UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 12

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

Theoretical Approaches to Protein Folding

I. SIMON
Institute of Enzymology
Hungarian Academy of Sciences
Budapest
Hungary

Theoretical Approaches to Protein Folding

István Simon

Institute of Enzymology

Hungarian Academy of Sciences

H-1518 Budapest PO Box 7, Hungary

E-mail: simon@enzim.hu

A hypothesis or theory is clear, decisive and positive, but it is believed by no one but the man who created it. Experimental findings, on the other hand, are messy, inexact things, which are believed by everyone except the man who did the work.

Harlow Shapley

Theoretical Approaches to Protein Folding

1. Statistical approaches

    1.1. Nonrandomness in the primary structure of proteins

    1.2. Heterogeneity of the databases

        1.2.1 " Rapid evolution" of the amino acid composition of proteins

        1.2.2. Interresidue interactions in protein classes

    1.3. Relations among amino acids

    1.4. Developing prediction methods

        a. disulfide bonds

        b. long-range interactions

        c. transmembrane helices

2. The energy calculation approach

    2.1. Description of the 3D structure of protein

    2.2. Forcefield and energy calculation

    2.3. Low energy conformations of oligopeptides (protein segments)

    2.4. Energy calculation on protein, the use of statistically obtained information

4.

Regularities in the primary structure of proteins

M. CSERZÖ and I. SIMON

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Received 21 September 1988, accepted for publication 23 January 1989

In this paper the latest protein database consisting of more than a million amino acids is analyzed to characterize the short range regularities in the primary structure. The amino acid distributions along the polypeptide chain and among the proteins have been studied first. Their influence on the amino acid pair statistics was taken into account. We are primarily interested in the distances of the covalent structure, where the amino acid pair frequencies show non-random characters. The amino acid pairs separated by at least 20 residues in the covalent structure exhibit an exact Gaussian distribution. We found that there is a range of non-random pairing in the covalent structure. We conclude that the pair preference characters are different for each of the 20 × 20 amino acid pairs. The range of the non-random pairing varies from pair to pair, and in most cases it does not extend beyond the 9th neighbour. The preferences of a certain pair in a certain position can not be derived from the character of that pair in another position. The preference values of 400 amino acid pairs are listed for up to the pairs in 9th neighbour position. Some fields of potential application of these data have also been discussed.

Abundance of amino acids in the database and in the subsequent decades of proteins

	summa	%	1	2	3	4	5	6	7	8	9	10
A	82250	7.75	8601	8300	7660	7896	8363	8298	8080	8343	8619	8090
C	22656	2.14	2059	2496	2367	2180	2069	2094	2292	2442	2431	2226
D	54759	5.16	5052	5512	5310	5637	5566	5904	5588	5462	5658	5080
E	64298	6.06	5881	6436	6471	6374	6272	6698	6541	6505	6512	6608
F	42080	3.97	4026	4187	4600	4249	4335	4063	4102	3823	4431	4264
G	77922	7.35	7256	8600	8149	7941	7869	7911	7548	7528	7859	7261
H	25052	2.36	2177	2454	2418	2543	2560	2609	2834	2666	2317	2474
I	53791	5.07	5322	5309	5204	5218	5462	5535	5546	5713	5171	5311
K	63297	5.97	6166	5983	5933	6215	6421	6508	6277	6217	6377	7200
L	96375	9.08	10790	9354	9290	9507	9521	9685	9910	9861	9084	9373
M	24106	2.27	3527	2037	2151	2286	2202	2249	2299	2389	2326	2100
N	44855	4.23	4129	4253	4563	4743	4593	4785	4681	4519	4382	4207
P	56053	5.28	5786	5844	5785	5959	5416	5506	5455	5284	5759	5259
Q	42665	4.02	4386	4317	4319	4421	4177	4235	4129	4016	4380	4285
R	54248	5.11	4976	5216	5614	5424	5165	5360	5567	5437	5540	5949
S	75360	7.10	7938	7681	7613	7449	7201	7578	7392	7318	7297	7893
T	62624	5.90	6368	6277	6849	5968	6201	6120	6497	6305	6163	5876
V	69314	6.53	7008	6910	6778	6774	7001	6983	6824	7372	6535	7129
W	14723	1.39	1628	1355	1515	1456	1552	1463	1240	1325	1631	1558
Y	34397	3.24	2797	3317	3736	3550	3502	3376	3343	3357	3721	3698

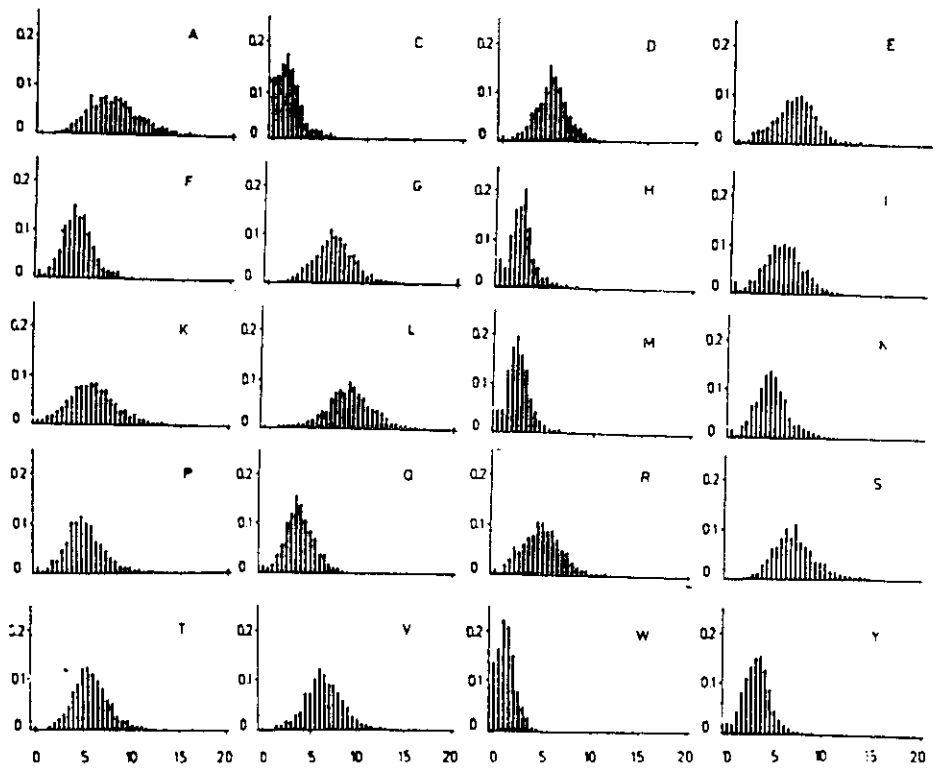


FIGURE 1

Total number of amino acids of those proteins that contain 0%, 0.5%, 1.0% .... 19.5% and 20% or more of a certain amino acid. The unit is 1 million amino acids.

Average abundance of amino acid pairs in the 21-st, 22-nd, ..., 30-th neighbour position.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6171	1310	3710	4345	2682	5368	1683	3490	4107	6460	1521	2961	3715	2874	3741	4785	4145	4733	898	2175
C	1283	818	953	1076	736	1344	448	913	1080	1525	352	771	1028	729	929	1331	1127	1201	269	639
D	3661	979	2689	3226	1971	3440	1218	2461	2893	4251	997	2105	2388	1913	2447	3283	2759	3076	651	1552
E	4346	1062	3126	4019	2155	3906	1302	2840	3543	5083	1209	2314	2700	2335	2994	3813	3190	3613	733	1763
F	2659	738	1976	2192	1667	2587	934	1948	2067	3421	828	1611	1852	1454	1801	2557	2138	2453	545	1294
G	5266	1339	3546	3973	2652	6507	1617	3311	3831	5740	1389	2826	4032	2622	3423	4766	3889	4388	970	2250
H	1678	440	1194	1304	922	1551	664	1081	1259	2129	451	935	1131	853	1075	1515	1269	1378	298	720
I	3505	854	2502	2832	1994	3287	1080	2785	2787	4435	1115	2152	2293	1845	2344	3235	2875	3155	651	1615
K	4119	1048	2890	3525	2199	3786	1292	2802	3856	5020	1166	2353	2648	2074	2665	3577	3174	3413	684	1776
L	6632	1595	4329	5222	3532	5862	2148	4497	4851	8357	1871	3593	4348	3438	4413	5960	5048	5718	1234	2732
M	1637	415	1109	1305	879	1543	515	1194	1271	2014	564	954	1039	885	1086	1476	1263	1391	292	694
N	2830	798	2142	2364	1657	2801	873	2129	2395	3538	884	2011	2012	1615	1915	2802	2443	2587	565	1402
P	3798	1028	2509	2787	1900	4047	1138	2278	2657	4277	932	1980	3531	2027	2502	3519	2958	3089	688	1534
Q	2831	715	1961	2299	1434	2683	874	1881	2132	3353	792	1590	2012	1719	2019	2720	2196	2340	520	1229
R	3669	959	2453	3003	1732	3335	1112	2289	2674	4185	1001	1924	2533	1935	2313	3243	2664	3023	656	1528
S	4780	1418	3493	3940	2604	4739	1562	3329	3647	5909	1362	2870	3559	2705	3324	5134	3963	4275	953	2274
T	4105	1187	2825	3216	2144	3978	1317	2887	3209	4877	1141	2400	3013	2198	2755	3927	3582	3733	777	1877
V	4687	1161	3153	3718	2483	4418	1468	3071	3492	5441	1277	2622	3105	2345	3097	4203	3595	4221	824	1996
W	898	252	637	748	595	907	309	667	732	1202	274	530	717	505	650	921	787	838	241	457
Y	2138	643	1617	1763	1219	2105	687	1579	1720	2603	639	1370	1518	1160	1464	2073	1739	1918	423	1074

Standard deviation of abundance of amino acid pairs in the 21st, 22nd, . . . 30th neighbour position

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	137	67	140	116	147	237	138	90	87	173	55	71	103	76	110	114	95	194	58	108
C	76	52	53	70	34	86	57	64	70	102	28	58	65	32	72	91	53	67	36	27
D	170	50	65	123	165	92	95	73	70	203	51	72	148	109	113	95	100	103	61	47
E	139	51	98	129	140	168	94	92	112	70	34	61	74	58	92	97	65	77	66	84
F	97	80	112	77	76	139	115	88	93	123	47	61	44	108	69	102	82	68	41	90
G	303	51	161	184	167	1222	104	107	96	168	80	97	495	162	135	111	143	94	77	120
H	129	42	104	102	116	111	99	29	73	181	51	55	99	25	55	80	69	123	34	83
I	92	51	130	80	92	129	54	93	95	152	53	73	84	75	76	84	104	78	41	57
K	112	84	126	158	141	167	113	140	168	257	44	73	92	90	116	100	144	136	48	55
L	171	59	127	165	143	212	194	170	191	218	95	112	105	154	141	150	144	161	96	88
M	71	50	76	38	42	80	65	43	50	85	24	61	47	48	62	79	43	75	25	31
N	99	34	61	76	80	83	58	58	106	71	24	91	48	56	51	72	104	76	45	55
P	133	75	120	60	105	547	94	75	160	156	42	87	258	94	114	102	122	138	54	105
Q	89	37	57	91	118	92	77	67	87	137	40	45	64	68	55	115	57	86	43	64
R	93	50	86	76	52	128	50	64	117	99	36	68	75	52	61	84	39	122	46	54
S	93	95	136	112	146	134	89	86	99	114	70	91	92	99	51	129	55	112	57	88
T	118	50	92	97	81	136	69	104	115	125	55	80	54	94	65	110	106	174	55	88
V	181	64	119	98	124	102	137	112	158	232	90	77	119	89	79	127	84	133	63	113
W	59	33	44	49	76	91	42	45	87	84	29	29	96	44	61	71	72	73	53	40
Y	130	34	62	95	63	106	44	73	102	127	42	87	108	66	70	98	54	57	36	43

Abundance of the first neighbour amino acid pairs (dipeptides) normalized by the pair average matrix shown in Table 2

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.13	0.93	0.98	1.05	0.96	1.02	1.02	1.02	0.91	1.08	0.98	0.90	0.96	0.98	0.93	1.02	0.92	1.03	0.96	0.86
C	1.02	0.79	1.05	0.76	1.02	0.97	1.37	0.86	0.95	1.12	1.06	1.01	1.12	1.08	1.08	1.01	0.96	0.90	0.97	1.14
D	1.00	0.90	0.99	0.93	1.05	1.04	0.72	1.17	0.97	1.15	0.97	0.98	1.08	0.88	0.88	0.99	0.89	1.02	1.12	1.06
E	1.12	0.95	1.02	1.19	0.92	0.94	0.88	1.05	1.07	1.02	1.07	1.11	0.80	0.99	1.04	0.80	0.96	0.98	1.06	0.93
F	0.85	0.89	1.09	0.88	0.95	1.05	0.98	1.01	0.99	1.03	0.80	0.98	1.08	1.00	0.97	1.24	1.08	0.93	0.85	1.00
G	0.97	0.91	1.01	1.06	1.05	0.86	1.06	1.06	1.18	0.95	1.05	1.04	0.95	1.00	1.03	1.04	1.01	0.98	0.91	0.97
H	0.90	1.35	0.71	0.84	1.21	1.28	1.09	0.94	0.87	0.99	0.82	1.01	1.30	0.97	0.98	1.01	0.89	0.98	1.31	0.98
I	0.98	1.19	1.00	0.90	0.90	0.93	1.04	0.95	0.97	0.96	0.94	1.10	1.11	1.07	0.98	1.13	1.16	0.93	0.91	0.95
K	1.02	0.96	0.99	1.03	0.86	1.00	1.17	1.01	1.03	0.97	0.89	1.03	0.96	0.93	1.04	0.91	1.02	1.13	0.82	1.08
L	0.98	0.97	1.00	1.00	0.92	0.97	0.99	0.92	1.04	1.02	0.97	1.00	1.01	1.04	1.03	1.12	1.12	0.94	0.80	0.89
M	1.17	0.72	1.08	1.08	1.05	1.01	0.87	0.89	1.13	0.90	1.06	1.04	0.92	0.88	1.04	0.97	1.06	1.01	0.92	0.86
N	0.97	1.10	0.80	0.91	1.04	1.02	1.00	1.08	0.96	1.04	1.05	1.05	1.25	0.93	0.94	0.95	0.94	1.02	1.14	1.07
P	0.98	0.82	0.98	1.24	0.87	1.13	0.97	0.90	0.97	0.91	0.89	0.97	1.00	0.96	0.93	1.10	0.96	1.07	1.17	0.91
Q	1.06	1.09	0.89	1.03	0.92	0.99	1.00	0.99	1.00	1.01	1.06	0.97	0.95	1.21	1.04	0.93	0.96	1.03	1.01	0.94
R	0.97	1.04	1.01	0.94	1.18	1.00	1.09	1.00	0.96	1.06	1.06	1.00	0.95	1.08	1.10	0.89	0.88	1.00	1.05	1.00
S	0.99	1.09	0.92	0.91	1.06	1.20	0.99	0.96	0.93	1.04	0.90	0.91	1.00	0.98	1.01	1.10	0.99	0.95	1.01	0.90
T	1.00	1.12	0.93	0.89	1.09	1.01	0.94	1.11	0.89	1.07	1.02	0.94	1.04	1.01	0.84	0.99	0.99	1.09	1.25	1.00
V	0.97	1.08	1.16	0.98	0.96	0.86	0.91	0.98	1.03	1.09	1.03	0.99	0.99	0.96	0.93	0.99	1.16	0.97	0.97	0.95
W	0.92	1.07	1.02	1.00	0.88	1.15	0.98	0.96	1.07	0.98	1.24	1.12	0.69	1.13	0.88	0.87	1.16	1.04	1.02	1.09
Y	0.84	1.02	1.00	0.82	1.14	0.98	0.98	0.98	0.93	1.02	1.01	1.03	1.04	1.11	1.10	1.06	1.03	0.96	1.13	1.08

Amino acid pair preference matrices (deviation units) from the 1st neighbours up to the 9th ones

1st neighbour matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	6.0	−1.3	−0.4	1.8	−0.7	0.5	0.2	0.7	−4.0	3.2	−0.5	−4.2	−1.3	−0.7	−2.4	0.7	−3.3	0.8	−0.7	−2.8
C	0.3	−3.4	0.8	−3.6	0.4	−0.5	2.9	−2.0	−0.8	1.9	0.8	0.2	1.9	1.8	1.0	0.1	−0.8	−1.8	−0.2	3.3
D	0.0	−2.1	−0.4	−1.9	0.6	1.4	−3.6	5.7	−1.2	3.2	−0.5	−0.5	1.3	−2.2	−2.6	−0.3	−3.0	0.6	1.3	2.1
E	3.7	−1.0	0.6	5.9	−1.3	−1.5	−1.6	1.5	2.2	1.2	2.6	4.0	−7.2	−0.3	1.4	−7.8	−1.7	−1.0	0.6	−1.5
F	−4.1	−1.0	1.6	−3.4	−1.1	1.0	−0.1	0.2	−0.1	0.9	−3.6	−0.4	3.5	−0.1	−0.8	6.1	2.0	−2.6	−2.0	0.0
G	−0.5	−2.2	0.1	1.4	0.7	−0.7	0.9	1.9	7.1	−1.6	0.8	1.2	−0.4	0.0	0.8	1.7	0.2	−1.1	−1.1	−0.5
H	−1.3	3.6	−3.3	−2.0	1.6	3.9	0.6	−2.1	−2.2	−0.1	−1.6	0.1	3.4	−0.9	−0.3	0.3	−2.1	−0.3	2.7	−0.2
I	−0.8	3.1	0.0	−3.6	−2.2	−1.8	0.8	−1.6	−0.8	−1.1	−1.4	2.9	3.0	1.8	−0.6	5.2	4.5	−3.0	−1.4	−1.5
K	0.8	−0.5	−0.2	0.6	−2.2	0.8	1.9	0.3	0.7	−0.5	−2.8	1.1	−1.2	−1.7	1.0	−3.4	0.5	3.2	−2.5	2.5
L	−0.8	−0.8	0.0	0.1	−2.0	−0.8	−0.1	−2.2	0.9	0.8	−0.5	0.0	0.5	0.9	0.9	4.9	4.1	−2.1	−2.5	−3.5
M	3.9	−2.3	1.2	2.7	1.0	0.2	−1.0	−3.2	3.3	−2.4	1.4	0.6	−1.7	−2.3	0.7	−0.6	1.7	0.1	1.0	3.1
N	−0.9	2.4	−6.9	−2.8	0.8	0.6	0.0	3.1	−1.0	1.9	2.0	1.2	10.3	−2.1	−2.4	−1.9	−1.4	0.6	1.7	1.7
P	−0.6	−2.4	−0.5	10.9	−2.4	1.1	−0.3	−3.0	−0.4	−2.4	−2.5	−0.6	−0.1	−0.8	−1.6	3.5	−0.9	1.5	2.2	−1.3
Q	1.9	1.7	−3.7	0.7	−1.0	−0.3	0.0	−0.2	−0.1	0.2	1.1	−1.0	−1.5	5.4	1.3	−1.6	−1.5	0.7	0.1	−1.1
R	−1.2	0.7	0.2	−2.5	6.0	−0.1	2.0	−0.1	−0.9	2.7	1.7	−0.1	−1.7	2.9	4.8	−4.3	−8.1	0.0	0.7	0.0
S	−0.7	1.4	−2.0	−3.2	1.0	7.1	−0.1	−1.5	−2.4	1.9	−1.9	−2.7	0.1	−0.5	0.4	4.2	−1.1	−1.9	0.1	−2.5
T	−0.1	2.8	−2.2	−3.6	2.4	0.2	−1.1	3.0	−3.2	2.6	0.4	−2.0	2.0	0.2	−6.8	−0.4	−0.2	1.9	3.5	0.1
V	−0.9	1.5	4.3	−0.8	−0.9	−6.1	−1.0	−0.5	0.8	2.1	0.4	−0.3	−0.4	−1.0	−2.7	−0.3	6.8	−0.9	−0.4	−0.9
W	−1.2	0.5	0.3	0.0	−1.0	1.5	−0.1	−0.6	0.6	−0.3	2.3	2.1	−2.3	1.5	−1.3	−1.6	1.7	0.5	0.1	1.0
Y	−2.6	0.4	−0.1	−3.3	2.8	−0.4	−0.2	−0.4	−1.2	0.3	0.1	0.5	0.6	1.8	2.1	1.3	0.9	−1.3	1.6	1.9

2nd neighbour matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4.2	0.9	−3.6	−1.3	0.2	2.5	−0.4	1.1	1.7	1.0	0.1	−2.4	−2.4	−2.4	−0.7	−0.5	−1.4	−0.3	−0.8	0.0
C	3.0	−4.6	1.1	0.0	−2.9	0.8	−0.2	−1.3	0.5	0.9	−2.1	0.5	0.4	2.7	−0.8	−0.4	−0.1	1.5	−2.1	−2.6
D	−1.1	−1.8	−2.0	0.0	1.0	−3.7	−2.2	5.2	−2.9	2.2	1.2	−1.2	−1.6	−1.2	−2.7	1.7	−0.1	5.4	0.9	2.3
E	−1.6	−0.3	−3.5	0.7	2.7	0.5	−0.6	4.5	−2.0	13.0	5.8	−4.1	−3.8	1.8	−1.4	−6.4	−3.4	2.5	0.6	−0.2
F	−2.0	−0.9	1.2	1.5	−1.9	2.2	0.6	−1.5	0.6	−2.3	−1.6	3.5	3.6	0.8	1.2	2.1	−0.3	−3.4	−2.7	−1.7
G	−0.5	1.6	−1.3	−2.0	1.2	−0.6	0.5	1.9	−0.2	0.1	−0.2	−1.7	1.5	−1.8	−0.7	1.7	2.2	2.4	0.6	−0.4
H	−1.3	−0.5	3.1	−0.4	−0.1	0.4	0.0	1.8	0.2	−0.6	−0.9	1.0	−1.0	0.9	−1.1	−1.5	−0.1	0.8	1.6	0.2
I	0.4	1.3	1.6	5.1	−3.4	2.4	2.0	−6.2	−0.1	−2.8	−1.9	5.3	−0.5	1.2	1.7	3.1	−0.1	−4.4	−1.2	−2.6
K	−2.1	0.8	−3.6	−2.5	−0.1	0.3	2.6	3.4	0.5	2.1	−0.9	−1.5	−2.4	−1.7	−1.0	−1.9	−0.1	1.9	1.9	2.2
L	0.8	4.8	5.3	4.3	−3.7	1.8	−0.6	−3.5	1.3	−2.3	−3.1	2.3	−0.4	1.6	2.7	1.3	−2.9	−2.6	−1.4	−5.2
M	−0.1	−0.4	−0.3	1.0	−0.1	−2.2	−0.9	−1.8	2.8	−1.7	1.8	2.3	−1.0	1.8	1.9	0.5	0.6	−1.0	1.7	−0.4
N	−2.1	−1.5	0.1	−1.6	−1.0	−4.9	−0.6	1.9	3.6	1.2	2.4	0.7	−3.0	−0.4	1.6	−1.5	3.4	0.1	0.4	0.6
P	0.2	0.7	−0.6	−3.2	0.0	0.3	−0.4	0.0	−0.3	−0.2	0.4	0.8	0.9	−1.1	−3.2	0.0	2.8	0.4	−1.7	0.0
Q	−0.9	−1.0	−6.7	−0.9	1.5	−2.3	0.7	3.2	0.0	3.3	3.3	−3.5	3.8	1.8	−0.3	−2.2	−4.1	2.4	0.3	−2.2
R	−2.6	−1.3	−3.5	−5.8	9.8	−0.6	−0.2	2.0	−3.0	8.0	0.2	−2.0	−1.6	0.3	4.0	−0.9	−1.8	0.6	0.9	1.5
S	−0.4	−0.6	−1.3	−2.7	0.7	2.1	−2.1	0.2	−3.6	4.1	−0.6	−2.7	−0.3	2.3	−2.1	5.8	6.0	−0.8	−0.1	−0.7
T	2.3	−1.3	−0.2	1.9	0.1	−2.9	−1.7	−2.0	−1.1	0.4	−1.6	−0.3	0.8	0.8	−1.5	3.3	5.0	−1.8	−0.1	−0.5
V	2.8	−0.4	1.6	0.2	−2.2	3.2	1.1	−1.6	−1.3	−1.3	−0.6	−1.1	1.2	0.2	−3.0	1.8	−1.0	−0.8	2.3	−1.2
W	−0.6	0.2	0.5	−0.4	−1.5	−0.4	0.8	−2.0	2.4	−0.4	−0.8	2.4	−0.3	4.4	2.4	0.0	−1.5	−1.7	−0.5	−1.4
Y	0.4	3.4	2.7	−0.7	−3.5	0.8	0.2	−1.5	−0.7	−1.1	0.5	1.4	1.6	2.2	−1.3	0.1	−2.1	−2.6	1.0	0.1

3rd neighbour matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4.2	−0.3	−0.7	−0.4	−1.1	−1.4	1.4	0.1	4.7	0.1	0.1	−0.8	0.3	−0.9	1.5	0.8	−2.3	−1.3	−1.4	−1.2
C	−0.1	2.2	1.9	−0.7	−1.7	1.9	0.6	−1.0	−1.8	0.1	−0.4	1.0	0.4	−4.1	−1.5	1.4	−0.3	−0.1	0.1	−1.3
D	0.1	−1.9	−1.6	−2.0	0.0	−2.2	−0.6	2.0	8.9	−1.1	−0.1	2.2	−0.3	0.1	0.7	−2.2	1.6	0.0	−1.4	2.0
E	1.6	2.1	−1.2	3.2	−0.6	−0.5	−1.7	−0.3	6.0	−4.3	−3.3	1.0	−1.2	0.3	7.8	−3.9	−4.3	−3.4	−1.5	−2.2
F	−1.8	−1.4	1.1	−0.4	4.9	−1.4	−0.3	−0.4	−0.9	2.9	0.8	1.4	−1.7	−1.9	−1.0	3.5	1.4	−4.3	−0.4	−1.3
G	−0.9	0.6	−3.8	−2.8	−0.2	1.4	−1.7	0.1	−5.0	1.2	0.5	−3.2	−0.7	−1.5	−2.8	−0.1	2.1	9.9	0.8	1.5
H	−1.5	−0.5	0.5	0.4	−0.7	−0.1	0.6	1.8	2.9	−1.0	−0.7	0.5	1.4	−1.9	1.5	−1.0	0.9	−1.0	0.3	0.4
I	0.8	−0.9	−0.3	−2.5	−0.9	1.0	1.5	0.0	0.2	2.6	0.2	1.2	−1.1	−2.0	−0.8	−1.3	0.9	−1.2	0.9	−1.1
K	−1.1	0.4	0.3	2.4	−0.6	1.1	−0.7	2.4	−0.5	−1.1	−3.3	−1.1	0.7	1.9	−1.9	−0.6	−1.1	2.3	0.2	−4.4
L	0.9	0.2	1.2	−0.2	−0.6	0.3	0.0	−1.9	−1.1	5.8	1.6	−1.0	−0.7	−0.4	−3.0	0.1	−1.4	−2.7	−0.6	2.0
M	2.1	−0.4	−1.7	−3.3	0.7	−0.4	−0.7	0.4	−4.7	3.3	2.9	−0.6	−0.1	−0.9	1.2	−0.1	0.4	0.0	0.5	1.0
N	−1.9	−0.6	−3.3	−1.3	0.8	−0.8	−1.6	3.2	1.0	3.6	−2.7	0.5	−2.4	1.5	−2.0	−0.2	0.6	2.3	2.2	−1.7
P	−1.5	−0.2	−1.2	−1.6	0.9	−1.6	−0.5	2.3	0.9	0.4	−1.3	−2.3	2.6	1.7	−0.5	0.9	1.1	1.7	0.1	−0.7
Q	−0.1	2.2	−2.1	0.8	0.8	0.4	−0.3	−1.6	−0.8	2.3	−1.7	2.5	−1.2	3.9	−0.6	−0.8	−1.2	−1.1	−0.4	−2.9
R	−0.6	−0.8	−1.1	−1.8	3.5	1.1	−1.8	−0.6	−1.1	2.8	−1.1	−1.2	0.3	3.3	4.0	−2.3	−4.5	−0.2	1.3	0.3
S	0.5	−1.5	1.2	0.9	−0.8	−0.9	0.9	0.2	−3.6	−1.2	−1.5	−1.0	1.2	0.9	−4.3	5.9	4.9	−0.1	0.1	−1.8
T	−1.8	2.7	−0.7	2.2	−1.3	−0.2	−1.4	−0.4	−1.3	−2.8	0.1	1.9	2.9	2.0	−1.9	2.2	1.9	−0.2	0.1	−0.9
V	3.3	0.5	1.4	−1.4	−1.2	3.5	1.5	0.3	−2.6	0.2	−0.1	0.0	0.4	−0.9	−3.6	1.0	−1.7	−1.4	−0.1	−1.5
W	−1.5	−0.2	1.2	−0.4	−0.9	0.3	−0.3	−1.3	−1.7	−1.1	0.4	3.4	0.0	4.2	1.4	−1.5	−0.6	2.8	0.3	−0.8
Y	−1.5	−2.0	1.5	0.2	0.8	−2.3	2.6	−3.3	−0.5	0.3	1.7	0.2	0.1	−0.2	0.5	1.6	4.1	−3.2	1.9	2.0

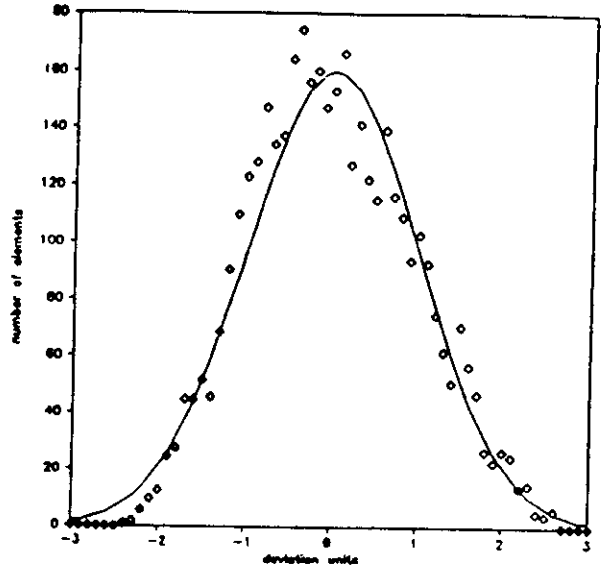


FIGURE 2

Abundance of the various pair preference matrix elements in the 21st, . . . 30th neighbour region. The solid line is the theoretical Gaussian distribution.

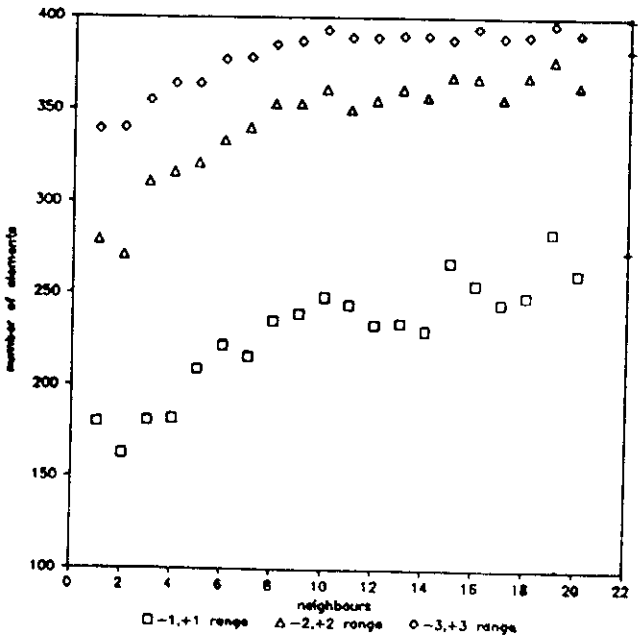


FIGURE 3

Number of pair preference matrix elements between $\{-1, +1\}$, $\{-2, +2\}$ and $\{-3, +3\}$ deviation units versus the sequential distance. Theoretical values of random pair occurrence within the $\{-1, +1\}$, $\{-2, +2\}$ and $\{-3, +3\}$ deviation unit range are marked on the right ordinate.

10.

'Rapid evolution' of the amino acid composition of proteins

While preparing a chapter entitled 'Frequency of Amino Acids' for the book *Proteins* (Landolt-Börnstein New Series, Vol. VII/2), we have noticed a rapid evolution in the amino acid compositions of the proteins which have been sequenced in the past decade.

Table I shows the overall composition of proteins derived from databases which were available in 1978[1], 1984[2], and 1986[3], as well as the amino acid composition of the so-called open-reading-frame (ORF) proteins of the 1986 database*.

Protein sequences that are derived from nucleic acid sequences have caused most of the systematic changes in composition shown in Table I. The applicability of this method does not depend on solubility or other features of the protein sequenced and thus new classes of proteins have been added to the data set.

A trivial change is the increasing amount of methionine coded by the start codon; N-terminal methionine belonging to a signal peptide is missing in an isolated protein. The ratio of hydrophobic and hydrophilic residues has increased, apparently because membrane-bound proteins, which cannot be sequenced as proteins by traditional methods, have been added to the database. There are also significant shifts in the relative abundances of some similar residues. For example the Arg:Lys ratio rises from 0.60 (data of 1978) to 0.86 (data of 1986), and it is 1.17 for ORF proteins which is closer to but still far from the ratio of the number of different codons coding for Arg and Lys[4,5].

A significant consequence of the fact that proteins appearing in the database do not represent all proteins uniformly is that structure-prediction methods based on statistical analysis of protein, such as the widely used Chou-Fasman method for predicting secondary structures[6], or our method for predicting domain boundaries

Table I. Amino acid composition (%) of proteins[a]

Amino acid	I	II	III	IV
A	8.3	8.11	7.75	7.28
C	3.1	2.28	2.14	1.77
D	5.5	5.13	5.16	4.07
E	5.7	5.97	6.06	4.87
F	3.8	3.83	3.97	4.63
G	8.9	7.57	7.35	6.18
H	2.4	2.38	2.36	2.34
I	4.1	4.98	5.07	6.14
K	7.0	6.25	5.97	4.85
L	7.5	8.76	9.08	10.93
M	1.6	2.25	2.27	2.99
N	4.0	4.28	4.23	4.49
P	5.5	5.10	5.28	5.76
Q	3.8	3.94	4.02	3.42
R	4.2	4.94	5.11	5.67
S	7.3	7.12	7.10	7.70
T	6.1	6.02	5.90	6.09
V	6.5	6.46	6.53	5.79
W	1.2	1.40	1.39	1.58
Y	3.4	3.29	3.74	3.45

[a]Derived from databases which were available in 1978 (I), in January 1984 (II), and in December 1986 (III) and from ORF proteins appearing in the 1986 database (IV) (see text).

*Baker, W. C., Hunt, L. C., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Seibel-Ross, E. I., Elzanowski, A., Hong, M. K., Ferrick, D. A., Bair, J. K., Chen, S. L. and Ledley, R. S. (1986) Protein Sequence Database, Release 11.0, December 1986 plus NEW.DAT file, National Biochemical Foundation. Georgetown University Medical Center, Washington DC, USA.

 0376-5067/90/$02.00

11.

of multidomain proteins[7], or even the very recently suggested prediction methods based on structural motifs[8,9], should be revised from time to time, or different data should be used for various sets of proteins. Unfortunately, classifications like water-soluble, membrane-bound, etc., do not necessarily lead to homogeneous groups from a structural point of view.

It is evident that some protein families are over-represented in the database because large numbers of phylogenetically related, homologous proteins have been sequenced. Unfortunately there are certain disadvantages of the selection of the database; some of these are discussed in Ref. 1. The main difficulty, however, arises from the under-representation of protein sets about which we will only learn after new sequencing methods are developed. It is clear that the small set of proteins for which three-dimensional structures are known represents the naturally occurring proteins even more poorly than does the larger set of the sequenced proteins. Therefore one should be very cautious when estimating the size of a data set sufficient for reliable structure prediction[9,10].

References

1 Vondervisztt, F., Matrai, Gy. and Simon, I. (1986) *Int. J. Peptide Protein Res.* 27, 483–492

2 Saroff, H. A. (1984) *Bull. Math. Biol.* 46, 661–672

3 Cserzö, M. and Simon, I. (1989) *Int. J. Peptide Protein Res.* 34, 184–195

4 King, J. L. and Jukes, T. H. (1969) *Science* 164, 788–798

5 Jukes, T. H., Holmquist, R. and Moise, H. (1975) *Science* 189, 50–51

6 Chou, P. Y. and Fasman, G. D. (1974) *Biochemistry* 13, 222–244

7 Vondervisztt, F. and Simon, I. (1986) *Biochem. Biophys. Res. Commun.* 139, 11–17

8 Unger, R., Harel, D., Wherland, S. and Sussman, J. L. (1989) *Proteins: Structure, Function and Genetics* 5, 355–373

9 Rooman, M. and Wodak, S. J. (1988) *Nature* 335, 45–49

10 Thornton, J. M. and Gardner, S. P. (1989) *Trends Biochem. Sci.* 14, 300–304

I. SIMON AND M. CSERZÖ

Institute of Enzymology, Hungarian Academy of Sciences, H-1502, PO Box 7, Budapest, Hungary.

12.

Interresidue Interactions in Protein Classes

Gugolya, Z., Dosztányi, Zs. and Simon, I.*

The free energy difference between folded and unfolded state is about the same for most protein and it is not more than the energy of a few noncovalent interactions. In addition to the numerous noncovalent interactions some protein contain one or more disulfide bonds which as covalent crosslinks significantly stabilize their tertiary structure. Correlation between the presence of disulfide bond(s), and the number noncovalent interresidue interactions of various kinds is analyzed here. The number of interaction per residue is almost the same for all protein. Also the number of long-range interactions per residue is the same in all protein. Proteins with S-S bond(s) (extracellular proteins) have more medium-range and less short-range interactions than those without S-S bonds. However the difference is independent of the number of these covalent crosslinks. We concluded that the different distributions of the various kinds of non-covalent interaction reflect the needs of proteins in the different environments, the extracellular and the intracellular ones, rather than the presence of the disulfide bond(s). We also pointed out that the observed differences in the distributions of short-, and medium-range interactions are in good agreement with different secondary structure compositions of extracellular and intracellular proteins.

TABLE 1 List of proteins*

PDB name	number of SS bonds	protein class†	number of residues	number of interactions			
				total	short	medium	long
155c	0	IN	121	466	339	71	56
1acx	2	EX	108	404	210	101	93
1alc	4	EX	122	464	345	86	33
1bbpA	2	EX	173	670	381	176	113
1cc5	1	IN	83	338	263	40	35
1eca	0	EX	136	517	469	15	33
1fkf	0	IN	107	432	246	69	117
1fnr	0	IN	296	1179	756	138	285
1gp1A	0	IN	184	701	470	83	148
1hdsB	0	IN	145	570	492	30	48
1hip	0	IN	85	303	205	45	53
1hoe	2	EX	74	293	153	62	78
1lrd4	0	IN	92	342	299	27	16
1paz	0	PP	120	476	280	83	113
1pcy	0	IN	99	382	207	69	106
1phh	0	IN	394	1577	1063	213	301
1rbp	3	EX	175	677	383	204	90
1rhd	0	IN	293	1128	769	109	250
1rnh	0	IN	148	593	383	83	127
1sn3	4	EX	65	268	155	65	48
1tpkA	3	EX	88	308	194	59	55
1wsyB	0	IN	385	1630	1099	223	308
256bA	0	PP	106	422	369	27	26
2alp	3	EX	198	864	423	218	223
2azaA	1	PP	129	515	290	72	153

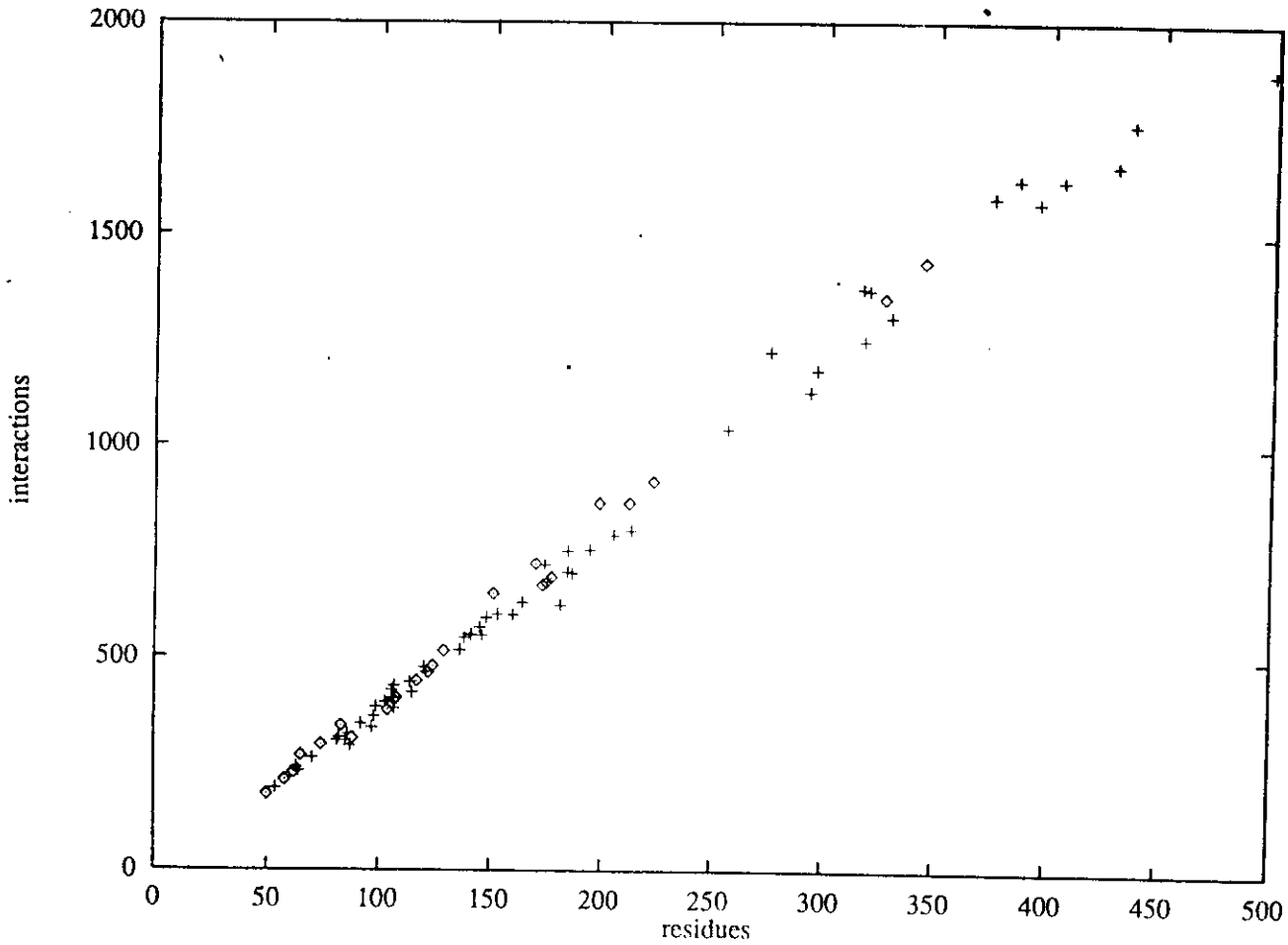


FIGURE 1 The total number of interactions as a function of the number of residues for proteins with disulfide bonds (◊), and without disulfide bonds (+).

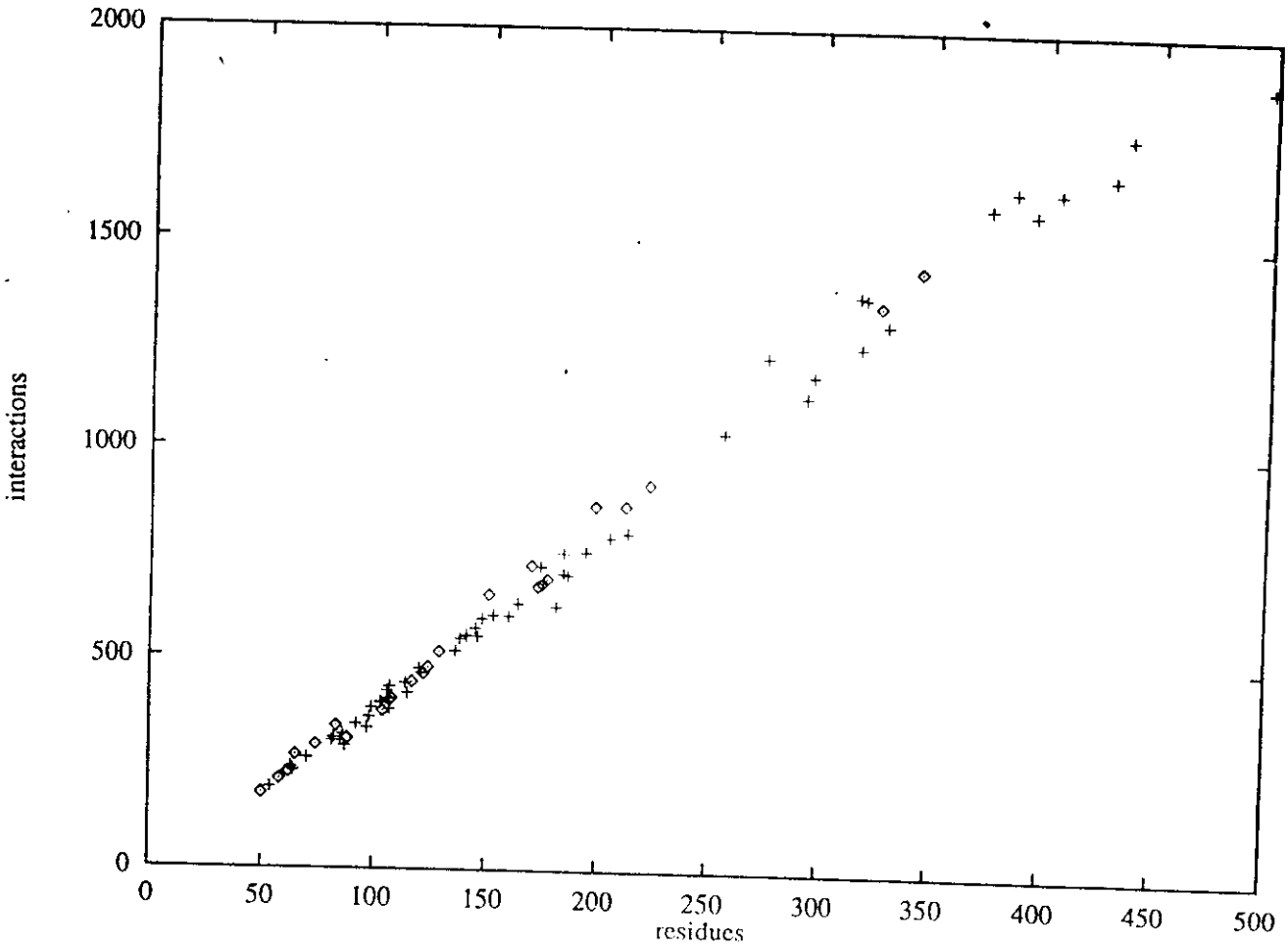


FIGURE 1 The total number of interactions as a function of the number of residues for proteins with disulfide bonds (◊), and without disulfide bonds (+).

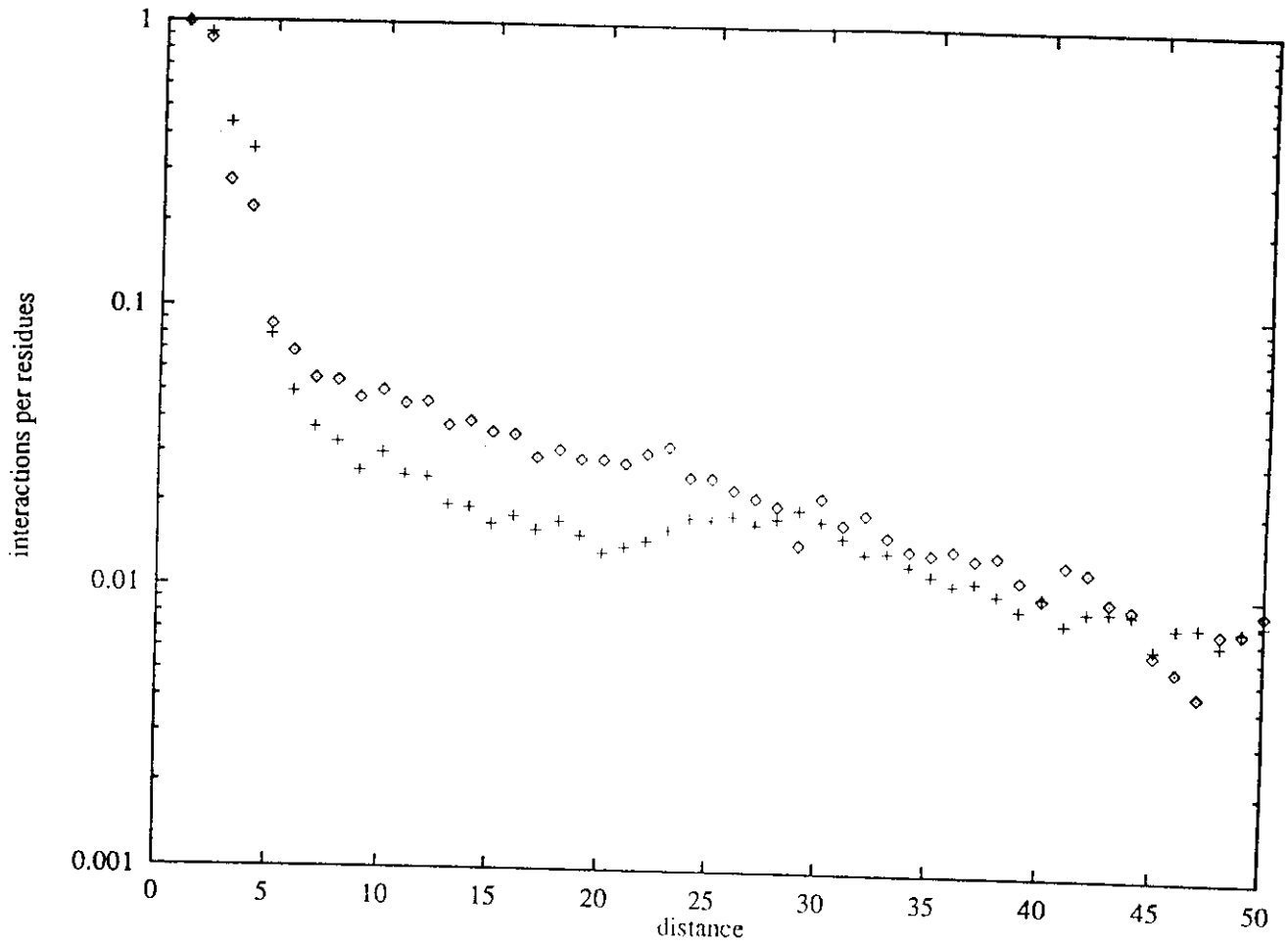


FIGURE 2 The average of the number of interactions per residues as a function of sequential distance, on semi-logarithmic scale, for proteins with (◊) and without (+) disulfide bonds.

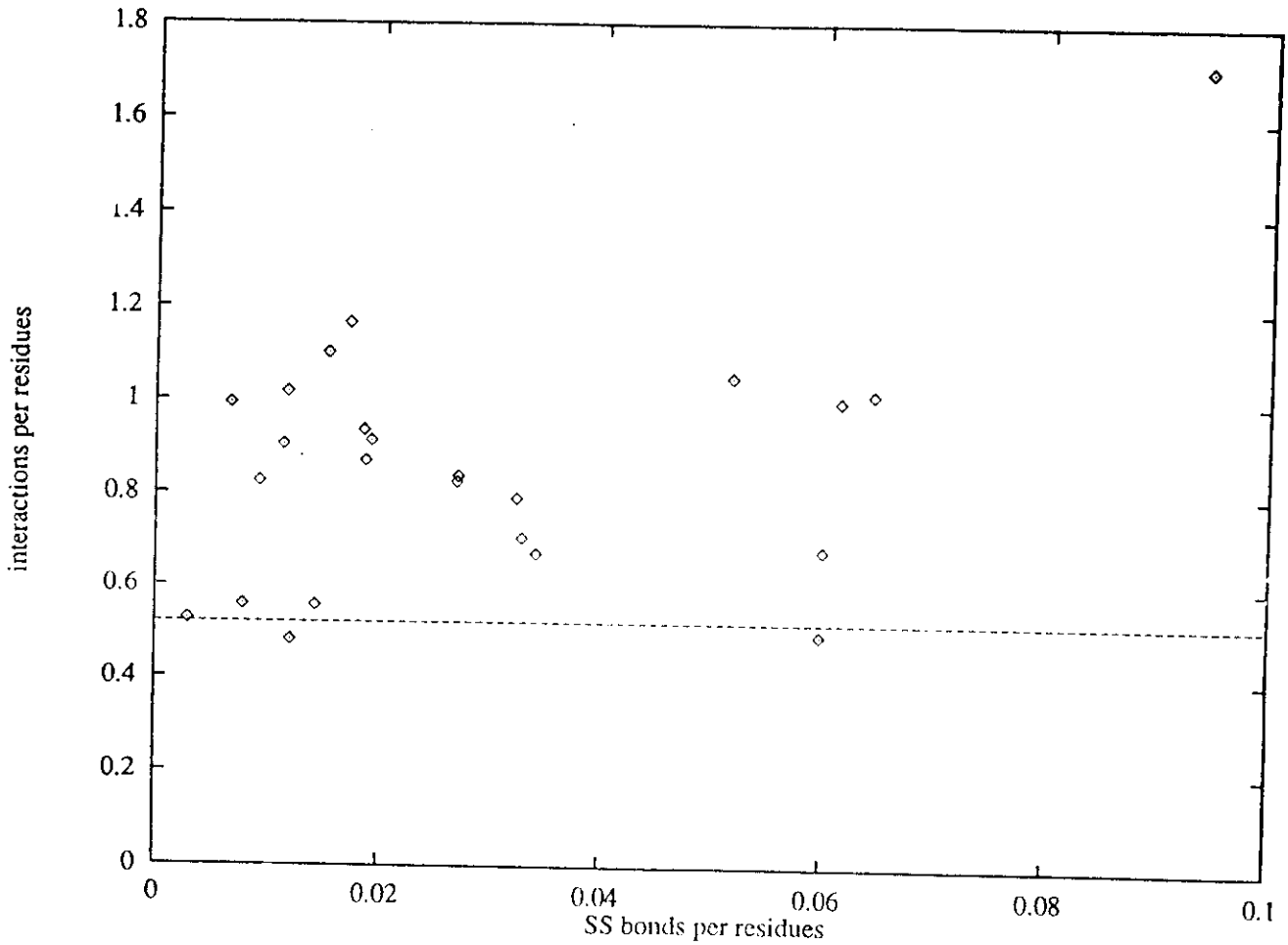


FIGURE 3 The number of interactions per residues for proteins with disulfide bonds as a function of the number of SS bonds per residues. The horizontal line marks the average number of interactions per residue for disulfide free proteins.

TABLE 2 The average number of different kinds of interactions.*

	proteins with SS bonds		disulfide free	
	intact	"modified"§	proteins	Δ‡
total	3.920	3.831	3.867	**0.036**
short	2.353	2.371	2.689	**0.318**
medium	0.860	0.778	0.520	**-0.258**
long	0.70.	0.683	0.658	**-0.025**

* The average number of the total, the short-, medium-, and long-range interactions for residues in proteins with disulfide bonds for the intact polypeptide chain, after the removal of all half cystine centered heptapeptides and in disulfide free proteins.

‡ The difference between the average number of interactions for disulfide free proteins and for proteins with disulfide bonds after the removal of all half cystine centered heptapeptides.

§ "modified" means parts of proteins left after the removal of all half cystine centered heptapeptides.

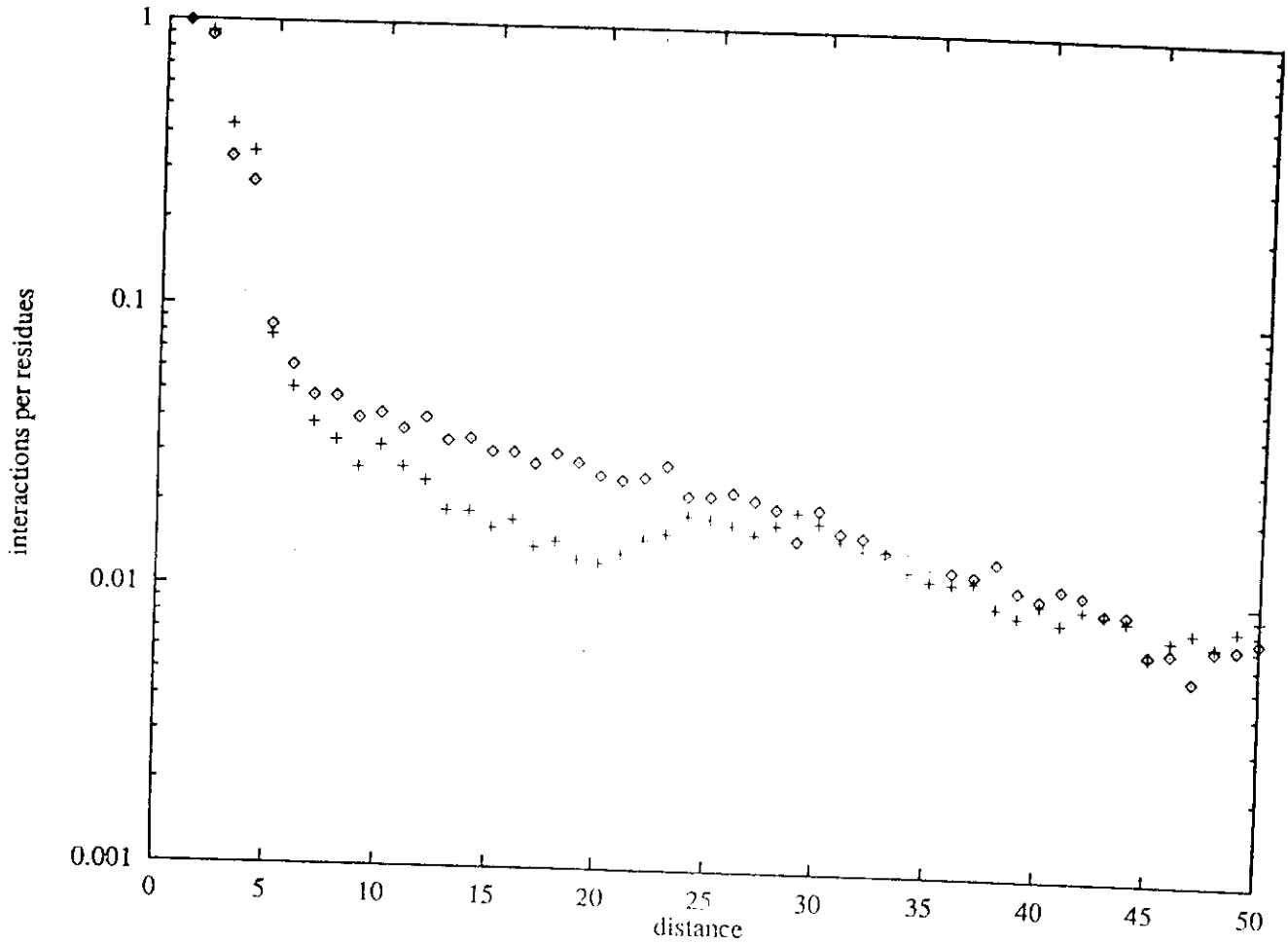


FIGURE 4 The average of the number of interactions per residues as a function of sequential distance, on semi-logarithmic scale, for extracellular (◊) and intracellular (+) proteins.

TABLE 3 The number and percentage of residues in different secondary structures for the studied extra- and intracellular proteins .*

all proteins			extracellular proteins			intracellular proteins		
h:	3955	28.63%	h:	1025	21.29%	h:	2930	32.56%
b:	2941	21.29%	b:	1328	27.59%	b:	1613	17.93%
t:	1717	12.43%	t:	626	13.00%	t:	1091	12.12%
c:	5199	37.64%	c:	1835	38.12%	c:	3364	37.39%

* h: helices, b: sheet, t: turn, c: coil.

Table 1

RReM scoring matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W
C	0.04																		
D	−0.08	−0.24																	
E	0.08	−0.51	0.37																
F	−0.25	0.30	−0.40	−0.50															
G	0.11	0.25	0.11	−0.23	−0.09														
H	−0.11	0.18	0.22	−0.19	−0.06	0.18													
I	−0.10	0.26	−0.54	−0.51	0.60	−0.22	−0.14												
K	−0.07	−0.43	0.18	0.54	−0.41	−0.23	−0.09	−0.40											
L	0.20	0.23	−0.55	−0.45	0.44	−0.14	−0.12	0.60	−0.34										
M	0.15	0.10	−0.25	−0.14	0.22	−0.01	−0.19	0.36	−0.16	0.38									
N	−0.14	−0.14	0.56	0.13	−0.33	0.22	0.32	−0.37	0.22	−0.39	−0.11								
P	−0.07	0.02	0.09	−0.11	−0.12	0.25	0.05	−0.19	−0.17	−0.20	−0.17	0.05							
Q	−0.01	−0.29	0.19	0.51	−0.44	−0.13	0.05	−0.43	0.52	−0.30	−0.18	0.20	−0.03						
R	−0.07	−0.20	−0.01	0.34	−0.31	−0.11	0.13	−0.29	0.47	−0.20	−0.17	0.12	−0.17	0.57					
S	0.14	0.10	0.22	−0.17	−0.01	0.29	0.14	−0.27	−0.24	−0.19	−0.17	0.14	0.27	−0.06	−0.16				
T	0.11	0.14	0.23	−0.27	0.02	−0.02	0.08	0.00	−0.24	−0.04	−0.02	0.13	0.20	−0.15	−0.32	0.45			
V	0.06	0.33	−0.48	−0.44	0.47	−0.12	−0.18	0.70	−0.36	0.49	0.23	−0.46	−0.17	−0.38	−0.29	−0.18	0.07		
W	−0.25	0.18	−0.41	−0.22	0.41	−0.08	−0.11	0.36	−0.22	0.22	0.16	−0.39	−0.13	−0.17	−0.03	−0.17	−0.12	0.29	
Y	−0.36	0.21	−0.34	−0.32	0.54	−0.24	0.08	0.43	−0.20	0.17	0.07	−0.27	−0.23	−0.28	−0.03	−0.21	−0.14	0.31	0.43

Protein Engineering vol.8 no.5 pp.417–423, 1995

22.

Independence divergence-generated binary trees of amino acids

Gábor E.Tusnády, Gábor Tusnády[1] and István Simon[2]

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1518, PO Box 7 and [1]Mathematical Institute, Hungarian Academy of Sciences, Budapest H-1364, PO Box 127, Hungary

[2]To whom correspondence should be addressed

The discovery of the relationship between amino acids is important in terms of the replacement ability, as used in protein engineering homology studies, and gaining a better understanding of the roles which various properties of the residues play in the creation of a unique, stable, 3-D protein structure. Amino acid sequences of proteins edited by evolution are anything but random. The measure of non-randomness, i.e. the level of editing, can be characterized by an independence divergence value. This parameter is used to generate binary tree relationships between amino acids. The relationships of residues presented in this paper are based on protein building features and not on the physico-chemical characteristics of amino acids. This approach is not biased by the tautology present in all sequence similarity-based relationship studies. The roles which various physico-chemical characteristics play in the determination of the relationships between amino acids are also discussed.

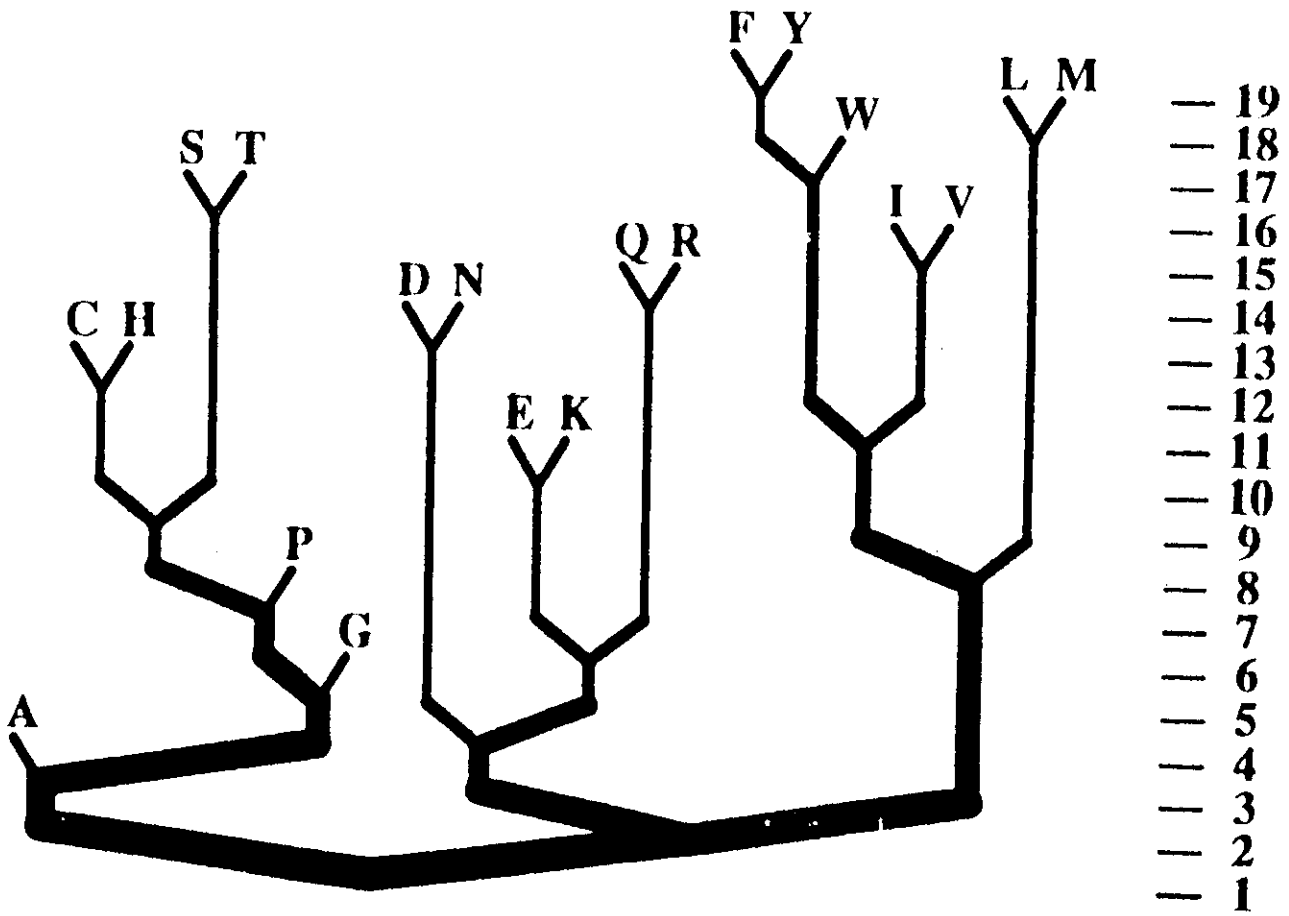
Key words: amino acid distance matrix/homology studies/ protein design/sequence analysis

Tsunoda, G. E. et al. (1995) Protein Eng. 8, 417

Divergence value: $D = \sum_i p_i \cdot \ln\left(\frac{p_i}{q_i}\right)$

p_i : observed probability

q_i : expected probability



Volume 302, number 2, 117–120 FEBS 11024 May 1992

24.

Hypothesis

Different sequence environments of cysteines and half cystines in proteins

Application to predict disulfide forming residues

András Fiser, Miklós Cserzö, Éva Tüdös and István Simon

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1518, PO Box 7, Hungary

Received 17 March 1992; revised version received 30 March 1992

Protein sequences are often derived by translating genetic information, rather than by classical protein sequencing. At the DNA level cysteines and half cystines are indistinguishable. Here we show that the sequential environments of 'free' cysteine and half cystine are different. A possible origin of this difference is discussed and a simple method to predict cysteines and half cystines from the amino acid sequence is also presented.

Prediction; Free cysteine; Half cystine; Sequential environment

25.

Different sequence environments of amino acid residues involved and not involved in long-range interactions in proteins

ÉVA TÜDÖS, ANDRÁS FISER and ISTVÁN SIMON

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Received 22 March, accepted for publication 17 July 1993

No method has yet been available to decode information, hidden in the protein primary structure, on long-range interactions of amino acids. Even a limited amount of information on long-range interactions could help in conformational energy calculations of protein structures and could lead to a better understanding of how the primary structure of proteins determines their conformation.

The sequence environments of amino-acid residues were compared from the viewpoint of their participation in long-range interactions. By using the simplest definition, residues were considered as partners in a long-range interaction if they were at least 20 residues apart in the sequence and their C_α distance was less than 7 Å.

In spite of this rather crude definition, an analysis of 88 unrelated proteins has shown that the sequence environments (10 residues on each side) of those amino acids which are involved in long-range interactions and of those which are not are significantly different according to the criteria of mathematical statistics. Moreover, in many cases the differences are so pronounced that the involvement of a given amino acid in long-range interactions can be predicted from its sequence environment.

Fiser, A. et al. (1992) FEBS Letters 302, 117 26.

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	Cys	1	2	3	4	5	6	7	8	9	10
A	1.20	0.73	1.11	0.72	0.59	0.99	0.99	1.18	1.17	1.12		1.74	1.26	0.95	1.15	0.65	0.88	0.70	0.63	0.74	1.54
C	0.77	1.64	0.62	1.65	1.25	1.44	1.44	0.50	0.44	0.69		0.81	0.41	0.61	1.33	1.54	1.30	1.79	0.78	1.39	0.87
D	0.94	0.81	1.16	1.60	1.94	0.62	0.82	1.62	1.08	1.04		0.90	0.73	1.17	1.60	1.35	1.06	1.09	1.16	1.31	1.28
E	0.66	1.27	0.51	1.88	1.27	1.06	1.03	1.25	0.81	0.97		0.94	0.81	0.71	0.44	0.74	0.53	0.57	0.59	0.49	0.46
F	0.90	0.17	0.41	0.43	0.88	0.49	0.52	0.79	0.65	0.80		0.97	0.49	0.99	0.41	1.21	0.59	1.11	0.75	0.96	1.08
G	1.55	2.07	1.35	1.08	1.18	1.94	1.90	1.25	1.24	1.29		1.95	1.75	1.38	1.53	1.04	1.46	1.54	1.25	0.92	1.53
H	0.78	0.72	0.68	0.43	0.43	0.67	0.22	1.00	1.08	1.31		0.22	0.23	0.78	2.25	0.76	0.69	0.35	0.54	0.51	0.44
I	0.70	0.98	0.58	0.96	1.16	1.06	0.47	0.65	1.30	1.02		0.42	1.31	0.94	0.76	1.30	1.31	0.90	1.36	1.13	0.78
K	0.67	0.77	0.87	1.08	0.81	1.26	0.67	0.90	1.11	0.72		1.00	0.94	1.15	0.79	1.16	0.64	1.12	1.02	1.01	0.87
L	0.60	0.57	0.71	0.54	0.44	0.56	0.61	0.21	0.78	0.50		0.48	0.59	0.35	0.74	0.45	0.46	0.94	0.81	0.30	0.61
M	0.66	0.80	0.34	0.53	0.76	0.62	0.33	0.46	1.99	0.44		0.54	0.75	0.48	0.53	0.37	0.59	0.54	0.36	0.57	0.44
N	1.60	1.44	1.55	0.87	1.06	0.88	1.91	2.08	2.03	1.25		1.80	1.58	1.65	1.47	1.07	2.35	1.89	1.52	1.17	1.03
P	0.93	0.72	1.18	1.03	0.87	0.59	0.99	0.89	0.33	0.81		0.70	0.61	0.70	0.74	0.84	1.14	1.38	1.05	0.81	1.57
Q	0.79	1.18	0.82	1.13	1.17	0.92	1.03	1.07	0.40	0.72		1.05	1.11	0.91	0.84	1.24	0.65	0.73	0.88	1.34	0.53
R	0.95	1.12	0.99	1.09	0.94	0.50	0.53	0.72	1.10	0.23		0.77	0.88	0.61	0.47	0.49	0.77	0.90	0.68	0.99	0.41
S	1.51	0.94	1.55	0.98	1.06	1.09	1.07	1.53	0.83	0.98		1.69	1.40	1.52	1.16	1.29	1.23	0.92	1.38	0.96	1.12
T	0.80	1.03	1.21	1.12	1.03	2.04	1.23	1.43	1.10	1.36		0.73	1.40	1.53	0.87	0.98	1.36	0.56	1.32	1.24	1.27
V	1.05	1.26	0.71	0.69	1.29	1.05	1.12	0.81	0.73	1.08		0.62	1.25	0.95	0.57	1.02	0.63	0.56	0.61	1.14	1.27
W	0.81	0.17	2.57	1.63	1.25	1.15	1.66	0.14	0.51	1.67		0.78	0.36	1.17	1.28	1.77	1.24	1.42	1.00	2.15	1.34
Y	1.63	1.28	2.05	1.55	1.31	1.03	1.37	1.25	2.73	1.31		2.21	0.89	1.81	1.60	2.22	1.63	1.36	2.50	1.59	1.03

Ratio of the abundance of residues in the vicinity of half cystines and "free" cysteines

Tüdös, E. et al. (1994) Int. J. Pept. Prot. Res. 43, 205

Amino acid	Total number of occurrences	Prediction power (%)	No. of residues with extreme potential	Prediction power (%)
Ala	[illegible]	59.61	[illegible]	70.58
Cys	[illegible]	61.69	[illegible]	66.08
Asp	[illegible]	60.85	[illegible]	56.31
Glu	[illegible]	59.44	[illegible]	69.31
Phe	[illegible]	56.70	146 (21.5%)	68.49
Gly	[illegible]	52.63	[illegible]	64.64
His	360	63.14	153 (42.5%)	67.97
Ile	699	59.84	165 (23.6%)	66.23
Lys	[illegible]	59.29	187 (13.6%)	62.84
Leu	[illegible]	57.89	171 (12.4%)	72.51
Met	332	57.14	158 (47.6%)	55.70
Asn	703	56.47	192 (27.3%)	62.50
Pro	556	59.41	157 (28.2%)	68.15
Gln	611	55.97	202 (33.0%)	67.82
Arg	684	54.68	131 (19.2%)	61.83
Ser	1171	58.67	172 (14.7%)	66.28
Thr	1013	55.38	151 (12.0%)	61.83
Val	1257	58.26	173 (14.2%)	68.97
Trp	239	54.58	127 (53.0%)	55.91
Tyr	541	54.58	149 (25.3%)	57.72

Prediction residues involved in long-range interactions in proteins

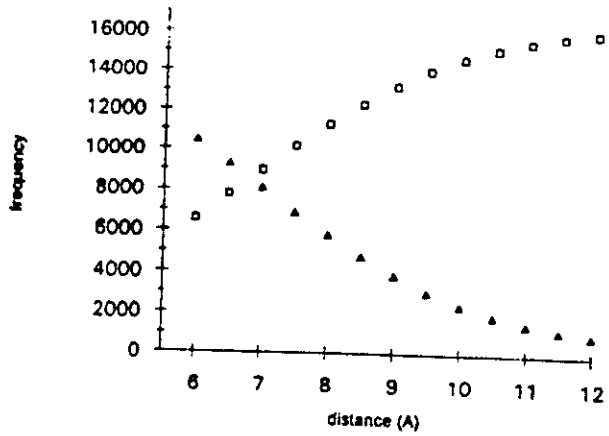
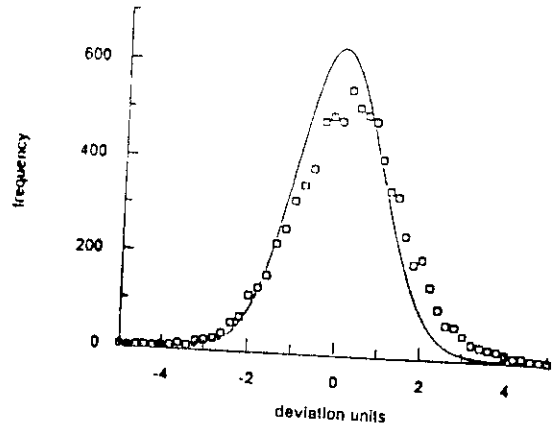


FIGURE 1

Number of residues recognized as interacting (□) and non-interacting (△) as a function of the upper limit of the C_α distances.



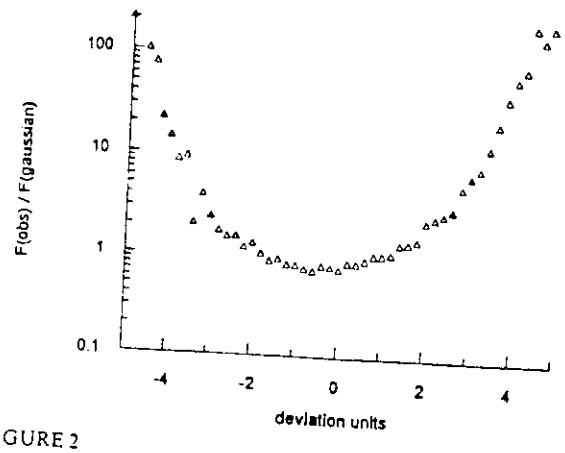


FIGURE 2

Comparison of the distribution of the theoretically expected values (Gaussian) and the observed 8000 values representing the differences between the sequence environments of residues depending on their involvement in long-range interactions in deviation units (a). The elements of the latter were calculated as $(Y_{i,j,k} - N_{i,j,k})/(DY_{i,j,k} + DN_{i,j,k})$; where $Y_{i,j,k}$ is the value of the ith residue in the kth relative position with the jth interacting residue, and $N_{i,j,k}$ is the same parameter for non-interacting residues. DY and DN are the standard deviation values of the corresponding matrix elements. (b) shows the ratio between the two functions of (a) on a semi-logarithmic scale.

J. Mol. Biol. (1994) **243**, 388–396

COMMUNICATIONS

New Alignment Strategy for Transmembrane Proteins

M. Cserzö[1,3], J.-M. Bernassau[2], I. Simon[3] and B. Maigret[1]

[1]*Laboratoire de Chimie Théorique*
URA CNRS No. 510 Université de Nancy-I-BP239, 54506 Vandoeuvre les Nancy Cedex, France

[2]*SANOFI Recherche*
rue du Professeur J. Blayac, 34082 Montpellier Cedex 04, France

[3]*Institute of Enzymology*
Biological Research Center, Hungarian Academy of Sciences, 1518 P.O. Box 7, Budapest, Hungary

In this paper an algorithm which locates helical transmembrane segments is described. It is shown that given the location of transmembrane helices of a protein, corresponding helices in another membrane related protein can be pinpointed. The method seems to be extremely insensitive to sequence identity but highly sensitive to the property of a sequence to assume transmembrane helical structure. As an example, using the present method, a sequence alignment between bacteriorhodopsin and human rhodopsin is carried out and it provides a good starting point for homology modeling of this G-protein coupled receptor. It is difficult to obtain this particular alignment using the traditional methods because of poor sequence homology. There are indications that hint at the broader range of applicability of the presented method.

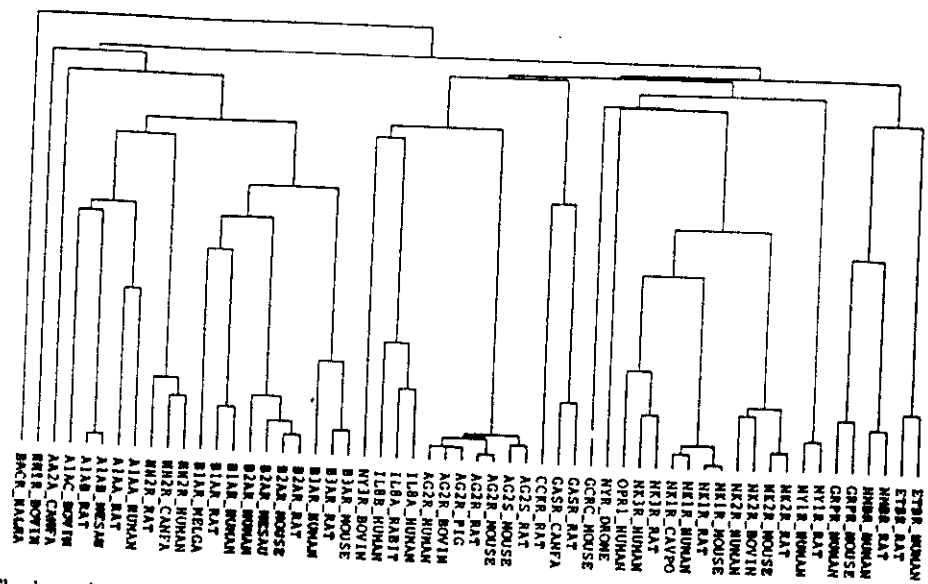
Keywords: transmembrane helix; G-protein coupled receptor; signal peptide; sequence alignment; homology modeling

Cserö et al. J. Mol. Biol (1994) 243, 388

RReM scoring matrix

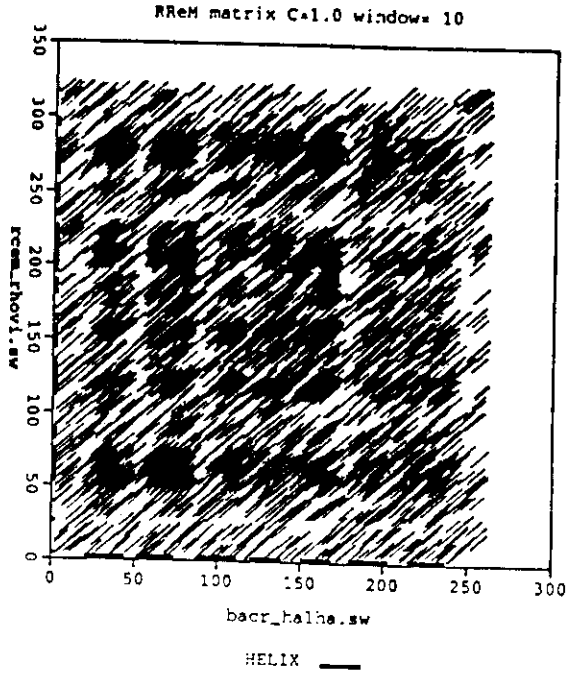
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W
C	0.04																		
D	−0.08	−0.24																	
E	0.08	−0.51	0.37																
F	−0.25	0.30	−0.40	−0.50															
G	0.11	0.25	0.11	−0.23	−0.09														
H	−0.11	0.18	0.22	−0.19	−0.06	0.18													
I	−0.10	0.26	−0.54	−0.51	0.60	−0.22	−0.14												
K	−0.07	−0.43	0.18	0.54	−0.41	−0.23	−0.09	−0.40											
L	0.20	0.23	−0.55	−0.45	0.44	−0.14	−0.12	0.60	−0.34										
M	0.15	0.10	−0.25	−0.14	0.22	−0.01	−0.19	0.36	−0.16	0.38									
N	−0.14	−0.14	0.56	0.13	−0.33	0.22	0.32	−0.37	0.22	−0.39	−0.11								
P	−0.07	0.02	0.09	−0.11	−0.12	0.25	0.05	−0.19	−0.17	−0.20	−0.17	0.05							
Q	−0.01	−0.29	0.19	0.51	−0.44	−0.13	0.05	−0.43	0.52	−0.30	−0.18	0.20	−0.03						
R	−0.07	−0.20	−0.01	0.34	−0.31	−0.11	0.13	−0.29	0.47	−0.20	−0.17	0.12	−0.17	0.57					
S	0.14	0.10	0.22	−0.17	−0.01	0.29	0.14	−0.27	−0.24	−0.19	−0.17	0.14	0.27	−0.06	−0.16				
T	0.11	0.14	0.23	−0.27	0.02	−0.02	0.08	0.00	−0.24	−0.04	−0.02	0.13	0.20	−0.15	−0.32	0.45			
V	0.06	0.33	−0.48	−0.44	0.47	−0.12	−0.18	0.70	−0.36	0.49	0.23	−0.46	−0.17	−0.38	−0.29	−0.18	0.07		
W	−0.25	0.18	−0.41	−0.22	0.41	−0.08	−0.11	0.36	−0.22	0.22	0.16	−0.39	−0.13	−0.17	−0.03	−0.17	−0.12	0.29	
Y	−0.36	0.21	−0.34	−0.32	0.54	−0.24	0.08	0.43	−0.20	0.17	0.07	−0.27	−0.23	−0.28	−0.03	−0.21	−0.14	0.31	0.43

PILEUP of 73, 7 transmembrane helices proteins

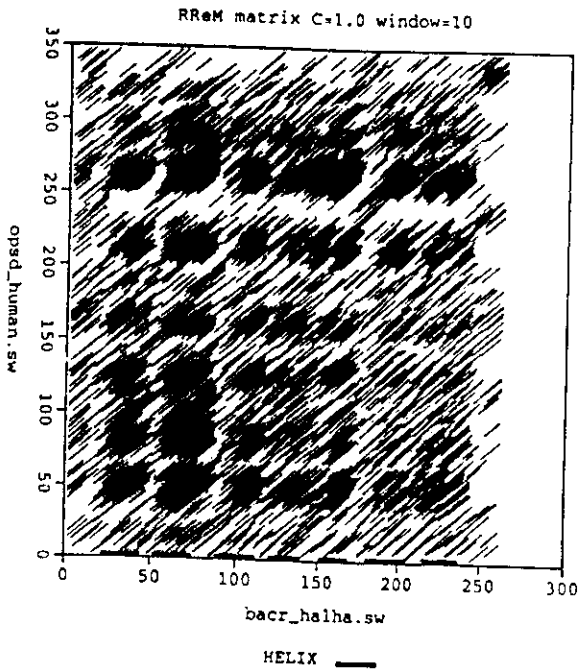


The homology clustering of 73 transmembrane proteins. The BR (SWISS-PROT code: bacr_halha) appears as a stand-alone branch on the tree. The Figure is made by the 'pileup' utility of the GCG package.

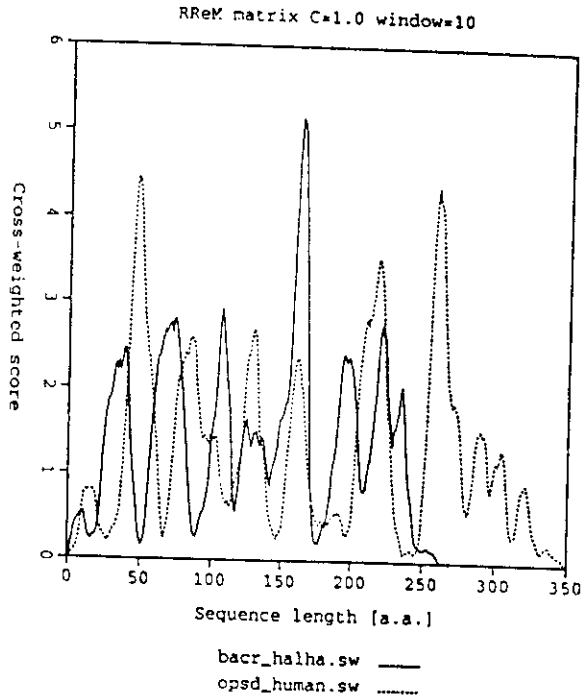
Berò et al. J.M.B. 1994



The alignment surface of the BR *versus* PRC-M. A 10 residue window size has been appli[illegible].



The alignment surface of the BR *versus* HR. A 10 residue window size has been applied.



Cross-weighted cumulative score profiles of BR and HR.

Conformational energy terms

Coulomb term:

$$U_e = \frac{332\, q_i\, q_j}{D\, r_{ij}}$$

Lennard-Jones 6-12 potencial:

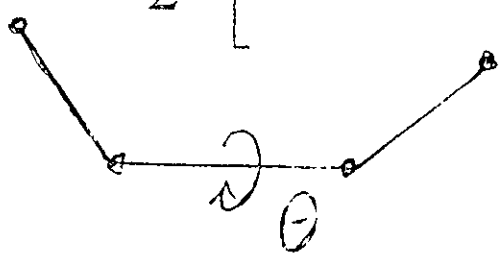
$$U_{6\text{-}12} = \varepsilon_{ij}\left[F\left(\frac{r_{ij}^o}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^o}{r_{ij}}\right)^{6}\right]$$

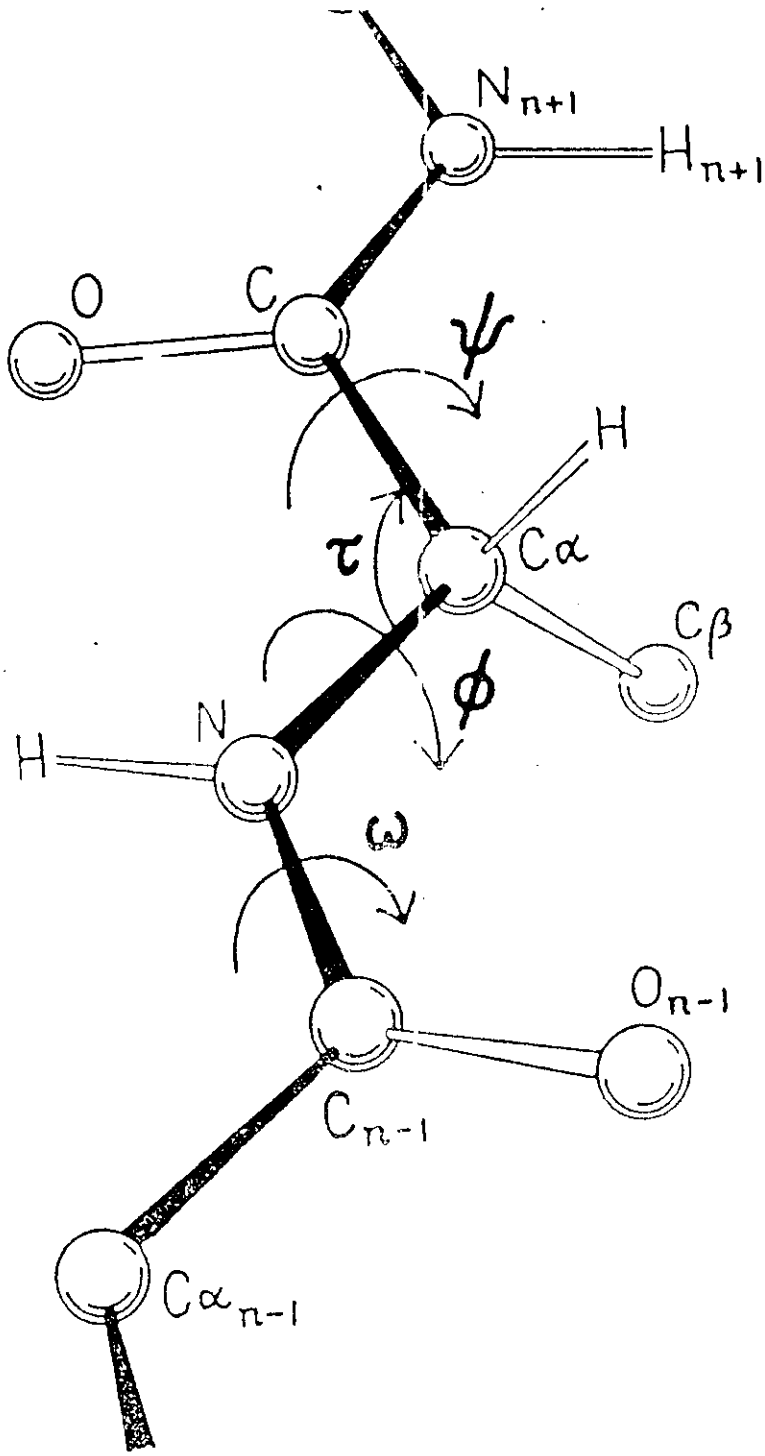
Hydrogen-bonding potencial:

$$U_{HB} = \varepsilon_{ij}\left[\left(\frac{r_{ij}^o}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^o}{r_{ij}}\right)^{10}\right]$$

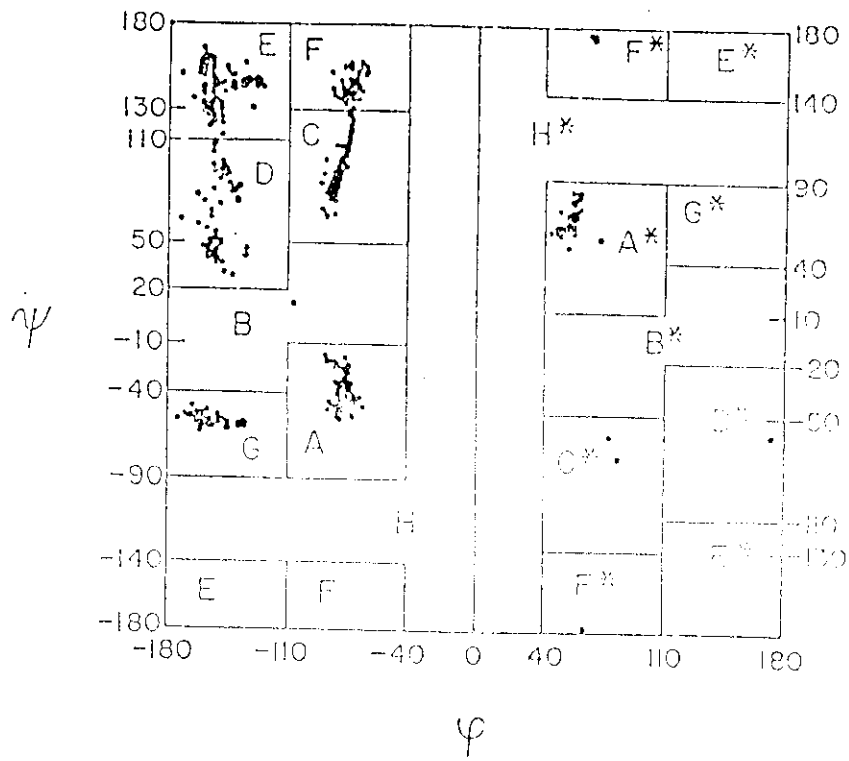
Intrinsic torsional potencial:

$$U_t = \frac{U_o}{2}\left[1 + k\cos(n\theta)\right]$$





Peptide bond



Energy minima in the ϕ,ψ map

Reprinted from **MACROMOLECULES**, Vol. 11, Page 797, July-August 1978

Conformational Energy Calculations of the Effects of Sequence Variations on the Conformations of Two Tetrapeptides[1]

István Simon,[2] George Némethy, and Harold A. Scheraga*

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received December 8, 1977

ABSTRACT: Conformational energy calculations were carried out on the two terminally blocked tetrapeptides *N*-acetyl-Thr-Asp-Gly-Lys-*N'*-methylamide and *N*-acetyl-Ala-Asp-Gly-Lys-*N'*-methylamide. The first peptide is a sequence variant of tetrapeptides studied earlier in this laboratory. The second peptide occurs in a bend at residues 94–97 in staphylococcal nuclease. A selection strategy is described which helps to accelerate the search of starting conformations used for energy minimization. The strategy involves exhaustive searches for conformations of fragments of the molecule which are stabilized by specific interactions and subsequent combination of fragments, prior to minimization. Several groups of low-energy conformations were found. They are compactly folded structures, but they differ from the "standard" chain reversals. One group, which is of low energy in both peptides, is stabilized by Asp···Asp and Asp···Lys backbone–side chain hydrogen bonds. Another group, of low energy in the Thr-containing peptides, is stabilized by a network of hydrogen bonds involving polar atoms of both backbone and side chains of the Thr, Asp, and Lys residues. The conformation corresponding to the sequence fragment in staphylococcal nuclease has relatively high energy, indicating that the bend observed in the protein is stabilized by interactions involving parts of the protein outside the tetrapeptide sequence.

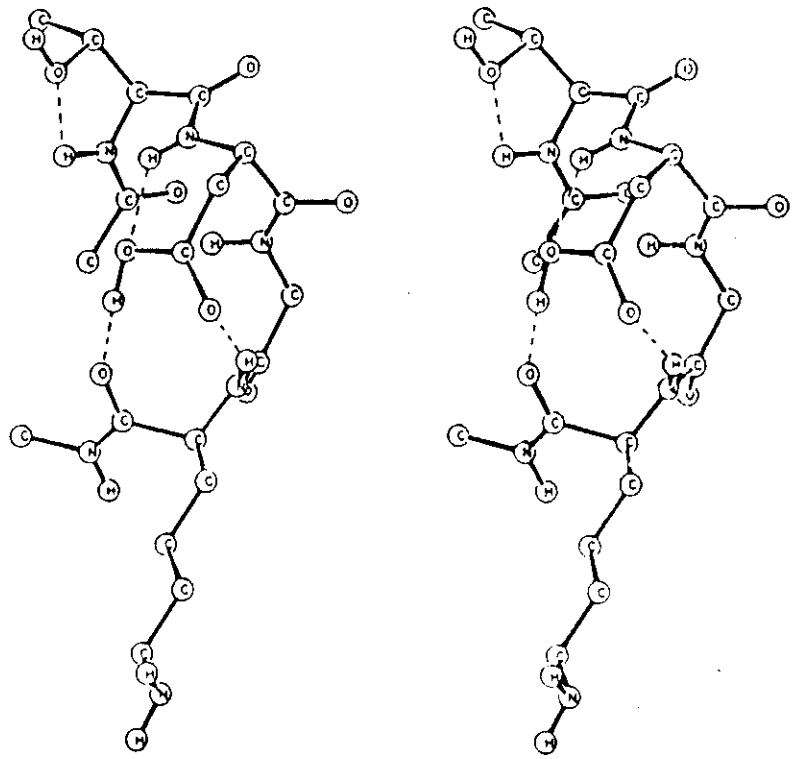


Figure 2. Stereoscopic illustration of the lowest energy conformation, (a), of peptide A and conformation (c) of peptide T. The positions of backbone and side-chain atoms are identical (within 0.1 Å) in both peptides, except for the $C^{\gamma 2}$, O^{γ}, and H^{γ} atoms of threonine in peptide T which do not occur in peptide A. Hydrogen bonds are indicated by dashed lines. Backbone and side-chain aliphatic hydrogen atoms have been omitted for clarity.

Table I. Minimum-energy Conformations of Ac-Ala-Asp-Gly-Lys-NMe

Classification[a]		Short-hand notation for dihedral angles[b] Ala (ϕ,ψ)	Asp (ϕ,ψ)	Asp χ^1	Asp χ^2	Asp $\chi^{3,2}$	Gly (ϕ,ψ)	Lys (ϕ,ψ)	Lys χ^1	Lys χ^2	Lys χ^3	Lys χ^4	Lys χ^5	Energy, δE kcal/mol
(i)	(a)	A	A	g^+	p	t	D[e]	F	t	t	t	t	g^+	0.0[c]
(i)L	(b)	A	A	g^+	p	t	D[e]	F	t	t	t	g^-	g^+	0.9
(ii)	(c)	A	A	g^+	n	t	C[e]	C	t	t	t	t	g^+	0.9
(i)	(d)	C	A	g^+	p	t	D[e]	C	t	t	t	t	g^+	1.4
(ii)		C	A	g^+	n	t	C[e]	C	t	t	t	t	g^+	1.7
(ii)L		A	A	g^+	n	t	C[e]	C	t	t	t	g^-	g^+	1.8
(i)		E	A	g^+	p	t	D[e]	C	t	t	t	t	g^+	1.9
(ii)		E	A	g^+	n	t	C[e]	C	t	t	t	t	g^+	2.1
(i)		D	A	g^+	p	t	D[e]	C	t	t	t	t	g^+	2.4
(ii)		G	A	g^+	p	t	D[e]	C	t	t	t	t	g^+	2.5
(ii)		D	A	g^+	n	t	C[e]	C	t	t	t	t	g^+	2.6
X-ray structure[d]		C	A	g^+	n	t	C[e]	C	t	t	t	t	g^+	3.0
Minimized[f]	(e)	D	A[e]	g^-	p	--[e]	D[e]	F	(see Table II)					$>10^3$
	(f)	C	A[e]	g^-	p	t	C[e]	C	g^-	t	g^+	t	t	6.9
"Standard" bends														
Type II with H-bond[g]		E	C	g^+	g^+	c	D[e]	E	g^+	t	t	t	g^+	6.7
Type I'		E	A[e]	t	g^+	c	D[e]	E	g^+	t	t	t	g^+	7.1
Y		D	D	g^-	g^+	t	C[e]	E	g^+	t	t	t	g^+	8.0
Type II		E	C	g^+	p	t	A[e]	E	g^+	t	t	t	g^+	8.1
Type I		E	A	g^+	p	t	B	E	g^+	t	t	t	g^+	11.5

Footnotes to Table I

[a] The classification is explained in the text. The letters (a) to (f) in parentheses refer to the columns in Table II.

[b] Backbone and side-chain dihedral angles are described in terms of the short-hand notation[16] described in Section II. Symbols not explained there are as follows: p: χ^2_{Asp} near 115°; n: χ^2_{Asp} near -50°; c: $\chi^{3,2}_{Asp}$ near 0° (cis).

[c] E_0 = -19.1 kcal/mol.

[d] Ref. 6.

[e] Not defined.

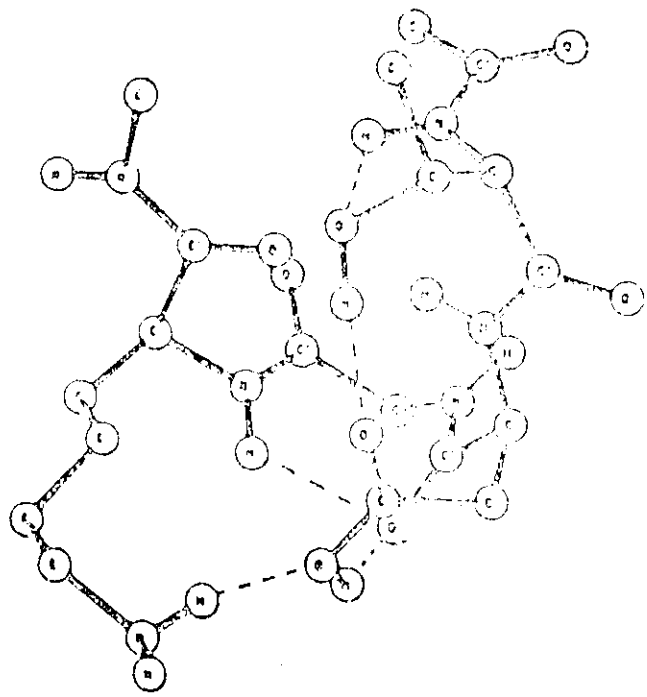
[f] This conformation was obtained by starting from the conformation of residues 94 to 97 in the X-ray crystallographic structure of staphylococcal nuclease, (e) in the Table, and minimizing the energy.

Table II. Dihedral Angles and Energies for Some Low Energy Conformations of Ac-Ala-Asp-Gly-Lys-NMe

Dihedral Angle	Conformation[a] (a)	(b)	(c)	(d)	(e) X-ray[b]	(f) X-ray minimized[c]
ϕ_i	-70	-70	-73	-83	-127	-81
ψ_i	-35	-39	-43	77	106	33
ω_i	179	179	180	179	179	179
ϕ_{i+1}	-73	-73	-78	-78	65	54
ψ_{i+1}	-33	-32	-24	-31	33	47
ω_{i+1}	-173	-174	-179	-174	180	178
χ^1_{i+1}	62	62	63	66	-62	-53
χ^2_{i+1}	114	115	-50	112	129	96
$\chi^{3,2}_{i+1}$	174	174	-179	176	-	-180
ϕ_{i+2}	139	137	109	134	119	108
ψ_{i+2}	-57	-57	-61	-59	-29	-39
ω_{i+2}	175	174	178	177	-165	-174
ϕ_{i+3}	-83	-81	-76	-83	-97	-79
ψ_{i+3}	131	130	100	128	138	93
χ^1_{i+3}	-174	-173	-174	-174	-68	-66
χ^2_{i+3}	176	177	178	177	144	175
χ^3_{i+3}	179	-176	179	179	9	76
χ^4_{i+3}	179	-69	179	179	-135	175
χ^5_{i+3}	62	75	62	62	-	177
	Energy, kcal/mol					
	0.0	0.9	0.9	1.4	$> 10^3$	6.9
	C^α distances, Å					
$C^\alpha_i \cdots C^\alpha_{i+3}$	7.5	7.5	8.1	7.6	-	4.8
$C^\alpha_{i-1} \cdots C^\alpha_{i+4}$	4.1	4.1	5.3	8.9	-	4.1

[a] The letters in parentheses refer to the classification in Table I. Dihedral angles are in degrees. χ for Ala was always within 1° of 60°.

[b] Ref. 6. [c] See footnote f of Table I.



Lowest energy structure of Thr-Asp-Gly-Lys

	Thr			Asp			Gly	Lys					ΔE KJ/mol
1.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	-g	g	0,0
2.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	-g	-g	6,3
3.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	g	t	6,7
4.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	t	g	7,1
5.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	t	t	7,6
6.	A	g	g	D	g	y	t	C^{π}	C	t	t	t	t	g	8,0
7.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	t	-g	8,4
8.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	g	g	8,4
9.	A	g	y	D	g	y	t	C^{π}	C	t	t	t	t	g	9,7
10.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	-g	t	10,5
11.	A	g	g	D	g	-g	c	C^{π}	F	t	t	t	t	g	11,3
12.	A	g	g	D	g	-g	c	C^{π}	F	g	t	t	g	-g	11,3
13.	A	g	t	A	g	y	t	D^{π}	F	t	t	t	t	g	12,2
14.	A	-g	t	A	g	y	t	D^{π}	F	t	t	t	t	g	12,6
15.	A	g	g	A	g	y	t	D^{π}	C	t	t	t	t	g	13,4
16.	E	t	t	A	g	y	t	D^{π}	C	t	t	t	t	g	13,4
17.	C	g	g	A	g	y	t	D^{π}	C	t	t	t	t	g	13,9
18.	A	g	g	D	g	y	t	C^{π}	C	-g	t	t	g	-g	13,9
19.	E	t	g	A	g	y	t	D^{π}	C	t	t	t	t	g	14,7
20.	C	g	g	A	g	-g	t	C^{π}	C	t	t	t	t	g	14,7
21.	A	g	t	A	g	-g	t	D^{π}	C	t	t	t	t	g	15,1
22.	A	-g	g	A	g	y	t	D^{π}	F	t	t	t	t	g	15,5
23.	E	t	t	A	g	-g	t	C^{π}	C	t	t	t	t	g	15,5
24.	E	t	g	A	g	-g	t	C^{π}	C	t	t	t	t	g	16,0
25.	A	g	g	D	g	-g	c	C^{π}	F	-g	-g	t	-g	t	16,0
26.	A	g	g	A	g	-g	t	C^{π}	C	t	t	t	t	g	16,4
27.	A	-g	t	A	g	-g	t	C^{π}	C	t	t	t	t	g	16,4
28.	A	g	g	D	g	-g	c	C^{π}	E	t	t	t	t	g	18,1
29.	A	g	g	E	g	-g	t	C	G	t	t	t	t	g	21,0

Low energy conformation of tetrapeptide Thr-Asp-Gly-Lys

Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 3661–3665, May 1991
Biophysics

39.

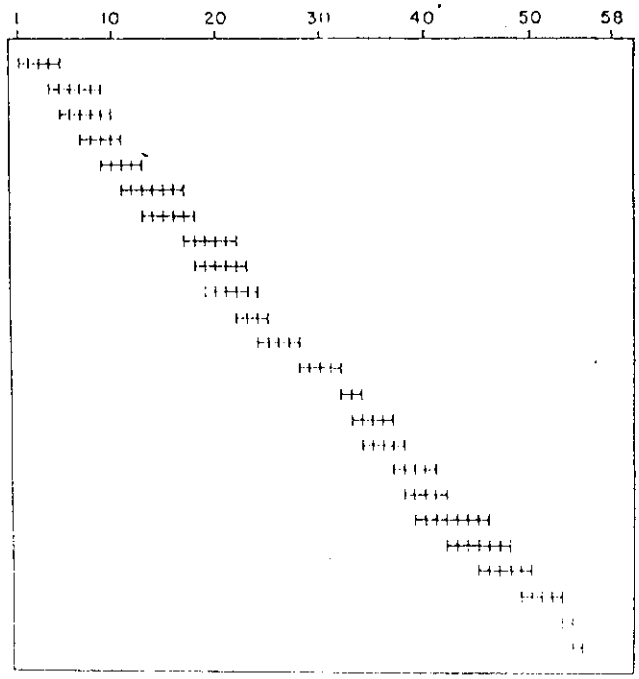
Calculation of protein conformation as an assembly of stable overlapping segments: Application to bovine pancreatic trypsin inhibitor

(conformational energy calculations/short-range interactions/build-up procedure/"conformon")

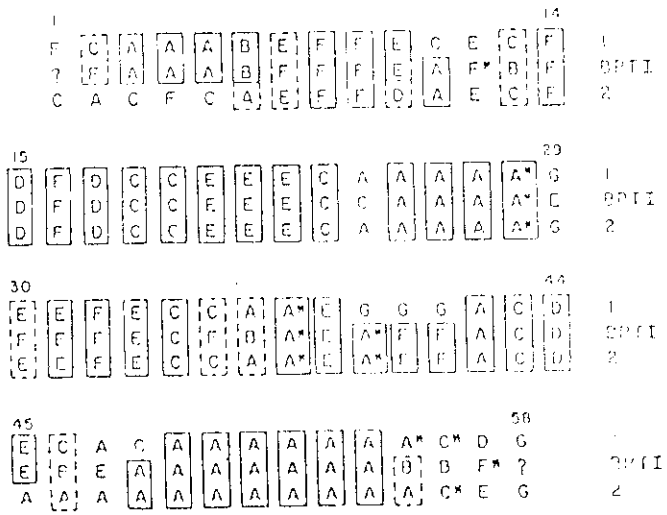
ISTVAN SIMON*, LESLIE GLASSER†, AND HAROLD A. SCHERAGA‡

Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853-1301

ABSTRACT Conformations of bovine pancreatic trypsin inhibitor were calculated by assuming that the final structure as well as properly chosen overlapping segments thereof are simultaneously in low-energy (not necessarily the lowest-energy) conformational states. Therefore, the whole chain can be built up from building blocks whose conformations are determined primarily by short-range interactions. Our earlier buildup procedure was modified by taking account of a statistical analysis of known amino acid sequences that indicates that there is nonrandom pairing of amino acid residues in short segments along the chain, and by carrying out energy minimization on only these segments and on the whole chain [without minimizing the energies of intermediate-size segments (20–30 residues long)]. Results of this statistical analysis were used to determine the variable sizes of the overlapping oligopeptide building blocks used in the calculations; these varied from tripeptides to octapeptides, depending on the amino acid sequence. Successive stages of approximations were used to combine the low-energy conformations of these building blocks in order to keep the number of variables in the computations to a manageable size. The calculations led to a limited number of conformations of the protein (only two different groups, with very similar structure within each group), most residues of which were in the same conformational state as in the native structure.



Sizes of low energy structures in bovine pancreatic trypsin inhibitor (BPTI)



Calculated (1,2) and X-ray (BPTI) structures