Structure-derived substitution matrices for alignment of distantly related sequences

Andreas Prlić, Francisco S.Domingues and Manfred J.Sippl[1]

Center of Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob-Haringerstrasse 3, A-5020 Salzburg, Austria

[1]To whom correspondence should be addressed. E-mail: sippl@came.sbg.ac.at

Sequence alignment is a standard method to infer evolutionary, structural, and functional relationships among sequences. The quality of alignments depends on the substitution matrix used. Here we derive matrices based on superimpositions from protein pairs of similar structure, but of low or no sequence similarity. In a performance test the matrices are compared with 12 other previously published matrices. It is found that the structure-derived matrices are applicable for comparisons of distantly related sequences. We investigate the influence of evolutionary relationships of protein pairs on the alignment accuracy.

Keywords: alignment accuracy/protein evolution/sequence alignment/structure alignment/substitution matrix

Introduction

By the end of 1999, 26 genomes have been fully sequenced and more than 100 additional genome projects are in progress, causing sequence databases to explode (Kyrpides, 1999). For a newly determined sequence, structural, functional and other biologically relevant information can be inferred if evolutionarily related sequences are found (Scharf *et al.*, 1994; Teichmann *et al.*, 1999). Sequence alignment is a standard method to search for such relationships (Pearson and Lipman, 1988; Altschul *et al.*, 1990, 1997).

The success and reliability of sequence comparisons depend on, among other ingredients, the substitution matrix used (Henikoff and Henikoff, 1993; Vogt *et al.*, 1995). Popular matrices such as PAM, BLOSUM and the GONNET matrix are based on sequence alignments (Dayhoff *et al.*, 1978; Gonnet *et al.*, 1992; Henikoff and Henikoff, 1992). The accuracy of these alignments is important for the quality of the matrices. This in particular is the case for large evolutionary distances, where sequence alignments become less reliable.

Structure alignments are generated independently of sequence similarity. They are reliable even in the case of distant evolutionary relationships. Several structure-based matrices have been published (Risler *et al.*, 1988; Johnson and Overington, 1993; Naor *et al.*, 1996; Russell *et al.*, 1997). An important issue in structure comparison is the definition of structural equivalence of residue pairs. Here, structural equivalence is defined by close C^{α} and C^{β} distances, corresponding to residues that occupy similar positions in the structures and resemble each other in their side-chain orientation.

A data set of superimposed protein pairs (Domingues *et al.*, 2000) is used for the derivation of amino acid substitution matrices. Since the goal is to exploit structural information rather than sequence information, the pairs used in this analysis were chosen to have high structural but low sequence similarity. From this dataset a substitution matrix is derived. To investigate the extent to which phylogenetic distances within the data set influence results, a second matrix is calculated by excluding pairs with no or unclear evolutionary relationships. The alignment accuracy of sequence alignments created using both matrices is compared with results from 12 other previously published matrices.

Structure comparison can have multiple solutions (Boutonnet *et al.*, 1995; Feng and Sippl, 1996; Godzik, 1996). It is difficult to judge which of the alternative solutions is the most relevant in terms of evolution. The question arises of which one should be used for the compilation of substitution matrices. Here we use the alignment having the largest number of equivalent residues to derive the matrices. However, in performance tests alignment accuracy is measured by comparing sequence alignments with all alternative solutions.

In the following sections we describe how the amino acid substitution matrices are derived and discuss differences between matrices derived from homologous pairs and matrices derived from homologous as well as analogous pairs. We apply the matrices in a performance test.

Methods

The data set of protein pairs

We need a data set for two purposes: to derive amino acid substitution matrices and to evaluate alignment accuracy. For the first task, sequence similarity among two proteins A and B that are structurally related has to be low. For the second task it is important that there is no sequence similarity of one pair A–B to any other pair C–D in the set. Otherwise, although in the jack-knife test the pair A–B is removed, there could still be a pair C–D related to A–B, resulting in in a statistical bias. A data set that correlates with this criterion was published recently (Domingues *et al.*, 2000). In summary, the procedure to derive the data set is as follows: (i) starting from the PDB (Berman *et al.*, 2000), a set of proteins is prepared, so that any proteins A and B taken from the set have <30% sequence identity; (ii) these proteins are grouped into structurally similar subsets using PROSUP (Feng and Sippl, 1996); and (iii) from these groups representative pairs of superimposed structurally related proteins are selected after visual inspection.

After this procedure, a data set of 122 protein pairs (Table I) is obtained having the following characteristics: (i) only 19 structurally related pairs have significant sequence similarity within the pair that can be detected using SSEARCH (Smith and Waterman, 1981; Pearson, 1991), but ID <30%. In terms of the statistics this means that there is a strong relation in terms of equivalent position and a very low if not negligible correlation in terms of sequence identity; (ii) there is no detectable sequence similarity between proteins that belong to

Table I. Set of protein pairs used in this study. The name of a protein is given by its PDB code (Berman *et al.*, 2000), its chain identifier and the model number

1	193l.–.–	153l.–.–	62	1hrd.A.–	1leh.A.–
2	1aba.–.–	1gp1.A.–	63	1i1b.–.–	4fgf.–.–
3	1acf.–.–	1pne.–.–	64	1idk.–.–	1air.–.–
4	1afi.–.1	1aps.–.1	65	1iow.–.–	1bnc.A.–
5	1agq.D.–	1tgj.–.–	66	1irs.A.–	1mai.–.–
6	1aiz.B.–	1rcy.–.–	67	1kpc.D.–	1hxq.B.–
7	1apm.E.–	1erk.–.–	68	1lea.–.–	1ruo.B.–
8	1apm.E.–	1irk.–.–	69	1lmb.3.–	1pou.–.1
9	1aps.–.1	1spb.P.–	70	1lpe.–.–	1nbb.B.–
10	1ash.–.–	1bin.A.–	71	1lpe.–.–	1vlt.A.–
11	1ash.–.–	1bvd.–.–	72	1lti.D.–	1asz.A.–
12	1ash.–.–	1cpc.A.–	73	1lti.D.–	1bcp.L.–
13	1asz.A.–	1bcp.L.–	74	1lti.D.–	1prt.B.–
14	1bbh.B.–	1nbb.B.–	75	1lti.D.–	1tii.D.–
15	1bbp.D.–	1hbq.–.–	76	1ndh.–.–	1fnb.–.–
16	1bcp.L.–	1prt.B.–	77	1ndh.–.–	2pia.–.–
17	1bdi.A.–	2dri.–.–	78	1oun.A.–	1std.–.–
18	1bdm.B.–	1bhs.–.–	79	1phd.–.–	2hpd.B.–
19	1bdm.B.–	6ldh.–.–	80	1plq.–.–	2pol.A.–
20	1bfm.A.1	1taf.B.–	81	1pot.–.–	1sbp.–.–
21	1bin.A.–	1bvd.–.–	82	1ptv.A.–	1ytn.–.–
22	1bin.A.–	2hbg.–.–	83	1qpa.A.–	2cyp.–.–
23	1btn.–.–	1dyn.B.–	84	1ris.–.–	1spb.P.–
24	1btn.–.–	1irs.A.–	85	1rnl.–.–	1dts.–.–
25	1btn.–.–	1mai.–.–	86	1ryt.–.–	1afr.F.–
26	1bvd.–.–	2hbg.–.–	87	1ryt.–.–	1×ik.A.–
27	1cew.I.–	1mol.A.–	88	1ryt.–.–	1×sm.–.–
28	1cew.I.–	1oun.A.–	89	1sbp.–.–	2abh.–.–
29	1cnv.–.–	1nar.–.–	90	1spb.P.–	1nue.A.–
30	1cpc.A.–	1col.A.–	91	1ste.–.–	1tss.A.–
31	1cpc.A.–	1cpc.B.–	92	1ste.–.–	3ull.A.–
32	1cpc.A.–	2hbg.–.–	93	1taf.A.–	1bfm.A.1
33	1ctj.–.–	1cxc.–.–	94	1taf.A.–	1taf.B.–
34	1ctj.–.–	2mta.C.–	95	1tii.D.–	1asz.A.–
35	1dat.–.–	1afr.F.–	96	1tii.D.–	3ull.A.–
36	1dat.–.–	1ryt.–.–	97	1urn.A.–	1spb.P.–
37	1dat.–.–	1×ik.A.–	98	1vlt.A.–	1nbb.B.–
38	1den.–.1	1tcp.–.1	99	1wba.–.–	1i1b.–.–
39	1dvr.A.–	1dts.–.–	100	1wba.–.–	4fgf.–.–
40	1dyn.B.–	1irs.A.–	101	1wkt.–.1	1amm.–.–
41	1dyn.B.–	1mai.–.–	102	1×ik.A.–	1afr.F.–
42	1eaf.–.–	3cla.–.–	103	1×ik.A.–	1×sm.–.–
43	1ece.A.–	1edg.–.–	104	1×sm.–.–	1afr.F.–
44	1ecm.B.–	1csm.B.–	105	2alp.–.–	1hav.A.–
45	1elg.–.–	1hav.A.–	106	2blt.B.–	3pte.–.–
46	1elg.–.–	2alp.–.–	107	2dri.–.–	1rnl.–.–
47	1erk.–.–	1irk.–.–	108	2hhm.A.–	1spi.D.–
48	1esl.–.–	1lit.–.–	109	2pia.–.–	1fnb.–.–
49	1eur.–.–	2sim.–.–	110	2pii.–.–	1aps.–.1
50	1fb4.H.–	1fna.–.–	111	3nll.–.–	1qrd.B.–
51	1fb4.H.–	1tup.A.–	112	3ull.A.–	1asz.A.–
52	1fec.A.–	1nhq.–.–	113	3ull.A.–	1bcp.L.–
53	1fmb.–.–	1sme.B.–	114	6ldh.–.–	3nll.–.–
54	1fna.–.–	1msp.A.–	115	1tdj.–.–	2tys.B.–
55	1gmf.B.–	1rcb.–.–	116	1tdj.–.–	1psd.A.–
56	1gsa.–.–	1bnc.A.–	117	1ak1.–.–	2dri.–.–
57	1gsa.–.–	1iow.–.–	118	1fui.A.–	1bhs.–.–
58	1gtq.A.–	1gtp.A.–	119	1aoy.–.1	2dtr.–.–
59	1hce.–.–	1i1b.–.–	120	1agj.A.–	1elg.–.–
60	1hce.–.–	4fgf.–.–	121	1ulo.–.–	2ayh.–.–
61	1hfc.–.–	1iag.–.–	122	1bnk.A.–	1fmt.A.–

one pair A–B and to those from another pair C–D; this ensures that there is no statistical bias in the jack-knife test.

Compilation of substitution matrices

When two protein structures are aligned, frequently alternative structure alignments are obtained. The one with the highest number of structurally equivalent residues is used to calculate amino acid substitutions. If the C^{α} and C^{β} atoms of an amino acid pair A_i, B_j are separated by less than 5 Å, they are considered to occupy equivalent positions. The frequency of occurrence f_{ij} of these equivalent positions corresponds to the observed substitution frequency of amino acid pair A_i, B_j.

To calculate the matrix the formalism of Henikoff is used (Henikoff and Henikoff, 1992).

The relative frequency q_{ij} of occurrence of an amino acid pair A_i, B_j is

$$q_{ij} = \frac{f_{ij}}{\sum\limits_{i=1}^{20} \sum\limits_{j=1}^{i} f_{ij}} \tag{1}$$

where f_{ij} is the observed substitution frequency of pair A_i, B_j. The frequency of occurrence of amino acid A_i in a pair A_i, B_j is

$$p_i = q_{ii} + \frac{\sum\limits_{j \neq i} q_{ij}}{2} \tag{2}$$

and the expected frequency e_{ij} for a substitution of a pair A_i, B_j is then

$$\text{if } i = j\text{: } e_{ij} = p_i p_j;$$

$$\text{if } i \neq j\text{: } e_{ij} = p_i p_j + p_j p_i = 2\, p_i p_j$$

Finally, the logarithm of the odds matrix is calculated by

$$s_{ij} = \log_2 \left(\frac{q_{ij}}{e_{ij}} \right) \tag{3}$$

The relative entropy H of a matrix, also called the average mutual information per residue, is calculated according to Altschul (1991):

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij}\, s_{ij} \tag{4}$$

The matrices

The data set of 122 protein pairs yields 13 908 structurally equivalent amino acid pairs. The values s_{ij} are derived from these substitutions according to Equation 3, resulting in the Structure Derived Matrix (SDM, Table II).

CATH classifies structures according to different level of structural similarity and evolutionary relationship (Orengo *et al.*, 1997). Proteins that share the same H-level ('H' for homology) are assumed to be evolutionarily related and share a similar fold. Such proteins are called 'homologous'. Proteins that share the same fold, but for which no evidence for an evolutionary relationship can be found, are classified in different H-levels. These are called 'analogous'.

A second matrix, called the Homologous Structure Derived Matrix (HSDM, Table III), is calculated from the subset of 77 proteins, which are classified as homologous by CATH (9947 amino acid pairs).

Using PHYLIP (Felsenstein, 1985), a phylogenetic inference package, a tree diagram is derived to display the relationships among amino acids.

Performance test

The 122 protein pairs used to derive SDM and HSDM are applied in a jack-knife test, to estimate the accuracy of alignments obtained from these matrices. Hence, for a pair of proteins being aligned, unique SDM and HSDM matrices are generated where substitution counts from this particular pair

Table II. Lower left diagonal, observed amino acid substitutions over all of the data set of structure alignments; upper right diagonal, the structure-derived substitution matrix (SDM) derived from these observed frequencies (values in the matrix are scaled by a factor of 2)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
	2.09	–0.50	–0.57	–0.73	0.33	–0.75	–0.12	0.27	–1.42	–0.97	–0.39	–0.38	–0.04	–0.76	–0.53	0.34	0.13	–0.66	–1.25	0.02	A
		2.87	0.60	0.13	–1.30	0.13	0.99	–0.96	0.54	–1.40	–1.19	1.42	–0.63	–1.40	0.21	–0.06	–0.15	–0.04	–0.75	–1.52	R
A	243		3.60	1.78	–2.08	0.33	–0.16	0.79	0.76	–2.43	–2.10	0.83	–2.01	–2.25	–1.10	0.40	0.30	–2.89	–0.36	–2.17	N
R	103	86		4.02	–2.51	0.34	1.20	–1.20	–0.01	–2.77	–2.65	0.66	–2.58	–2.19	0.72	0.71	–0.75	–1.91	–1.21	–2.02	D
N	81	63	72		6.99	–0.83	–1.97	–2.11	–1.50	0.13	–0.31	–2.19	1.04	1.13	–2.19	0.31	–0.59	–0.76	0.13	0.34	C
D	100	70	100	142		2.60	1.23	–0.12	–0.46	–1.47	–1.49	0.92	–0.13	–2.31	0.24	1.04	0.60	–0.81	–0.61	–1.38	Q
C	44	13	8	9	37		2.97	–0.41	–0.62	–1.81	–2.11	1.11	–1.86	–1.61	–0.26	0.31	–0.21	–2.70	–1.64	–1.84	E
Q	74	52	45	59	12	48		4.36	–0.40	–2.93	–1.98	–0.71	–1.86	–2.67	–0.04	0.29	–0.81	–1.21	–1.62	–1.96	G
E	148	113	61	128	13	96	141		5.89	–1.76	–0.93	0.31	–1.04	–0.22	–1.44	–0.74	–0.52	–1.48	–0.12	–0.35	H
G	191	65	96	63	14	68	99	292		2.76	1.56	–1.81	0.99	0.76	–2.00	–1.75	–0.96	0.25	0.08	1.94	I
H	37	38	33	33	6	21	32	39	60		2.43	–1.96	1.61	1.23	–1.56	–2.30	–0.86	–0.14	0.70	0.81	L
I	123	55	31	36	30	42	60	46	24	163		2.91	–1.62	–2.41	–0.19	–0.06	–0.10	–1.94	–1.72	–1.27	K
L	216	85	50	54	37	60	78	92	46	310	302		3.75	0.80	–1.09	–1.34	–1.58	0.87	–0.41	0.61	M
K	135	131	86	106	12	86	148	89	44	60	82	138		3.28	–0.91	–1.11	–0.69	2.29	1.96	0.51	F
M	66	28	14	15	16	26	23	26	12	69	123	25	35		5.45	–0.29	0.93	–5.34	–1.98	–1.11	P
F	84	35	21	28	27	20	41	32	26	104	176	31	41	79		2.36	1.20	–1.18	–1.56	–1.11	S
P	64	43	22	54	6	34	46	56	12	28	47	47	15	26	83		2.04	–0.57	–0.41	0.05	T
S	164	74	70	102	27	85	106	119	29	58	69	93	26	46	43	102		6.96	2.15	–1.09	W
T	146	69	65	59	19	70	85	78	30	73	109	88	23	51	63	131	84		3.95	0.21	Y
W	31	20	6	11	5	12	10	19	6	31	39	13	15	40	2	16	19	36		2.05	V
Y	63	39	36	35	17	32	36	41	24	73	130	35	24	89	16	35	50	34	79		
V	204	62	40	55	38	51	70	76	46	289	281	85	71	112	45	85	122	23	90	177	

Table III. Lower left diagonal, observed amino acid substitutions over the subset of homologous pairs; upper right diagonal, the substitution matrix derived from homologous structural pairs (HSDM) (values in the matrix are scaled by a factor of 5)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
	5.50	–2.24	–1.77	–2.38	0.45	–2.16	–0.47	0.63	–3.01	–1.72	–1.09	–1.22	0.16	–2.42	–1.11	1.27	0.60	–2.61	–4.22	0.16	A
		8.59	0.24	–0.33	–6.29	–0.74	2.83	–3.39	0.70	–3.93	–2.83	3.89	–1.43	–4.36	1.31	–0.50	0.34	1.02	–1.01	–3.80	R
A	181		10.00	4.07	–6.53	1.42	–0.39	1.16	1.77	–5.78	–5.64	1.64	–4.67	–6.22	–3.23	1.54	1.14	–6.29	–0.93	–5.65	N
R	63	72		11.01	–6.98	1.10	2.41	–3.91	0.32	–6.18	–7.41	1.53	–7.88	–5.06	0.81	2.34	–1.36	–5.63	–3.85	–6.10	D
N	58	39	65		19.05	–2.47	–4.70	–5.70	–5.95	–0.13	–0.82	–6.65	3.50	1.72	–6.70	1.08	–1.89	–3.01	–0.44	1.32	C
D	68	46	73	122		7.85	3.16	–0.24	–2.24	–3.26	–4.56	3.24	–1.76	–5.54	1.30	2.59	1.08	–4.30	–1.73	–4.97	Q
C	30	6	5	6	33		8.43	–1.80	–1.29	–5.89	–5.62	3.08	–3.94	–4.44	–0.43	0.42	–0.61	–6.28	–4.50	–4.23	E
Q	50	31	36	44	8	40		11.64	–1.24	–8.58	–6.55	–1.82	–5.29	–7.46	–1.79	0.63	–2.24	–4.77	–4.34	–5.32	G
E	97	78	43	81	9	64	102		15.72	–4.44	–2.49	–0.17	–3.66	0.25	–3.55	–2.38	–1.14	–5.71	1.17	–1.63	H
G	144	42	68	43	10	51	63	259		6.74	3.86	–4.82	2.94	2.30	–4.04	–4.67	–3.03	–0.26	–0.08	5.23	I
H	27	23	23	24	3	12	21	27	44		6.38	–5.91	4.32	3.90	–2.88	–6.22	–2.40	–0.58	1.81	2.28	L
I	96	36	24	29	20	31	33	29	16	112		8.23	–5.47	–6.19	–1.21	–0.27	–0.37	–5.45	–4.03	–3.57	K
L	150	60	35	35	26	37	49	55	30	215	218		10.21	2.66	–2.02	–3.92	–5.18	4.28	–4.95	1.18	M
K	89	92	58	73	7	66	99	64	25	39	48	103		9.14	–2.96	–5.03	–4.00	6.49	5.38	0.52	F
M	49	20	11	9	13	15	17	18	7	52	90	14	28		13.32	–1.28	2.44	–11.46	–7.41	–2.31	P
F	54	21	14	21	16	14	25	21	19	75	134	20	31	60		6.35	3.09	–4.44	–4.17	–2.69	S
P	52	37	17	38	4	29	35	37	9	25	42	32	13	18	69		6.33	–3.55	–2.92	–0.23	T
S	123	49	56	80	20	59	67	88	18	39	45	62	17	23	31	76		18.08	6.79	–2.13	W
T	110	54	52	47	13	47	57	58	21	48	75	60	14	26	51	95	73		10.92	0.66	Y
W	19	16	5	7	3	6	7	11	3	19	26	8	14	30	2	9	10	27		5.28	V
Y	39	31	27	23	11	22	23	30	20	50	93	25	10	66	9	24	28	29	66		
V	153	45	30	36	30	30	51	56	29	223	212	57	50	72	39	63	87	18	68	138	

are excluded from the statistics. In order to test the performance of different substitution matrices, an implementation of the Needleman and Wunsch algorithm is used to construct the sequence alignments (Needleman and Wunsch, 1970; Gotoh, 1982), where end gaps are not penalized. Alignment accuracy is measured in terms of residue pairs that are placed at equivalent structure positions and are also aligned in the sequence alignment. The accuracy is expressed as a percentage of the length of the structure alignment (Vogt *et al.*, 1995).

Structure alignments are calculated using the program PROSUP (Feng and Sippl, 1996). Each alternative is considered as a possible solution for a structure alignment in the evaluation. The structure alignment (L_x) which yields the highest number of correctly aligned residues, when compared with the sequence alignment, is used to assess alignment accuracy.

The alignment accuracy is estimated in the same way for 12 previously reported matrices (Dayhoff *et al.*, 1978; Risler *et al.*, 1988; Gonnet *et al.*, 1992; Henikoff and Henikoff, 1992; Johnson and Overington, 1993; Naor *et al.*, 1996; Russell *et al.*, 1997). For each matrix the optimum combination of gap opening and gap extension penalties is determined as the one that corresponds to the maximum average alignment accuracy obtained from the pairs. Optimum gap opening penalties are tested for each matrix in the range from 0 and 20 and extension penalties between 0 and 10.

To investigate if there are differences between local and

Table IV. Frequency of occurrence (Equation 2) of amino acids in the substitutions used to calculate SDM compared with the values published by Johnson and Overington (1993)

	SDM	Johnson and Overington
A	9.2	8.4
R	4.8	3.7
N	3.9	4.7
D	5.0	6.0
C	1.5	1.2
Q	3.7	3.6
E	6.0	5.1
G	6.8	8.8
H	2.4	2.2
I	6.7	5.2
L	9.7	7.6
K	6.0	5.9
M	2.6	1.9
F	4.3	3.9
P	3.0	4.5
S	5.7	7.4
T	5.5	6.3
W	1.5	1.6
Y	3.8	3.8
V	7.9	7.8

global alignment algorithms, parameter optimization was done for a few matrices using a local alignment algorithm. The results are similar to those from global alignment, not penalizing end gaps (data not shown).

Results

The substitution matrices

The relative frequencies of the amino acids in the data set are similar to others reported in the literature (Table IV; Johnson and Overington, 1993). The clustering of the amino acids in the dendrogram tree is similar for the structure-derived matrices (Figure 1). Two groups are obtained, one consisting of hydrophobic amino acids and the other containing polar and charged amino acids. The most pronounced differences from the tree obtained for the GONNET matrix are the position of tryptophan and cysteine. In the GONNET matrix these two are grouped separately from the other amino acids, whereas they belong to the hydrophobic branch in the HSDM dendrogram.

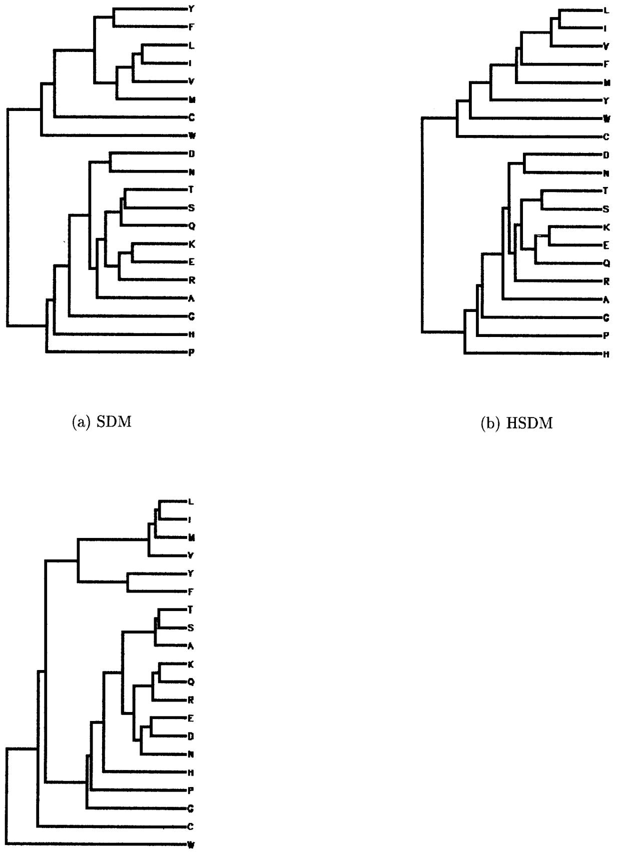
The relative entropy (Equation 4) of the HSDM of 0.28 bit is higher than the relative entropy of the SDM 0.22 bit. The matrix derived from homologous pairs appears to have a higher information content than matrices derived from homologous and analogous pairs.

Matrix performance test

Gap penalties are optimized for several matrices in order to obtain the maximum average alignment accuracy (Table V). The top-ranking matrices are the HSDM and SDM. On average they align ~34 and 33% of the alignment correctly. A similar performance is shown by the GONNET matrix (Gonnet *et al.*, 1992) and the NAOR matrix (Naor *et al.*, 1996), with ~33 and 31% correctly aligned. On average there is no clear gap between the top-ranking matrices, but for single protein pairs there are several cases where one matrix yields a much better performance than the other (Figure 2). The correlation coefficient between GONNET and HSDM is 0.87.

The alignment accuracy results collected here are comparable to those published by Vogt *et al.* (1995). In their study, the top-ranking matrices showed similar averages, with the GONNET matrix performing best.

The results above refer to the parameter set that gives the best alignment performance on average over the whole set of



(a) SDM

(b) HSDM

(c) GONNET

Fig. 1. Tree diagrams for (**a**) Structure-Derived Matrix (SDM), (**b**) HSDM and (**c**) GONNET matrix.

Table V. Results of parameter optimization[a]

Open	Extension	% Correctly	Name
19.0	0.8	33.7	HSDM
6.6	0.6	33.0	SDM
8.5	0.8	32.6	GONNET
14.5	2.0	31.2	NAOR
11.5	1.1	30.7	BLOSUM30
8.5	1.8	30.6	BLOSUM40
3.0	0.2	29.5	RISLER
8.0	1.2	29.1	REMOTEHOMOS
6.0	2.0	28.8	BLOSUM50
10.0	1.0	28.8	PAM250
5.5	0.8	27.8	BLOSUM62
9.5	1.2	25.5	JOHNSON
3.5	5.5	14.5	COMBINED
3.0	3.0	8.4	ANALOGOUS

[a] Open, gap opening penalty; Extension, gap extension penalty; % Correctly, pairs found to be aligned in the structure alignment as well in the sequence alignment in % of the length of the structure alignment; Name, name of matrix. Matrices used: PAM250 (Dayhoff *et al.*, 1978), RISLER (Risler *et al.*, 1988), GONNET (Gonnet *et al.*, 1992), BLOSUM30, BLOSUM40, BLOSUM50, BLOSUM62 (Henikoff and Henikoff, 1992), NAOR (Naor *et al.*, 1996), REMOTEHOMOS, COMBINED, ANALOGOUS (Russell *et al.*, 1997), JOHNSON (Johnson and Overington, 1993).

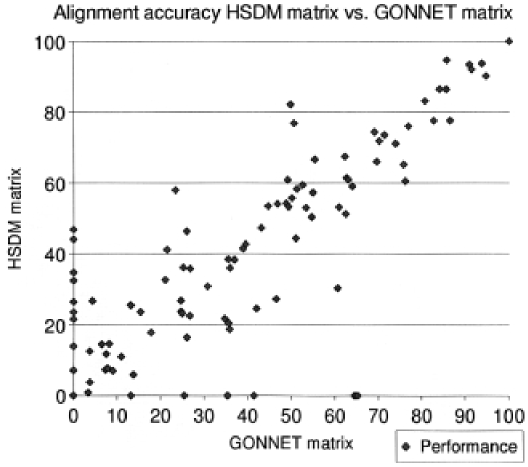


Fig. 2. Alignment accuracy of HSDM and GONNET matrices for different sequence pairs. Alignments are calculated using optimum gap penalties on average over all of the data set. There is some correlation between the two matrices, but in some cases they perform differently.

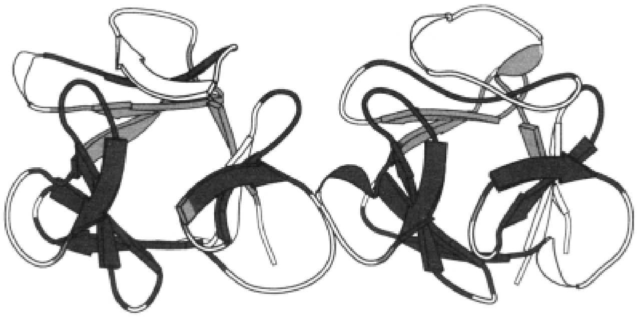


Fig. 3. Structures 1hce and 4fgf. Using the average optimum gap penalties over the whole data set, the GONNET matrix aligns 65% of the alignment correctly, whereas HSDM does not correctly align any residue. Optimizing the alignment accuracy for this pair, HSDM achieves 74% (open penalty 8.0, extension 5.0) and GONNET 72% (open penalty 11.0, extension 0.4). Black residues correspond to structurally equivalent and correctly aligned in sequence alignment, using HSDM and optimum parameters for this pair and dark gray to residues structurally equivalent and not correctly aligned.

protein pairs. Another question is what the alignment accuracy is that can be obtained by optimizing parameters for each pair individually. Here, the average performance increases ~15% for the HSDM, SDM and GONNET matrices. For some protein pairs a large difference in alignment accuracy can be observed.

The structure alignment of the pair hisactophilin and fibroblast growth factor has several small gaps (see Figure 3). The average best-performing gap penalties are too high to allow the opening of the multiple gaps. Using the optimized parameters for this case the opening penalty is much lower and a better alignment is achieved.

Evolutionary relationship of sequences

The pairs of proteins used for testing can be classified either as homologous or as analogous. It is interesting how good analogous pairs can be aligned, as no evidence of an evolutionary relationship can be found for these cases. Here we find that there is a clear difference in the distribution of alignment accuracy between these two groups (Figure 4). In almost half of the analogous structures not a single amino acid pair can be aligned correctly. This only happens in <10% of the homologous proteins. Differences in alignment accuracy between analogous and homologous proteins have also been observed by Russell *et al.* (1998).

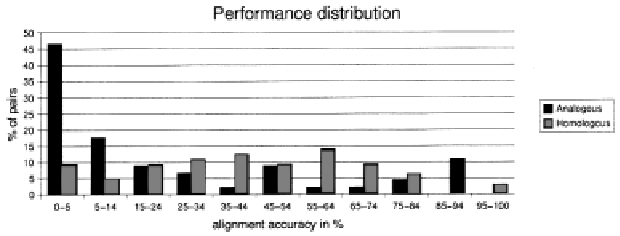


Fig. 4. Alignment accuracy of analogous and homologous structural pairs. Almost 50% of the analogous structural pairs show a performance <5%. Homologous structural pairs show an even distribution over all ranges of performance.

Table VI. Average performance of the top-ranking matrices excluding analogous structures from the test set[a]

Open	Extension	% Correctly	Name
19.0	0.8	43.2	HSDM
7.0	0.6	42.8	SDM
9.0	0.7	41.6	GONNET

[a] For details, see Table V.

These results show that it is more problematic to align analogous proteins. We investigated how the performance changes after analogous pairs are excluded from the test set (Table 6). The average alignment accuracy increases by ~10%. The average optimum gap penalties do not change much. Also, the ranking of the matrices remains the same.

Alternative alignments

Multiple solutions for structure alignments can be found in 88% of the 122 protein structure alignments. For 63% more than four solutions exist. To estimate the similarity of these alternative alignments, they are grouped into different clusters (Lackner, Koppensteiner, Sippl, Domingues, in preparation). Half of the pairs show more than one cluster, one third more than two clusters.

Owing to this finding, it is necessary to consider all the alternative solutions $L_1 \ldots L_n$ when measuring the accuracy of a sequence alignment. The one that shares the highest number of correctly aligned residues with the sequence alignment is L_x. In 20% of the alignments, L_x is different from L_1, the alternative with the highest number of equivalent residues. Using only L_1 in the performance test, the average performance decreases for all of the matrices by 3–4%.

Discussion

The goal of this study was to investigate the possibility of deriving substitution matrices suitable for sequence alignment of distantly related sequences. We derive matrices based on a data set of 122 structure alignments. The superimpositions are produced considering the C^{α} distances as well as the orientation of the side chains. Matrices are derived from structurally equivalent residues of homologous or analogous protein pairs.

Two amino acid substitution matrices are derived. One is calculated from the whole data set of protein structure alignments, SDM. The second, HSDM, is computed after proteins of unclear evolutionary relationship are excluded from the data set. The information content of HSDM is higher, although the total number of observed substitutions is 30% smaller than those of SDM. Also, the average alignment accuracy increases

(Table 5). A similar clustering of substitution values can be observed in both matrices (Figure 1).

These structure-derived substitution matrices, and also previously published matrices, are applied in sequence comparisons of sequences at the border of detectable similarity. The accuracy of the sequence alignments is evaluated by comparison with structure superimpositions. The best average alignment accuracy is observed with the new matrices HSDM and SDM, although the difference from the other top-ranking matrices GONNET and NAOR is small (Table 5).

A closer investigation of the top-ranking matrices shows that the GONNET matrix performs fairly well in the alignment of distantly related sequences. It was based on exhaustive matching of an entire protein sequence database, resulting in 1.7×10^6 sub-sequence matches. In contrast to this large data set, only 77 protein structure comparisons are used to derive HSDM. This demonstrates that structure alignments provide a good source to derive substitution matrices. The good performance of the NAOR matrix was not expected by its authors (Naor *et al.*, 1996). Our data show that it is suitable for sequence alignments of distantly related sequences.

For several pairs the alignment accuracy using GONNET and HSDM is completely different (Figure 2). Also in several cases no correct sequence alignment is obtained when average optimum gap penalties are used. As an example, it is shown how the alignment accuracy could be improved enormously by applying different gap penalties (Figure 3). It has been reported previously that the choice of gap penalty and substitution matrix used considerably affects the sequence comparison results (Henikoff and Henikoff, 1993; Johnson and Overington, 1993; Pearson, 1995; Vogt *et al.*, 1995). The results obtained in this study reiterate one of the basic problems of sequence alignment: if gap penalties are kept constant over all of the sequence, the biologically most meaningful alignment often cannot be found. It would be interesting to investigate the alignment quality of algorithms that do not require gap penalties (Morgenstern *et al.*, 1996) or use a position-dependent gap penalty (Flöckner *et al.*, 1997; Sanchez and Sali, 1997).

Our approach provided us with a high-quality data source to observe amino acid substitutions. The matrices derived here are among the best-performing ones, although they are based on a fairly small number of amino acid replacements. Further improvements could be achieved by collecting data from additional homologous structure alignments. Finding a way to estimate which of the alternative structure alignments is the biologically most meaningful might also improve the quality of the matrices.

The matrices can be downloaded from our web site at http://www.came.sbg.ac.at.

Acknowledgements

We thank Hannes Flöckner and Markus Jaritz for valuable discussions. Thanks are due to Antonina Andreeva, Nora Benhabiles and Peter Lackner for carefully reading the manuscript. This work was supported by grant number P13710-MOB from the Fonds zur Förderung der wissenschaftlichen Forschung (Austria).

References

Altschul,S.F. (1991) *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Boutonnet,N.S., Rooman,M.J., Ochagavia,M.E., Richelle,J. and Wodak,S.J. (1995) *Protein Eng.*, **8**, 647–662.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. National Biomedical Research Foundation, Washington, DC, ed. Dayhoff,M.O., pp. 345–352.

Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) *J. Mol. Biol.*, **297**, 1003–1013.

Felsenstein,J. (1985) *Evolution*, **39**, 783–791.

Feng,Z.K. and Sippl,M.J. (1996) *Folding Des.*, **1**, 123–132.

Flöckner,H., Domingues,F.S. and Sippl,M.J. (1997) *Proteins*, Suppl 1, 129–133.

Godzik,A. (1996) *Protein Eng.*, **5**, 1325–1338.

Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) *Science*, **256**, 1433–1445.

Gotoh,O. (1982) *J. Mol. Biol.*, **162**, 705–708.

Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1993) *J. Mol. Biol.*, **233**, 716–738.

Johnson,M.S. and Overington,J.P. (1993) *J. Mol. Biol.*, **233**, 716–738.

Kyrpides,N.S. (1999) *Bioinformatics*, **15**, 773–774.

Morgenstern,B., Dress,A. and Werner,T. (1996) *Proc. Natl Acad. Sci. USA*, **29**, 12098–12103.

Naor,D., Fischer,D., Jernigan,R.L., Wolfson,H.J. and Nussinov,R. (1996) *J. Mol. Biol.*, **256**, 924–938.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.

Pearson,W.R. (1991) *Genomics*, **11**, 635–650.

Pearson,W.R. (1995) *Protein Sci.*, **4**, 1145–1160.

Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henault,A. (1988) *J. Mol. Biol.*, **204**, 1019–1029.

Russell,R.B., Saqi,M.A.S., Sayle,R.A., Bates,P.A. and Sternberg,M.J.E. (1997) *J. Mol. Biol.*, **269**, 423–439.

Russell,R.B., Saqi,M.A.S., Bates,P.A., Sayle,R.A. and Sternberg,M.J.E. (1998) *Protein Eng.*, **11**, 1–9.

Sanchez,R. and Sali,A. (1997) *Proteins*, Suppl 1, 50–58.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB), **2**, 348–353.

Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.

Teichmann,S.A., Chothia,C. and Gerstein,M. (1999) *Curr. Opin. Struct. Biol.*, **9**, 390–399.

Vogt,G., Etzold,T. and Argos,P. (1995) *J. Mol. Biol.*, **249**, 819–831.

Received February 25, 2000; revised June 20, 2000; accepted June 26, 2000