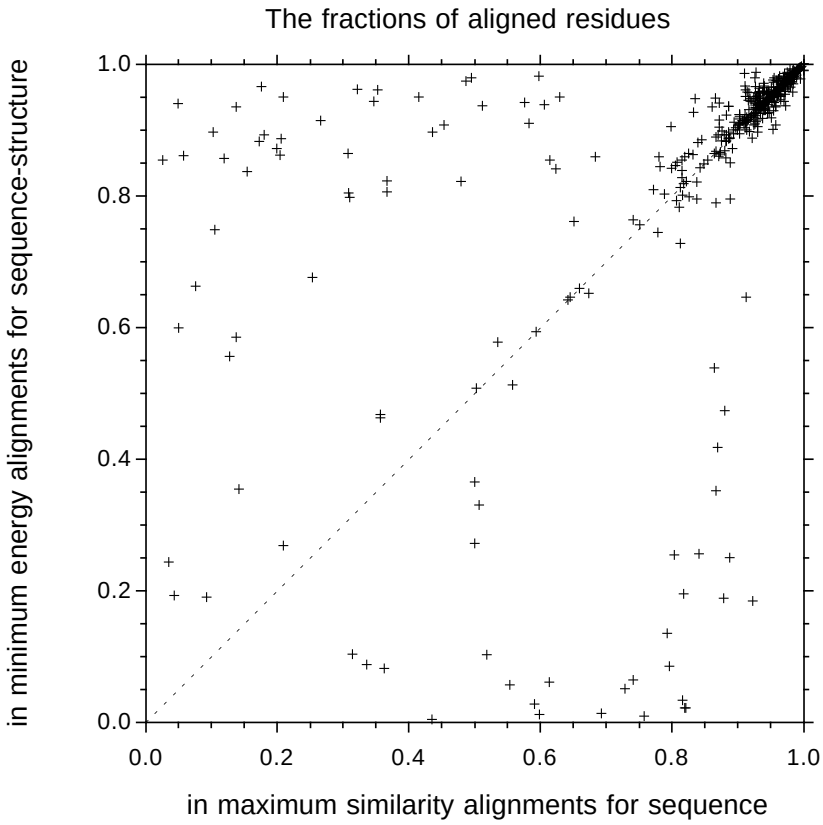
Sequence - structure alignments have similar overall characteristics to conventional sequence alignments, such as

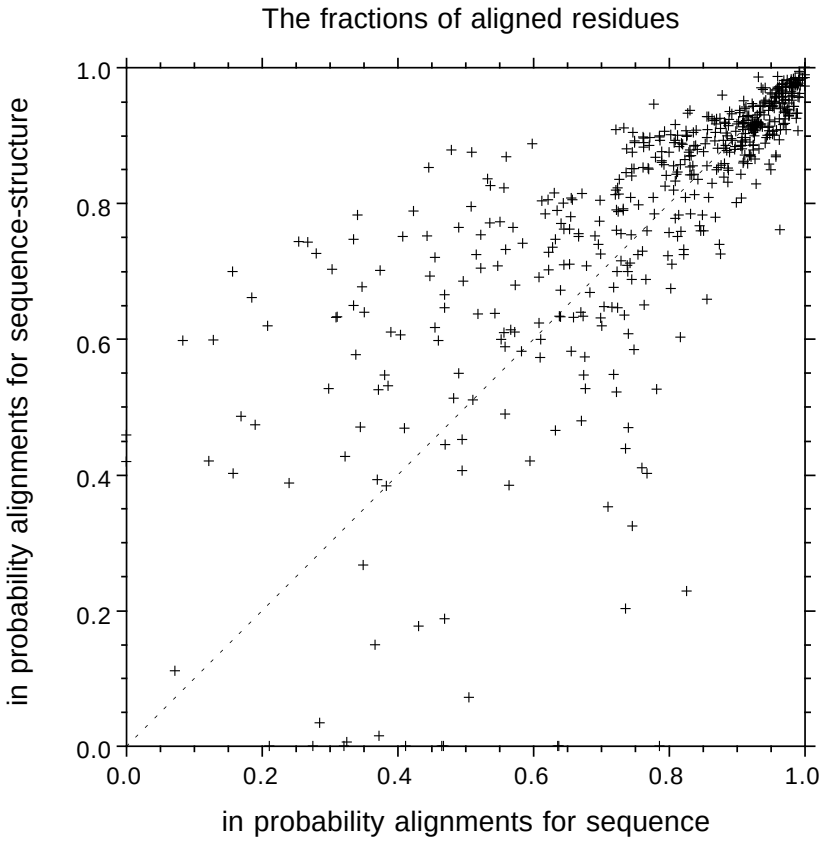
- the fraction of aligned residues,
- the fraction of identical residues,

indicating the values of gap parameters to be appropriate.

The relative temperature ($1/\beta$) has been adjusted so that similar fractions of residues are aligned in the probability alignments for both sequence-structure and sequence-sequence alignments.



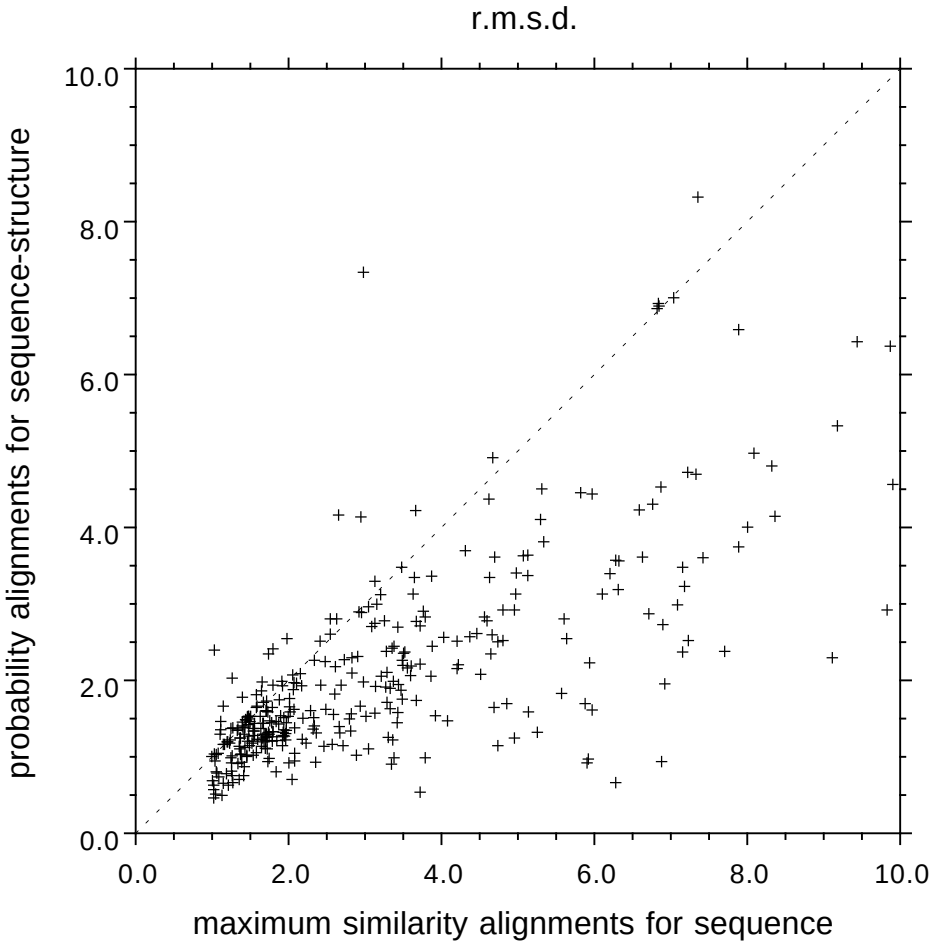

The fractions of aligned residues
in minimum energy alignments for sequence-structure
in maximum similarity alignments for sequence
0.0
0.2
0.4
0.6
0.8
1.0


The fractions of aligned residues
in probability alignments for sequence-structure
in probability alignments for sequence
0.0
0.2
0.4
0.6
0.8
1.0

RMSDs in superposition of aligned residues for homologous protein pairs

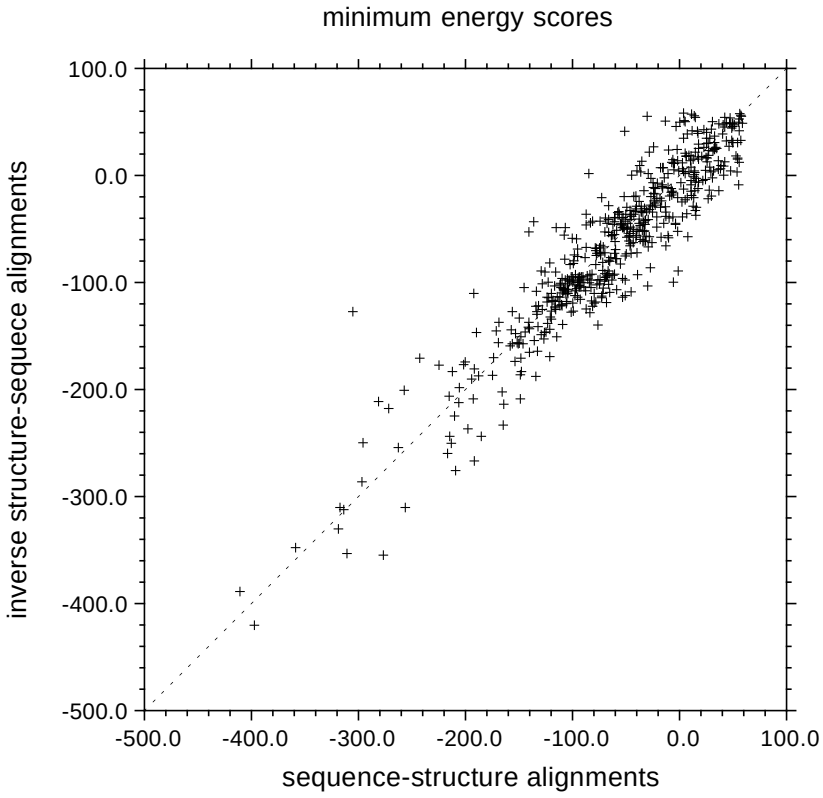
The RMSDs are improved in probability alignments for sequence-structure. than in maximum similarity alignments for sequence.

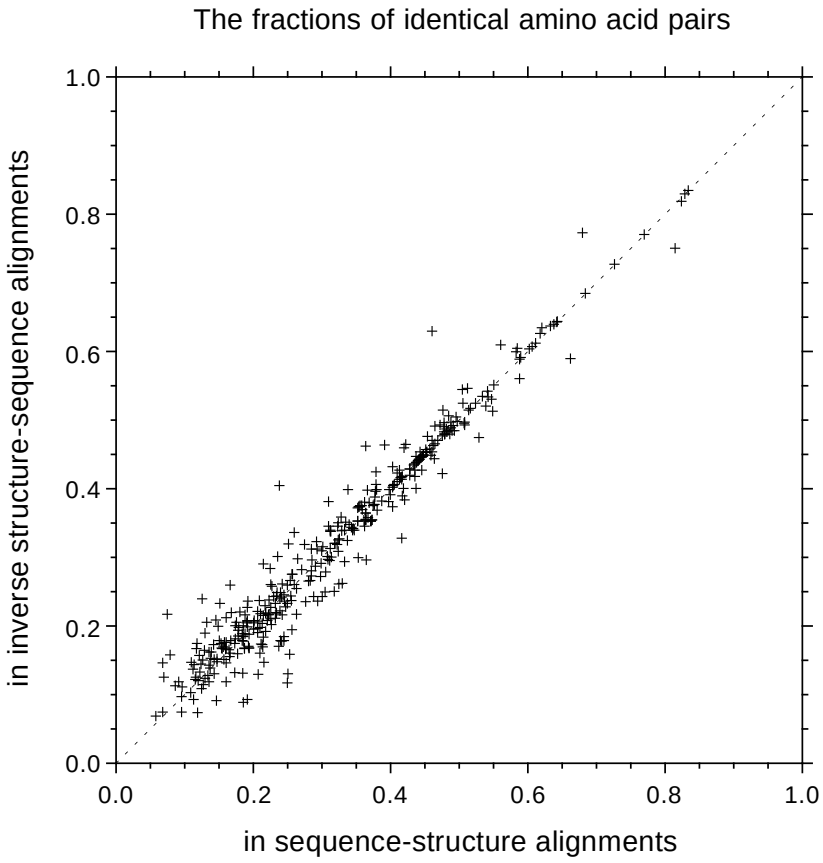
To reduce the effects of fewer aligned residue pairs on RMSD, 357 homologous protein pairs whose alignments have more than 50 aligned residue pairs are used here.



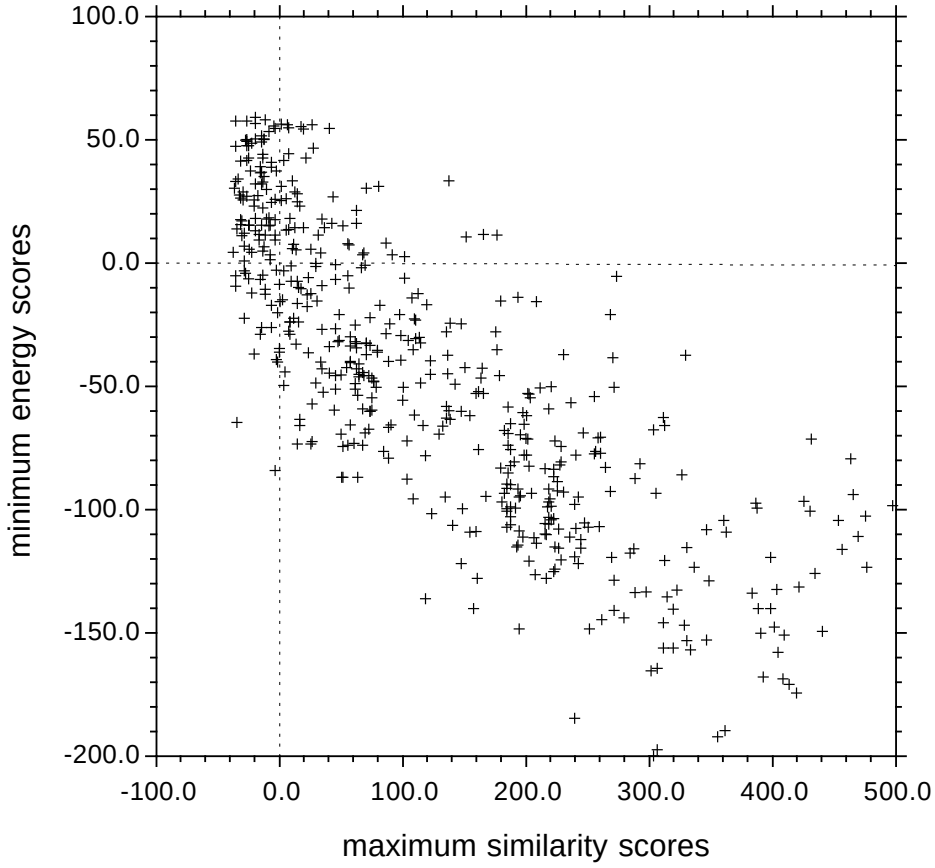
Comparison of two types of sequence - structure alignment

Overall characteristics and minimum energy scores of alignments are similar for both types of alignment, sequence-structure and inverse structure-sequence.

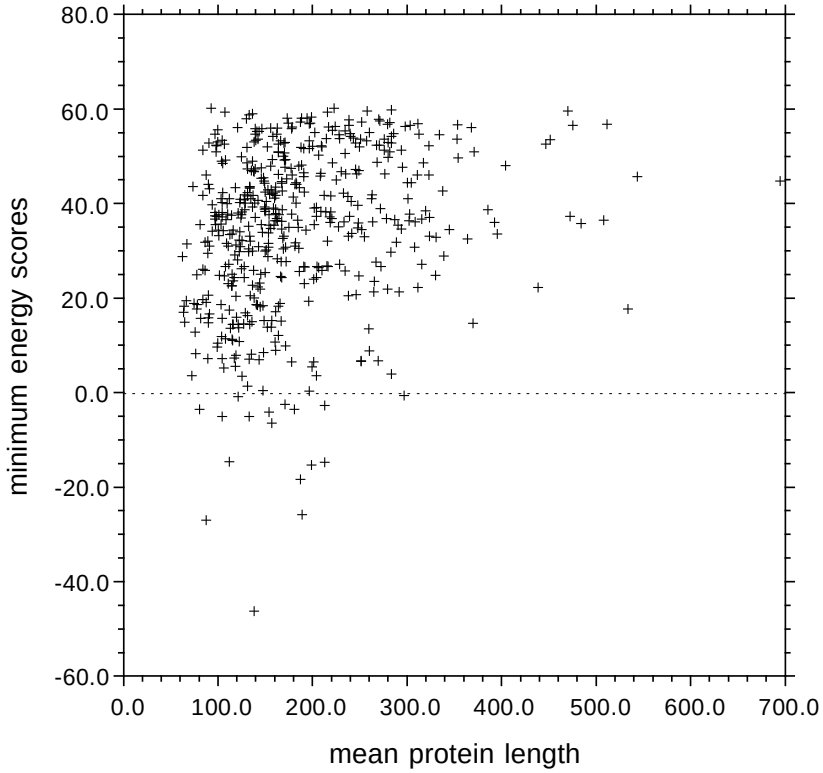




Detection of homologous proteins from dissimilar proteins



Dissimilar protein pairs:



Homologous protein pairs:

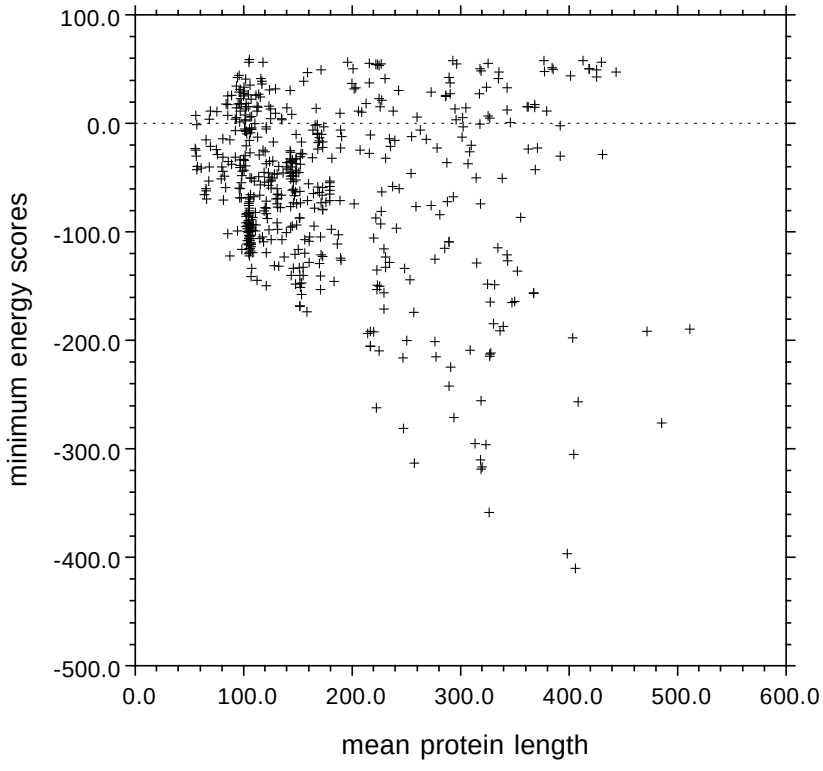


Table IV. Recognition of homologous protein pairs from dissimilar protein pairs by sequence-structure alignments.

false positives in 548 homologous protein pairs	false negatives in 505 dissimilar protein pairs	method
106	5	conventional sequence alignment
131	17	sequence-structure alignment
126	24	inverse structure-sequence alignment
191	19	sequence-structure alignment without secondary structure energies
164	26	inverse structure-sequence alignment without secondary structure energies

Table V. Protein pairs† whose compatibilities are not identified by sequence alignments but by sequence-structure alignments.

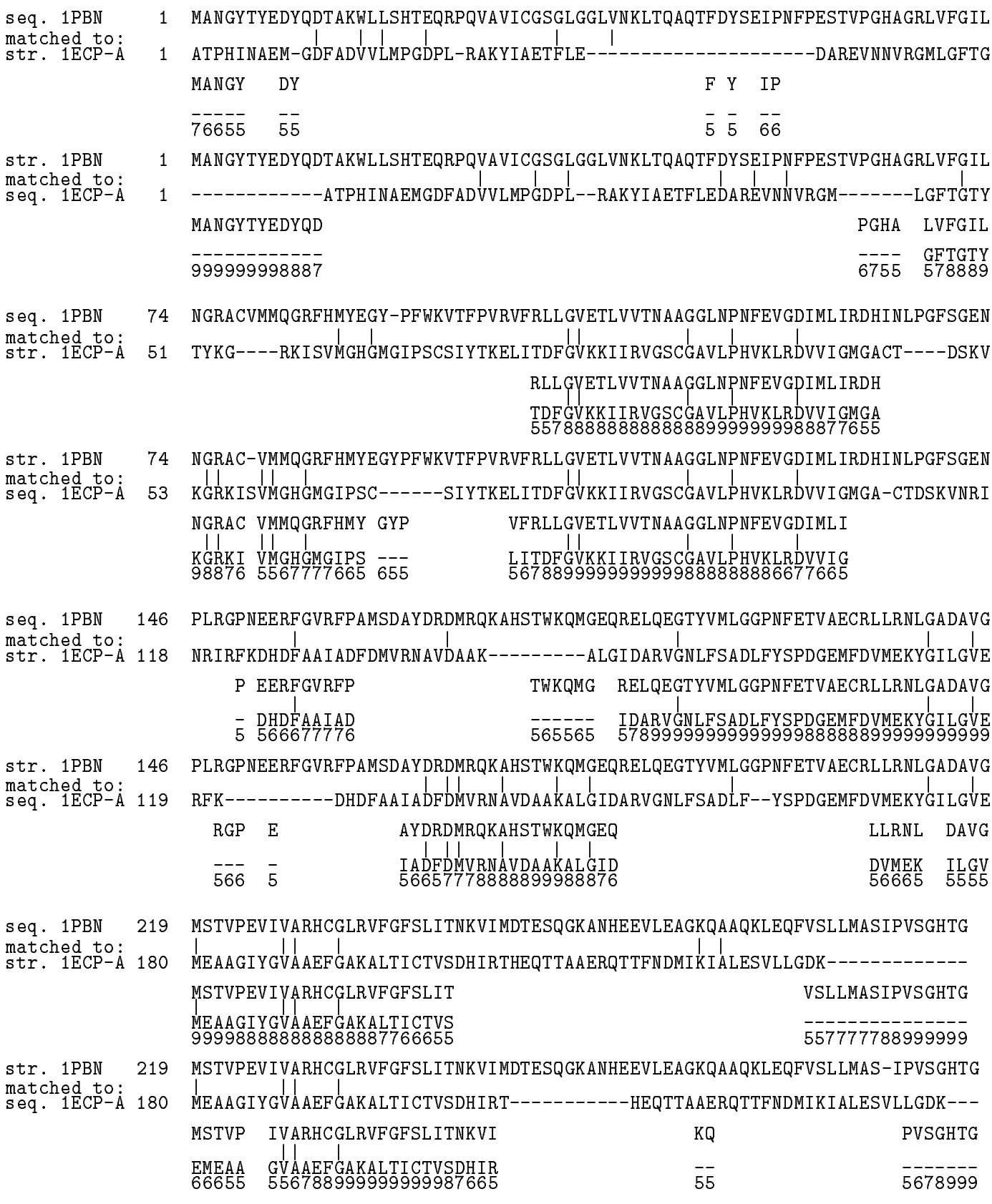
sequence	length	structure	length	sequence-structure			
				minimum energy alignment		probability alignment	
				minimum energy score	fraction of identical residues	no. of residues† aligned with probabilities ≥ 0.5	r.m.s.d.
1ECF-A:250-469	220	1HMP-A	214	−10.7	0.11	88	4.6
1NCX	162	2SAS	185	−17.3	0.10	85	9.1
1PBN	289	1ECP-A	237	−6.5	0.09	99	5.4
1PII:1-254	254	1TTQ-A	256	−12.3	0.12	62	11.8
1PTV-A	297	1YTS	278	−36.2	0.11	105	4.9
1XEL	338	1ENY	268	−3.1	0.05	57	10.9
1XEL	338	1FDS	282	−20.2	0.10	61	2.6
2DRI	271	2LBP	346	−26.4	0.13	157	7.3
2DRI	271	2LIV	344	−37.1	0.11	165	8.1
2HVM	273	1NAR	289	−84.2	0.11	103	4.0
2HVM	273	2EBN	285	−22.7	0.10	111	10.1
2OHX-A:175-324	150	1QOR-A:136-265	130	−40.2	0.21	99	4.9
3GRS:364-478	115	1NPX:322-447	126	−26.4	0.12	73	3.0
8FAB-A:3-105	103	1HNF:4-104	101	−39.3	0.11	61	2.8

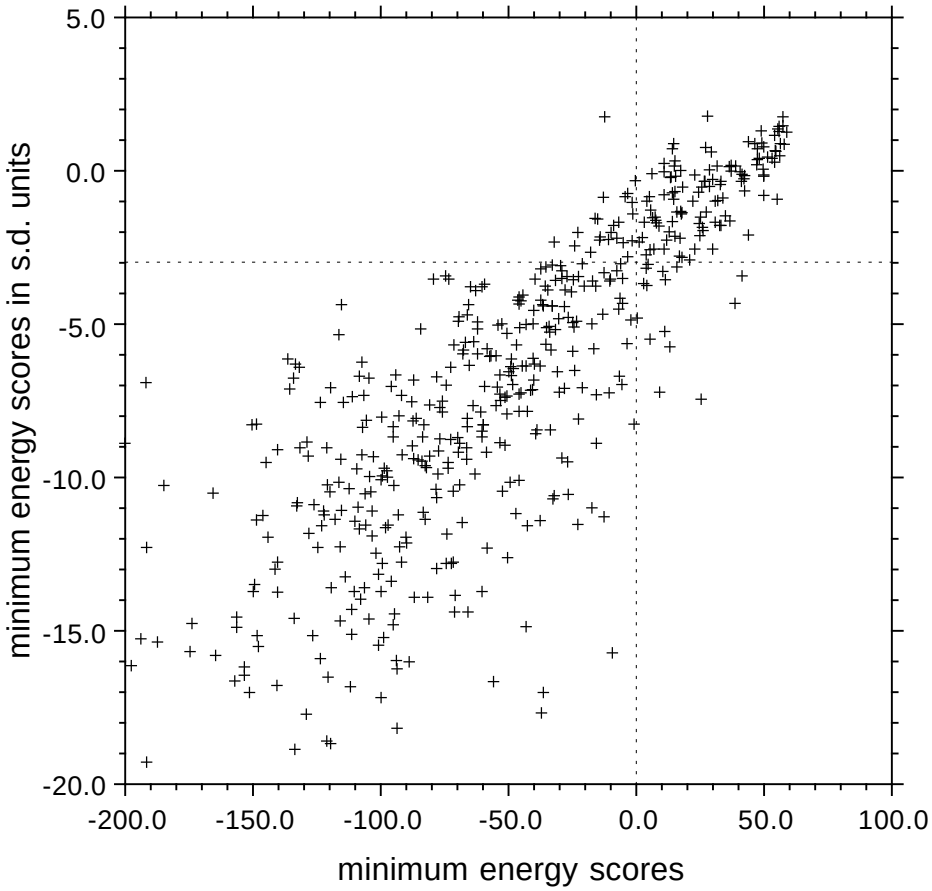
† Only protein pairs with 50 or more aligned residue pairs are listed in this table.

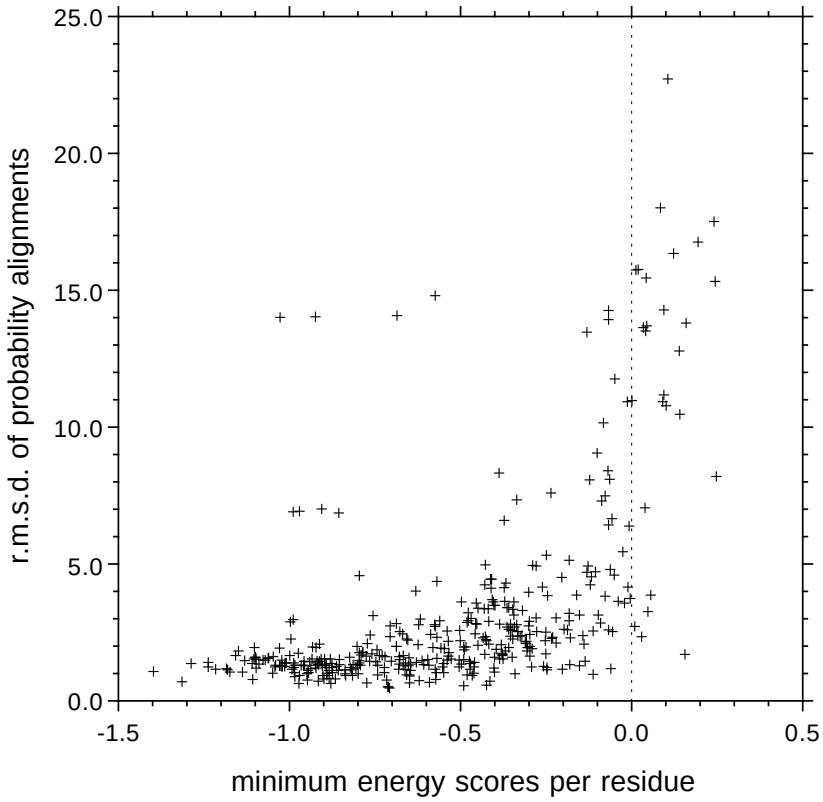
Sequence-structure pairs undetected by sequence alignments

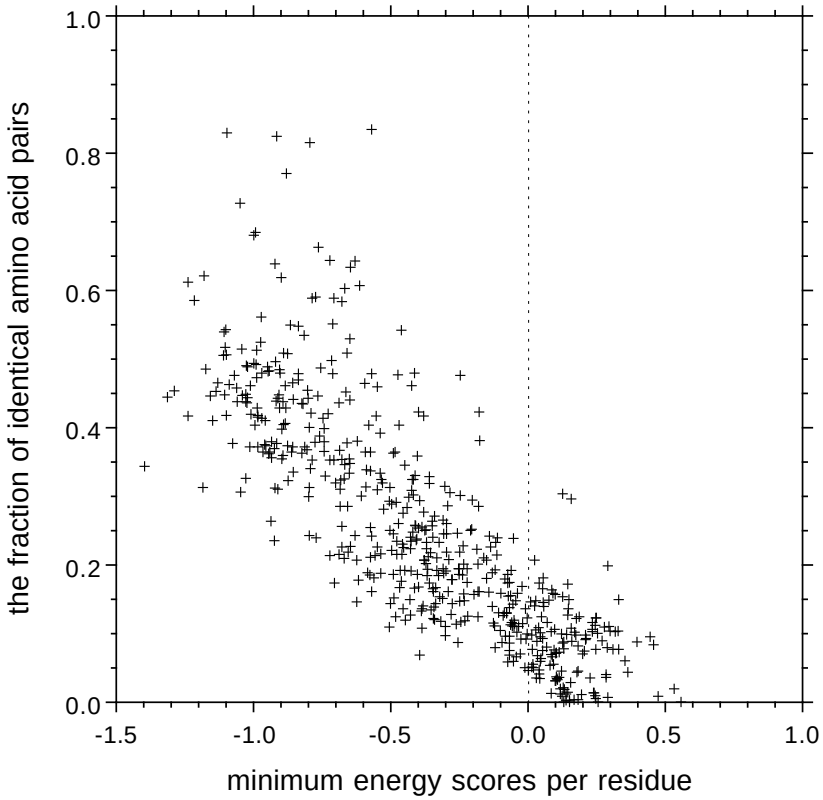
```
minimum energy alignment
    sequence  3GRS 364    YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                             |  |  |  | |    |                 | |
    structure 1NPX 322    GVQGSSGLAVFDYKFASTGINEVMA-QKLGK-ETKAVTVV -EDYLMDFNPDKQKAWF
probability alignment
    sequence  3GRS 364    YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                      ?      |  |  ?  | | |  |   ?             | |
    structure 1NPX 322    GVQGSSGLAVFDYKFASTGINEVM-AQKLGKE-TKAVT-V VEDYLMDFNPDKQKAWF
                          7777664334334698999887541577776424333203 34444444455566666

              1NPX 322       bbbbbbbbbbbbbbbbbaaaa aaaaa   bbbbb b bbbbb       bbbbbb
                            ######                 ######                   #########
              3GRS 364        bbbbbb b bbbbbbbaaaaaaaaa   bbbbbbbb bb            bbbb
minimum energy alignment
    structure 3GRS 364    YNNIPTVVFSHP PIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                                      | | |            |       | |
    sequence  1NPX 322    GVQGSSGLAVFD YKFASTGINEVMAQKLGKETKAVTVVE DYLMDF--NPDKQKAWF
probability alignment
    structure 3GRS 364  --YNNIPTVVFSHP-PIGTVGLTEDEAIHKYGIENVKTYSTS-FTPMYHAVTKRKTKCVM
    matched to:           ??          ?     |  |?    | | |       ?? ???       | |
    sequence  1NPX 322  GV--QGSSGLAVFDYKFASTGINE-VMAQKLGKETKAVTVVEDY---LMDFNPDKQKAWF
                        432233444444430345556554145667776543322222021112233566677777


minimum energy alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHMQG-LGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTSSEE   L
    matched to:         | |         |                  |     |  |         |
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKP   W
probability alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHM-QGLGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTS-SEE- L
    matched to:         | |         |   ?              |     |  |    ?    |      ?
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWN I
                        66666667777776540456799999988888888888888887643445543322221 2

              1NPX 376  bbbbb    bbbbbbbbbb      aaaaaaaaaa  aaaaaaaa              a
                        #####              ##########################
              3GRS 420  bbbbb b   bbbbbbbbbb      aaaaaaaaa
minimum energy alignment
    structure 3GRS 420  KMVCA-NKEEKVVGIHMQGLGCDEMLQGFAVAVKMGATKADFDNT----VAIHPTSSEEL
    matched to:         | |             |              |     |  |
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWNII
probability alignment
    structure 3GRS 420  KMVCANKEEKVVG-IHMQGLGCDEMLQGFAVAVKMGATKADFDN----TVAIHPTSSEEL
    matched to:         | |         |?  |              |     |  |       ?     ??????
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFF------------
                        776653443444303444334444445555567778888887643356222222111100


minimum energy alignment                               min.ene. rmsd #aligned identities
    sequence  3GRS 475  VTLR             -----------
    matched to:                                         -26.4    3.9  112       0.12
    structure 1NPX 433  NIIN             TAALEAVKQER
probability alignment
    sequence  3GRS 475  VTLR            ------------
    matched to:          ??                                      3.7  108       0.12
    structure 1NPX 435  I---            NTAALEAVKQER
                        2011            246789999999                     3.0   73

              1NPX 435  a               aaaaaaaaaa
                                            ########
              3GRS 475
minimum energy
    structure 3GRS 475  VTLR                --------
    matched to:          |                              -20.0    4.3  113       0.11
    sequence  1NPX 436  NTAA                LEAVKQER
probability alignment
    structure 3GRS 475  VTLR------------------------
    matched to:         ???                                      3.5   92       0.12
    sequence  1NPX 424  ----FQPAFDKPWNIINTAALEAVKQER
                        0112533343233344344577788999             3.0   45
```

Sequence-structure pairs undetected by sequence alignments



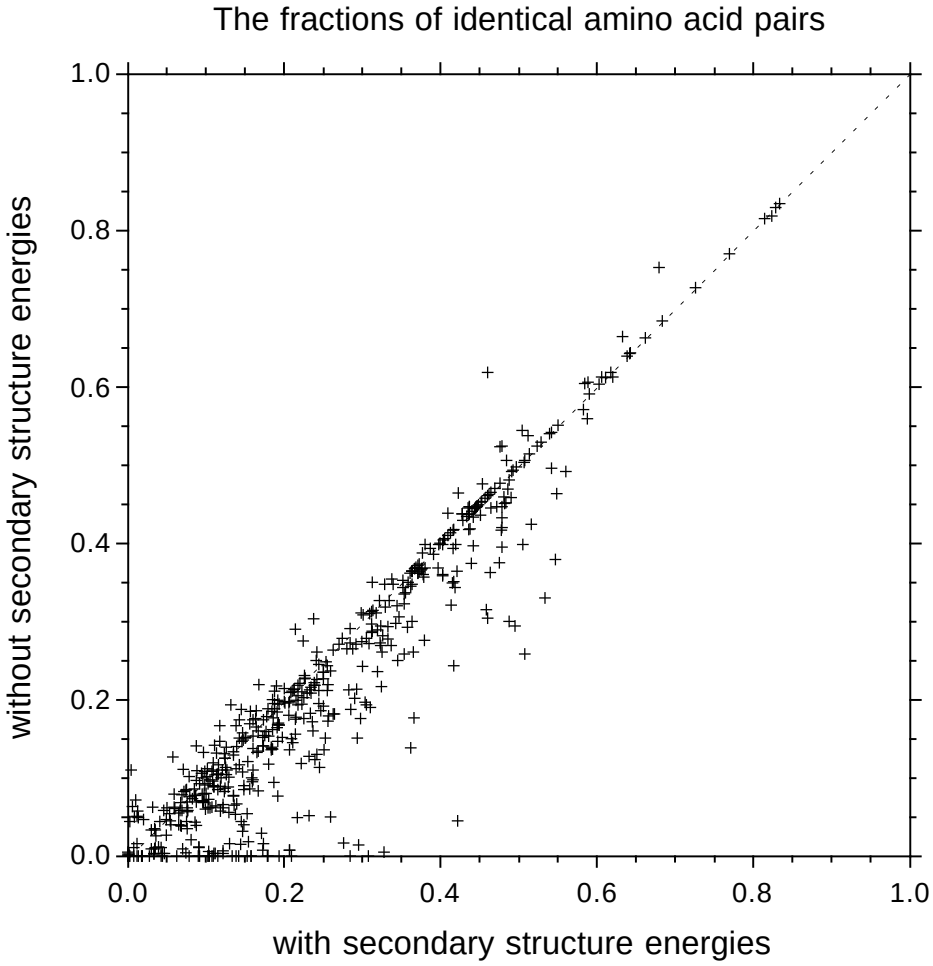

5.0
0.0
-5.0
-10.0
-15.0
-20.0
minimum energy scores in s.d. units
-200.0
-150.0
-100.0
-50.0
0.0
50.0
100.0
minimum energy scores


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


1.0
0.8
0.6
0.4
0.2
0.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
the fraction of identical amino acid pairs
minimum energy scores per residue

The effects of secondary structure potentials

Including secondary structure energies tends to increase identical amino acid pairs in the alignments of homologous protein pairs; secondary structure potentials are useful to yield correct positions of residues in alignments.



Conclusions

- 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.
- Some individual sequence-structure pairs are detected by this method having only 5-20 % sequence identity.

Probability alignment

(Miyazawa, S. (1995) *Prot. Eng.*, **8**, 999-1009)

Algorithm: by iteratively choosing a site pair with the maximum correspondence probability as follows.

1. Set $i_1 = 1$, $i_2 = m$, $j_1 = 1$, and $j_2 = n$.

2. Calculate a site pair (a_i, b_j) such that

   $p(a_i, b_j) = \max_{i_1 \le k \le i_2, j_1 \le l \le j_2} (p(a_k, b_l) \mid p(a_k, b_l) \ge p(a_k, \phi)$ and $p(a_k, b_l) \ge p(\phi, b_l))$.

3. If there is no such a site pair, align ϕ to all sites of $i_1 \le i \le i_2$ and of $j_1 \le j \le j_2$.

4. If (a_i, b_j) is such a site pair, choose it as one of residue-residue correspondences in the alignment. Then, repeat steps $2 - 4$ to align the remaining segments until all the sites are aligned.

A threshold, 0.5, of probability for reliable residue correspondences

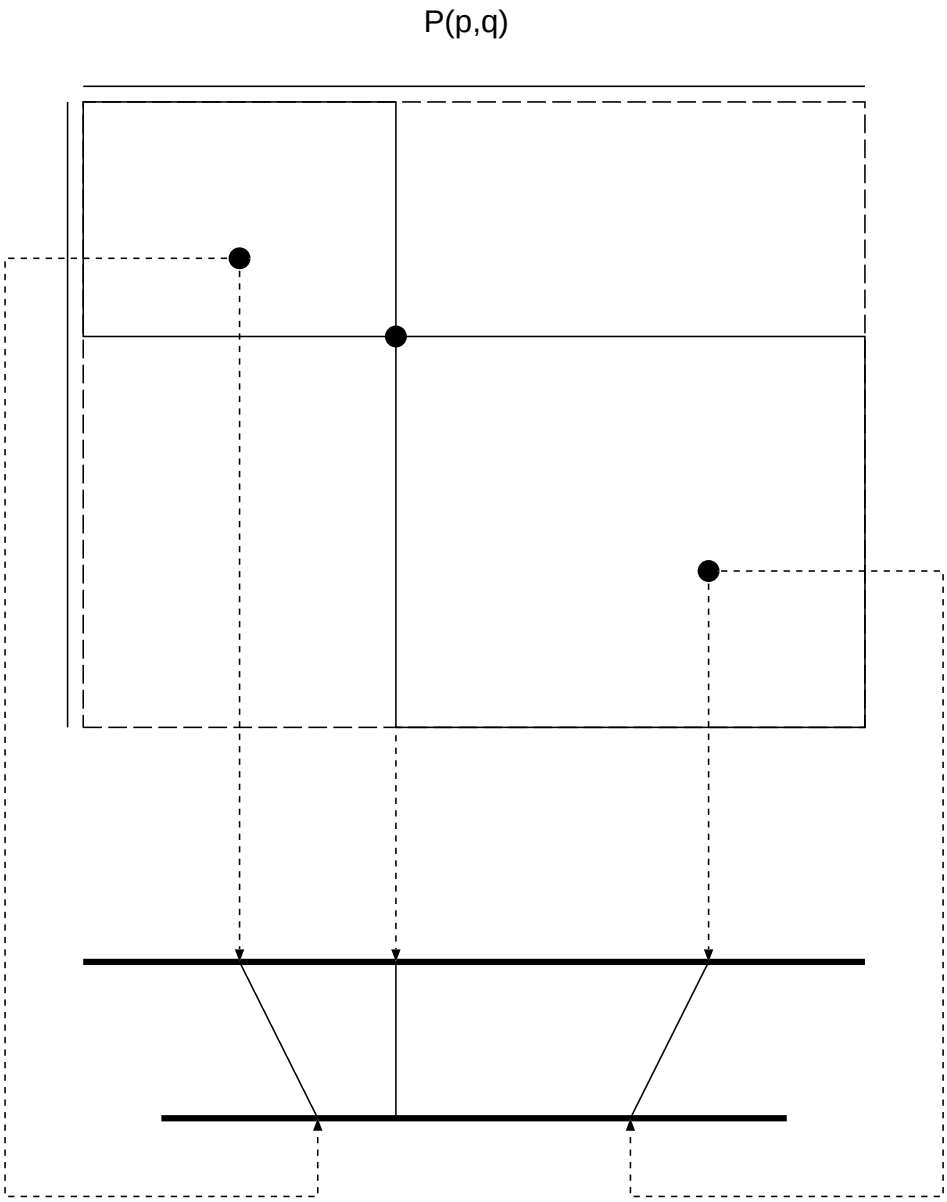
Site pairs with $p(a_i, b_j) > 0.5$, $p(a_i, \phi) > 0.5$, and $p(\phi, b_j) > 0.5$ can constitute an alignment, because

1. The number of b_j and ϕ such that $p(a_i, \{b_j, \phi\}) > 0.5 = 1$ or 0

2. Let $p(a_i, b_j) > 0.5$ and $p(a_k, b_l) > 0.5$. If $i < k$, then $j < l$.

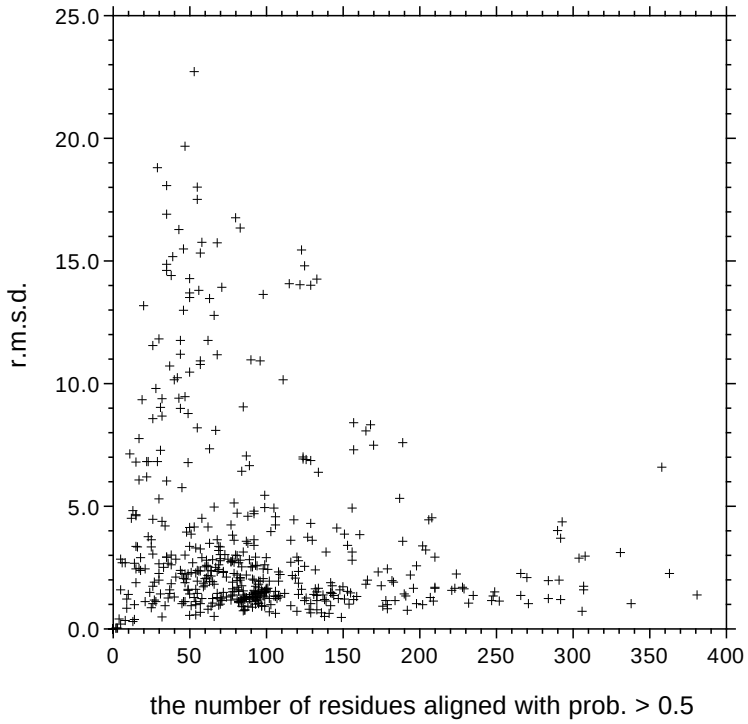
Probability alignment

(Miyazawa, S. (1995) *Prot. Eng.*, **8**, 999-1009)



The effects of the number of aligned residues on r.m.s.d.

in probability alignments



in minimum energy score alignments