



SMR: 1098/6

**WORKSHOP ON THE STRUCTURE OF
BIOLOGICAL MACROMOLECULES**

(16 - 27 March 1998)

"Protein Folding"

presented by:

István SIMON

Institute of Enzymology
Hungarian Academy of Sciences
H-1518 Budapest P.O. Box 7
Hungary

These are preliminary lecture notes, intended only for distribution to participants.

Protein Folding

István Simon

Institute of Enzymology

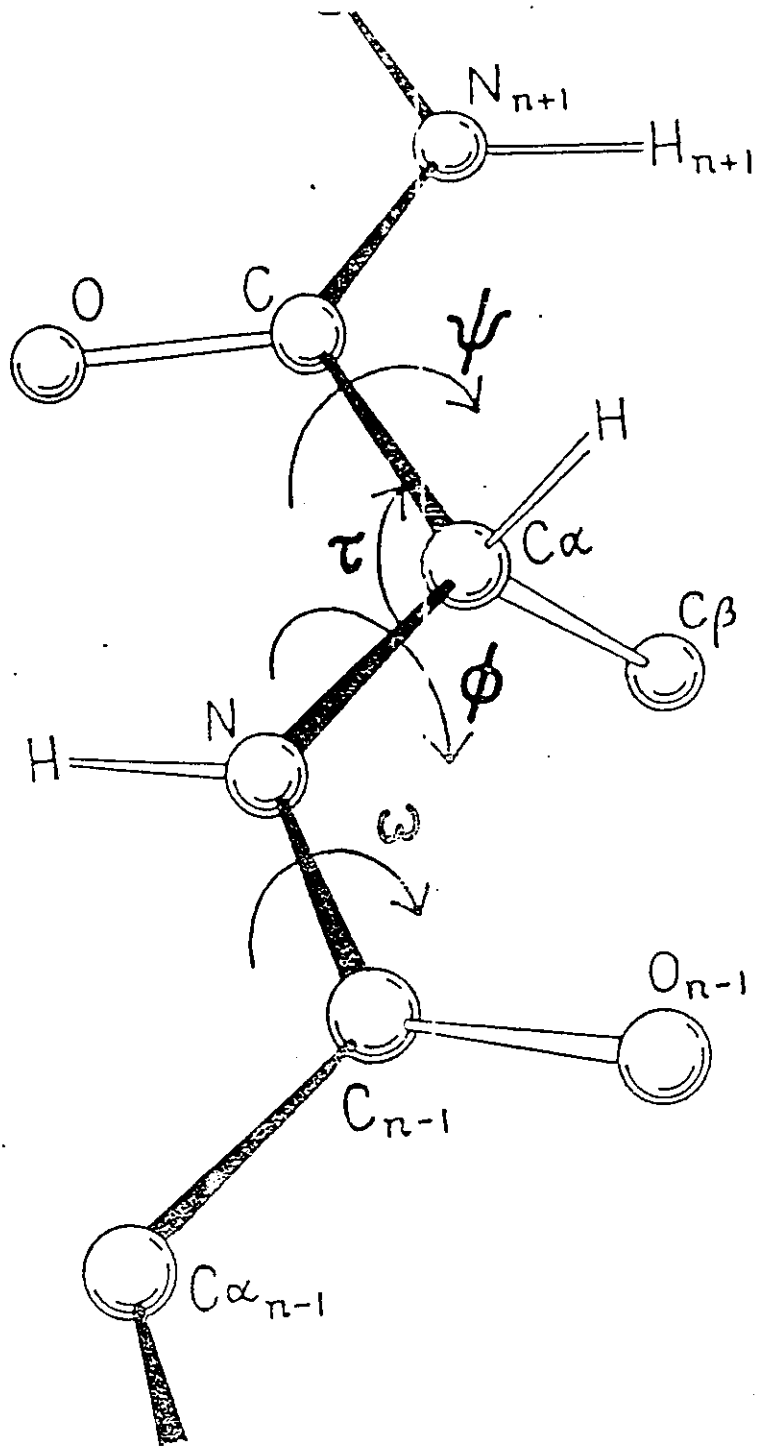
Hungarian Academy of Sciences

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



Peptide bond

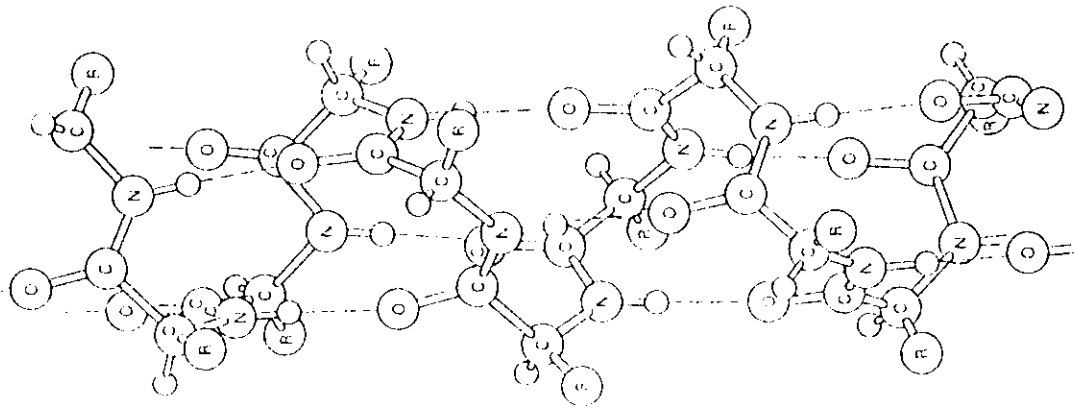


FIGURE 5.6
The classical right-handed α -helix.

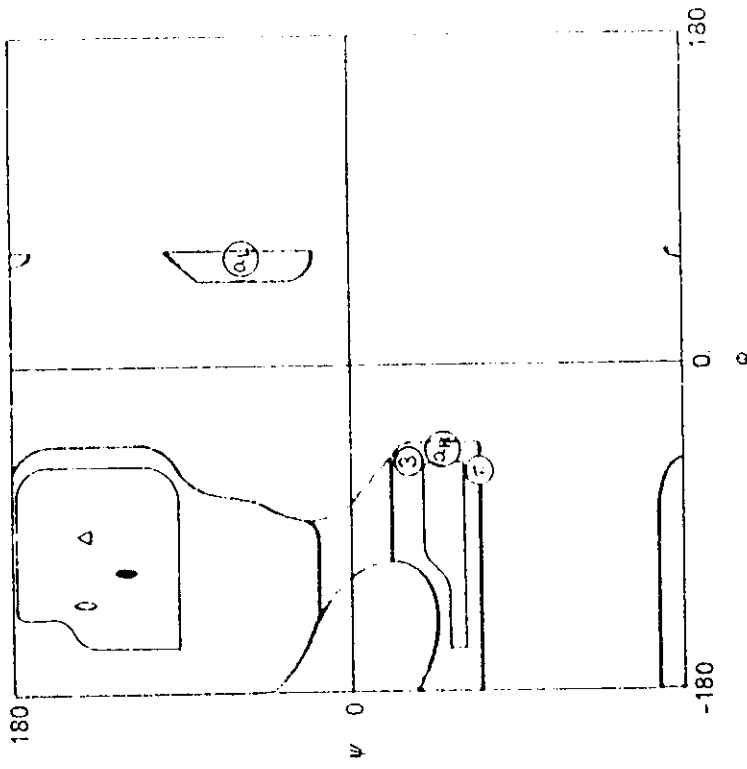


FIGURE 5.7

The positions of the regular conformations of polypeptides on a Ramachandran plot. The regular conformations are α_R , the right-handed α -helix; α_L , the left-handed α -helix; C, the antiparallel β -sheet; β , the parallel β -sheet; β' , the right-handed 3_{10} -helix; π , the right-handed π -helix; Δ , polyPro I, polyPro II, and polyGly II. (From G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.* 23:283-457, 1968.)

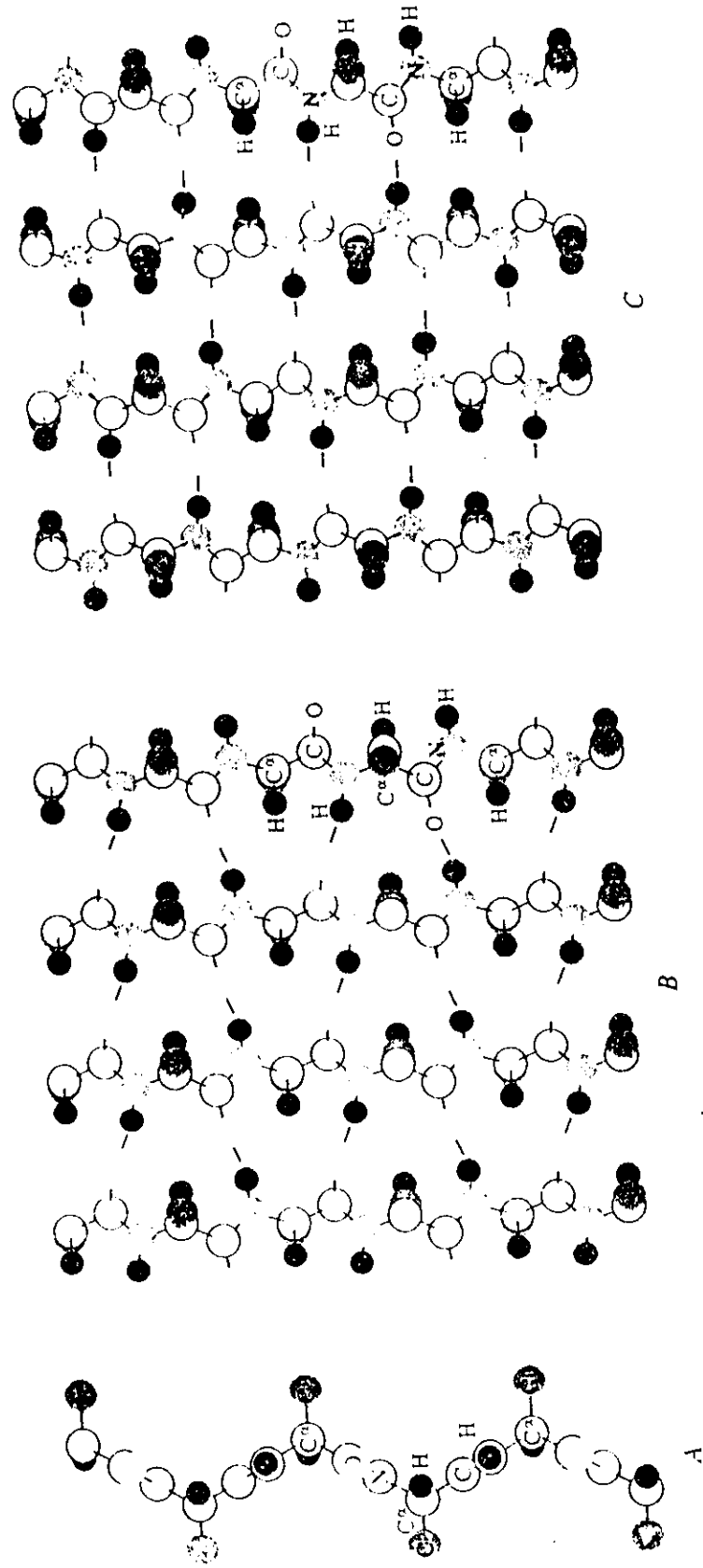
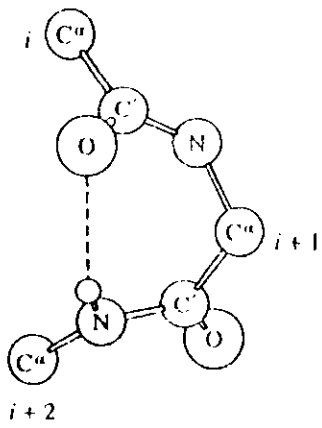
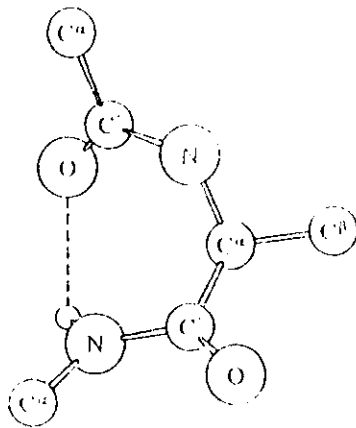


FIGURE 5.9 A single β -strand (A) and its incorporation into flat parallel (B) and antiparallel (C) β -sheets.

Classical γ turn

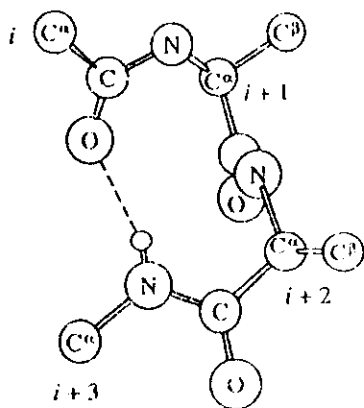


Inverse γ turn

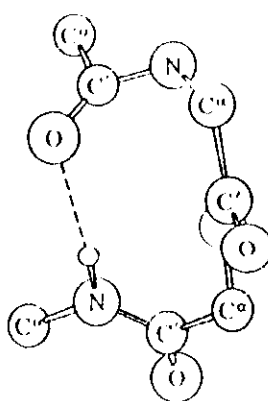


β Turns

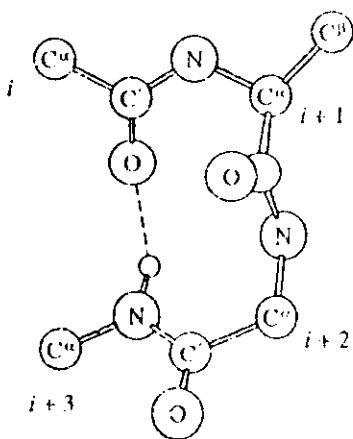
Type I



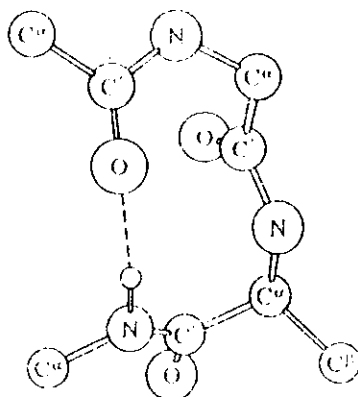
Type I'



Type II



Type II'



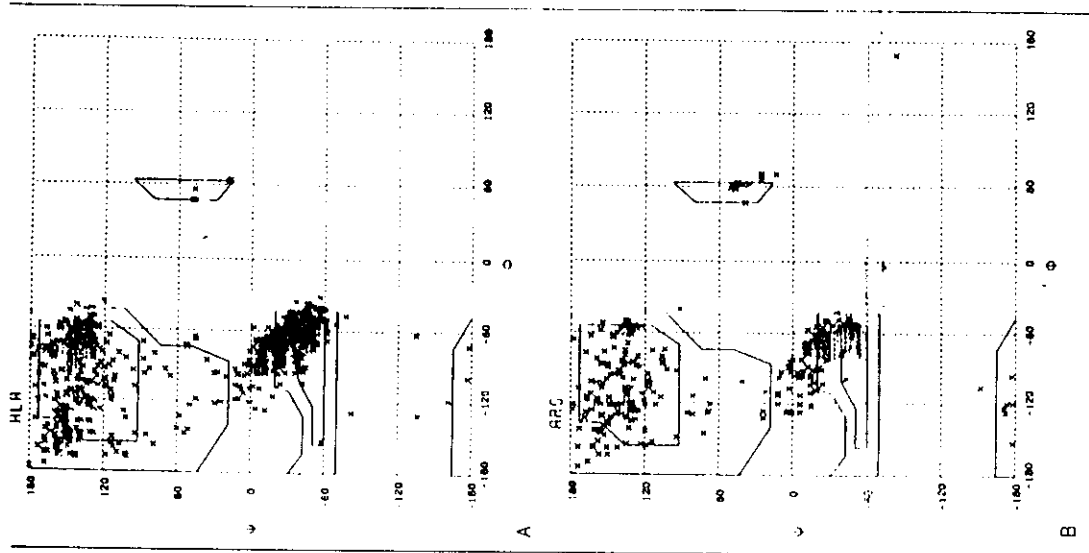
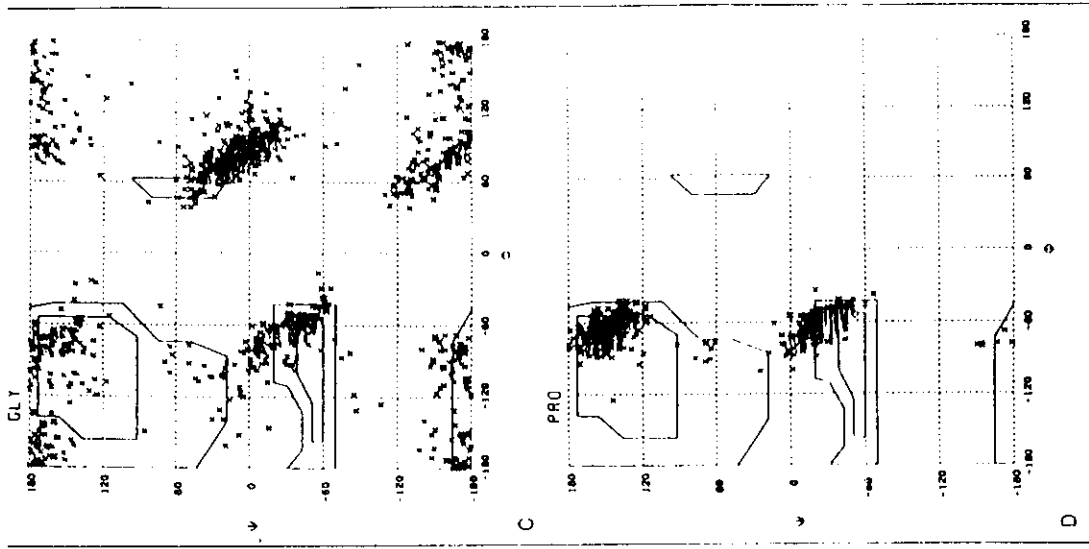


FIGURE 1-2. Ramachandran ϕ - ψ plots for four residue types as found in high-resolution X-ray structures of proteins. The allowed areas based on hard sphere atoms with two different sets of assumed radii are shown by the contours. The fits for all other residues not shown are



similar to those for the alanine and arginine distributions. The very restricted area for proline, ϕ close to -60° , and the highly expanded area around glycine, symmetrical about $\phi = 0$, $\psi = 0$, are shown in the bottom two panels. (These patterns were provided by M. Kozlowski using the program FRAGLE; Finelli et al., 1990.)

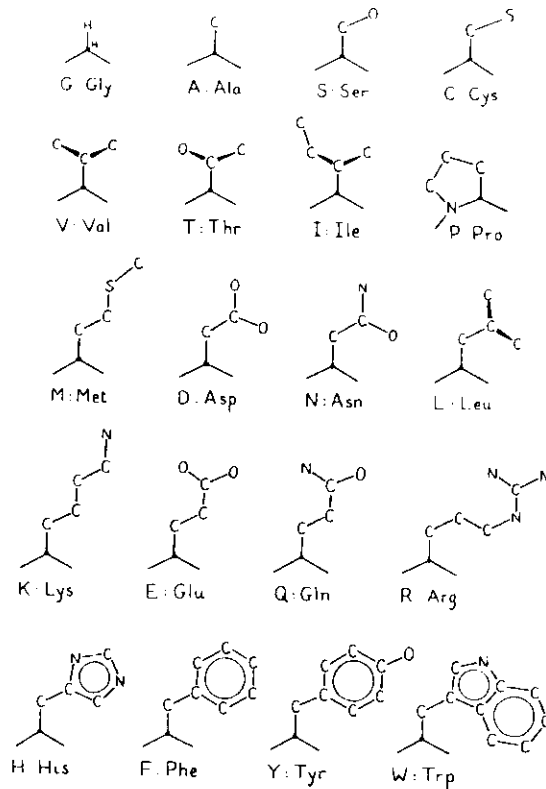


Figure 57. Structures of the 20 naturally occurring amino acid residues laid out according to size and shape of side chain, with three-letter and one-letter abbreviations: glycine, alanine, serine, cysteine; valine, threonine, isoleucine, proline; methionine, aspartate, asparagine, leucine; lysine, glutamate, glutamine, arginine; histidine, phenylalanine, tyrosine, tryptophan. Hydrogens are shown only for glycine. The α carbons are black dots; tapered bonds indicate the directionality at tetrahedral carbons; circles indicate aromatic rings.

Table II. Conformational Parameters for α -Helical, β -Sheet, and β -Turn Residues in 29 Proteins^a

	P_α		P_β		P_i		f_i		f_{i+1}		f_{i+2}		f_{i+3}		
Glu	1.51	H _α	Val	1.70	H _β	Asn	1.56	Asn	0.161	Pro	0.301	Asn	0.191	Trp	0.167
Met	1.45		Ile	1.60		Gly	1.56	Cys	0.149	Ser	0.139	Gly	0.190	Gly	0.152
Ala	1.42		Tyr	1.47		Pro	1.52	Asp	0.147	Lys	0.115	Asp	0.179	Cys	0.128
Leu	1.21		Phe	1.38		Asp	1.46	His	0.140	Asp	0.110	Ser	0.125	Tyr	0.125
Lys	1.16	h _α	Trp	1.37	h _β	Ser	1.43	Ser	0.120	Thr	0.108	Cys	0.117	Ser	0.106
Phe	1.13		Leu	1.30		Cys	1.19	Pro	0.102	Arg	0.106	Tyr	0.114	Gln	0.098
Gln	1.11		Cys	1.19		Tyr	1.14	Gly	0.102	Gln	0.098	Arg	0.099	Lys	0.095
Trp	1.08		Thr	1.19		Lys	1.01	Thr	0.086	Gly	0.085	His	0.093	Asn	0.091
Ile	1.08	I _α	Gln	1.10	i _β	Gln	0.98	Tyr	0.082	Asn	0.083	Glu	0.077	Arg	0.085
Val	1.06		Met	1.05		Thr	0.96	Trp	0.077	Met	0.082	Lys	0.072	Asp	0.081
Asp	1.01		Arg	0.93		Trp	0.96	Gln	0.074	Ala	0.076	Thr	0.065	Thr	0.079
His	1.00		Asn	0.89		Arg	0.95	Arg	0.070	Tyr	0.065	Phe	0.065	Leu	0.070
Arg	0.98	i _α	His	0.87	h _β	His	0.95	Met	0.068	Glu	0.060	Trp	0.064	Pro	0.068
Thr	0.83		Ala	0.83		Glu	0.74	Val	0.062	Cys	0.053	Gln	0.037	Phe	0.065
Ser	0.77		Ser	0.75		Ala	0.66	Leu	0.061	Val	0.048	Leu	0.036	Glu	0.064
Cys	0.70		Gly	0.75		Met	0.60	Ala	0.060	His	0.047	Ala	0.035	Ala	0.058
Tyr	0.69	h _α	Lys	0.74	B _β	Phe	0.60	Phe	0.059	Phe	0.041	Pro	0.034	Ile	0.056
Asn	0.67		Pro	0.55		Leu	0.59	Glu	0.056	Ile	0.034	Val	0.028	Met	0.055
Pro	0.57		Asp	0.54		Val	0.50	Lys	0.055	Leu	0.025	Met	0.014	His	0.054
Gly	0.57		Glu	0.37		Ile	0.47	Ile	0.043	Trp	0.013	Ile	0.013	Val	0.053

^a P_α , P_β , and P_i are conformational parameters of helix, β sheet, and β turns. $f_i, f_{i+1}, f_{i+2}, f_{i+3}$ are bend frequencies in the four positions of the β turn. H_α, H_β, etc., as defined previously (Chou and Fasman, 1974b). From Chou and Fasman (1977, 1978).

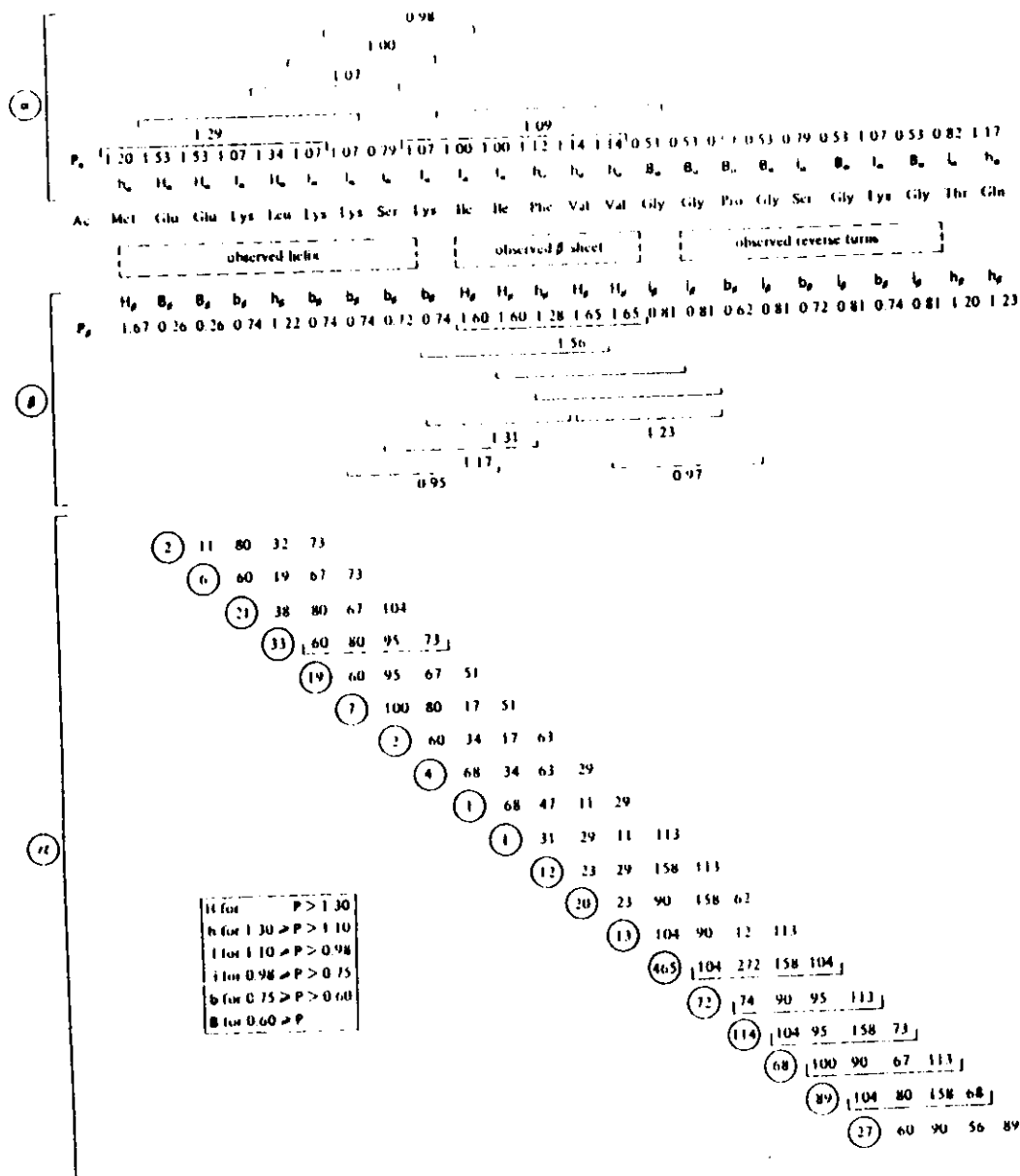


Figure 6-2. Prediction of α -helix, β -pleated sheet and reverse turns (*rt*) for the 24 N-terminal residues of adenylate kinase (389), following the procedure of Chou and Fasman (340). The α - and β -propensities P_α and P_β are taken from Tables 6-1 and 6-2. The symbols are defined in the inset. For α -predictions hexapeptides with average propensities above 1.00 are given in solid lines. Solid lines are also used for pentapeptides with average β -propensity above 1.00. The ends of helices and sheets are tested with tetrapeptides. Tetrapeptides with average propensities above 1.00 are given with dash-dot lines whereas those with averages below 1.00 are given with dashed lines. Note that in our nomenclature the average propensities could be called potentials. Reverse turn propensities depend on the position of the residue in question within the tetrapeptide to be tested. Thus, there are four propensities for each residue. All of them are multiplied by a factor of 10^{10} . The *rt*-potential is the product of the propensities in a tetrapeptide. These potentials are marked with circles and are given on the same line as the *rt*-propensities of the tetrapeptide. All potentials are multiplied by 10^{16} . Tetrapeptides with potentials above $50 \cdot 10^{-6}$ are marked with solid lines. A relative maximum which does not reach the threshold is indicated by a dashed line.

Amino acid	Residue position																
	$j-8$	$j-6$	$j-4$	$j-2$	j	$j+2$	$j+4$	$j+6$	$j+8$								
Gly	0	0	-5	-10	-20	-30	-50	-70	-100	-70	-50	-30	-20	-10	-5	0	0
Ala	0	5	15	20	30	40	45	55	60	55	45	40	30	20	15	5	0
Val	0	0	0	0	-5	-15	-10	-5	0	-5	-10	-10	-5	0	0	0	0
Leu	0	0	5	5	5	10	15	20	25	20	15	10	5	5	5	0	0
Ile	0	0	0	0	0	0	5	10	15	10	5	0	0	0	0	0	0
Ser	0	0	-5	-10	-15	-20	-25	-30	-35	-33	-30	-25	-20	-15	-10	-5	-5
Thr	0	-5	-10	-15	-20	-25	-35	-40	-45	-40	-35	-25	-20	-15	-10	-5	0
Asp	0	0	0	0	10	15	10	5	0	-10	-25	-35	-25	-10	-5	0	0
Glu	25	28	32	35	44	50	60	65	70	55	40	20	10	5	5	0	0
Asn	0	0	0	0	-5	-10	-20	-30	-42	-30	-20	-10	-5	0	0	0	0
Gln	0	0	0	0	0	0	0	0	15	20	25	25	20	10	0	0	0
Lys	5	5	5	5	5	5	10	15	28	40	55	58	50	45	42	40	40
His	0	0	0	0	0	0	0	0	5	15	25	32	36	40	42	43	44
Arg	0	0	-10	-18	-20	-7	10	20	30	36	36	30	20	10	0	0	0
Phe	0	0	0	0	5	15	22	28	30	28	22	15	5	0	0	0	0
Tyr	0	-3	-10	-15	-22	-30	-35	-47	-35	-20	-10	-10	-10	-12	-17	-25	-45
Trp	25	40	30	20	10	42	42	33	20	12	5	3	0	0	0	0	0
Cys	-15	-15	-15	-15	-15	-12	-10	-5	0	0	-10	-25	-35	-43	-46	-47	-37
Met	5	5	8	10	20	30	38	45	50	48	45	42	35	30	20	15	10
Pro	0	-5	-12	-20	-25	-35	-50	-66	-96	-177	-118	-90	-70	-52	-40	-30	-20

Table III. Directional Information Values $I(S_j = X:\bar{X};R_{j+m})$ for Extended Conformation $X = E$

Amino acid	Residue position																
	$j-8$	$j-6$	$j-4$	$j-2$	j	$j+2$	$j+4$	$j+6$	$j+8$								
Gly	5	12	22	35	40	33	22	-10	-45	-10	22	33	40	35	22	12	5
Ala	0	0	0	-5	-10	-15	-20	-30	-40	-30	-20	-15	-10	-5	0	0	0
Val	0	0	0	5	10	20	45	70	90	70	45	20	10	5	0	0	0
Leu	0	0	0	0	0	0	10	25	33	25	5	-15	-25	-20	-10	-5	0
Ile	-15	-25	-40	-15	0	15	35	60	70	60	35	15	0	-15	-40	-25	-15
Ser	25	25	25	22	18	15	5	0	-5	0	5	15	18	22	25	25	25
Thr	5	5	5	7	10	13	21	35	28	22	18	15	15	15	15	15	15
Asp	0	0	0	0	0	-15	-25	-105	-55	-15	0	0	0	0	0	0	0
Glu	-10	-10	-15	-20	-27	-35	-45	-55	-65	-77	-63	-50	-40	-30	-20	-15	-10
Asn	5	12	20	20	15	-10	-45	-90	-60	-30	-5	10	25	30	35	30	20
Gln	15	20	20	10	5	0	-5	-15	-20	-35	-20	-15	0	10	15	20	20
Lys	0	0	0	0	-5	-12	-23	-35	-53	-70	-58	-45	-37	-30	-25	-20	-15
His	0	-5	-15	-5	0	0	0	0	0	0	5	15	25	15	5	0	0
Arg	0	0	0	0	0	0	0	0	-5	-10	-20	-25	-28	-25	-10	-5	0
Phe	-20	-35	-60	-60	-45	-30	0	25	40	25	0	-30	-45	-60	-60	-35	-20
Tyr	0	0	0	0	7	15	27	40	40	27	15	0	0	0	0	0	0
Trp	-15	-25	-40	-45	-80	-15	0	10	15	17	20	20	20	20	20	20	20
Cys	0	0	-20	-60	-55	-40	-20	5	15	17	10	5	0	0	0	0	0
Met	-20	-65	-90	-80	-60	-30	-5	15	30	15	-15	-45	-50	-45	-40	-30	-25
Pro	20	20	20	20	10	0	-30	-65	-110	-65	-30	0	10	20	20	20	20

Table IV. Directional Information Values $I(S_j = X:\bar{X};R_{j+m})$ for β Turn $X = T$

Amino acid	Residue position																
	$j-8$	$j-6$	$j-4$	$j-2$	j	$j+2$	$j+4$	$j+6$	$j+8$								
Gly	0	0	0	-5	-10	-20	-30	0	70	80	10	0	0	0	0	0	0
Ala	0	0	0	0	0	-3	-5	-10	-15	-22	-28	-32	-20	-7	-3	0	0
Val	0	0	0	0	-5	-10	-30	-50	-95	-50	-30	-10	-5	0	0	0	0
Leu	0	0	0	0	0	-3	-10	-15	-50	-15	-10	-3	0	0	0	0	0
Ile	0	0	0	-5	-20	-30	-50	-80	-140	-80	-30	-15	0	0	0	0	0
Ser	0	0	0	0	0	15	30	40	30	5	-15	-30	-15	-5	0	0	0
Thr	0	0	-5	-8	-10	-15	-18	-25	-42	-15	10	20	10	5	0	0	0
Asp	0	0	0	0	5	10	30	35	30	15	5	0	0	0	0	0	0
Glu	0	0	0	0	-5	-10	-10	25	20	-5	-10	20	20	5	0	0	0
Asn	0	0	-5	-20	-5	10	30	45	55	55	30	15	0	-10	-20	-5	0
Gln	0	0	0	0	0	0	0	0	0	0	0	-25	-60	-25	-15	-10	-15
Lys	0	0	0	0	0	3	5	18	30	25	10	-10	-25	-20	-12	-5	0
His	0	0	0	3	10	25	38	5	-20	15	15	-5	10	35	15	0	0
Arg	0	0	0	5	10	25	35	20	0	-35	-5	0	0	0	0	0	0
Phe	0	10	20	25	20	10	-15	-55	-50	-40	-30	-15	0	10	20	25	0
Tyr	0	0	0	5	10	20	30	5	-25	-10	0	15	30	40	30	15	0
Trp	0	-15	-45	15	35	25	10	-15	-57	-20	-10	-5	0	0	0	0	0
Cys	0	0	5	30	50	60	40	5	10	25	25	10	45	30	10	0	0
Met	0	0	0	0	0	-5	-25	-55	-85	-55	-25	-5	0	0	0	0	0
Pro	0	0	0	0	0	20	54	40	-154	40	50	20	0	0	0	0	0

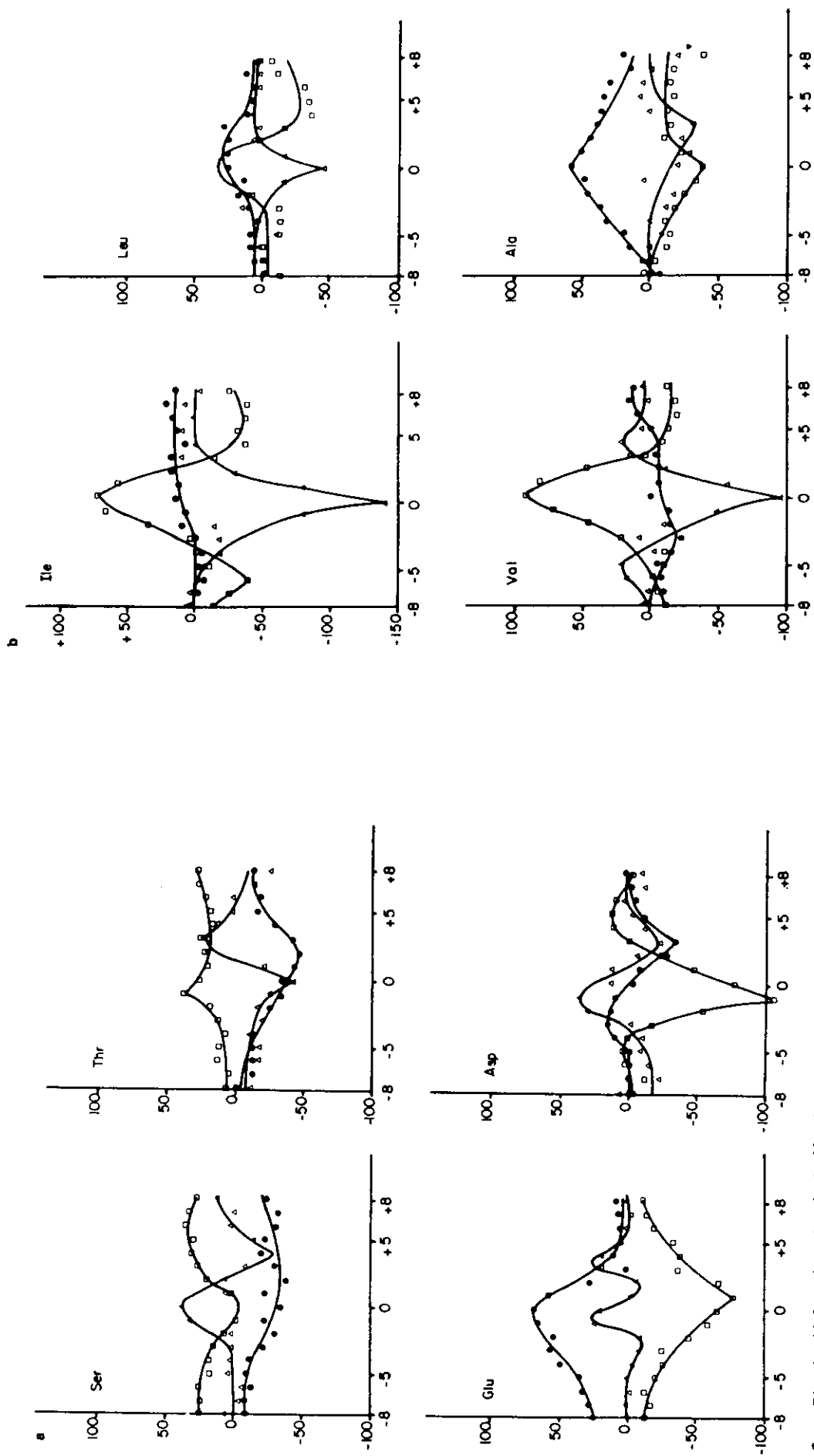
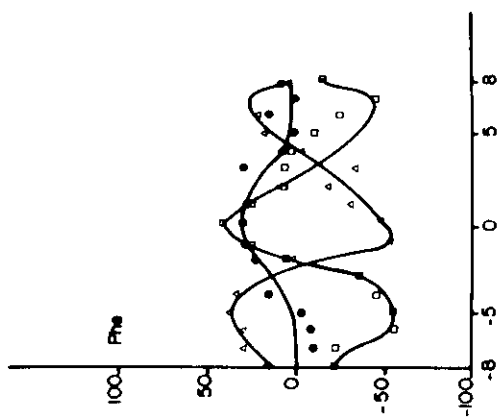
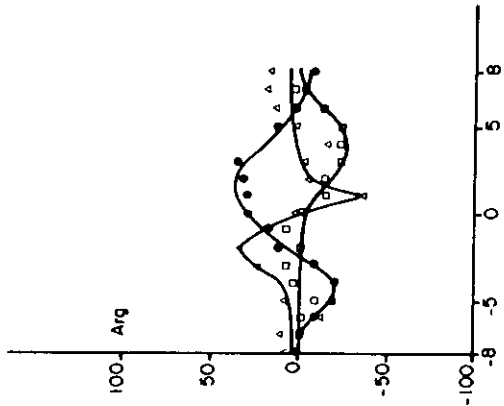
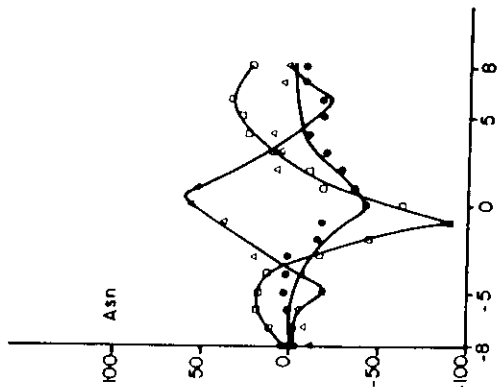
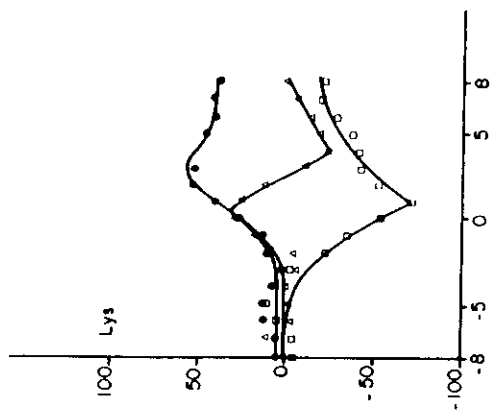
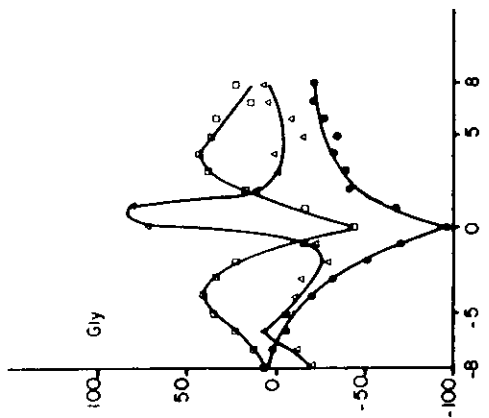
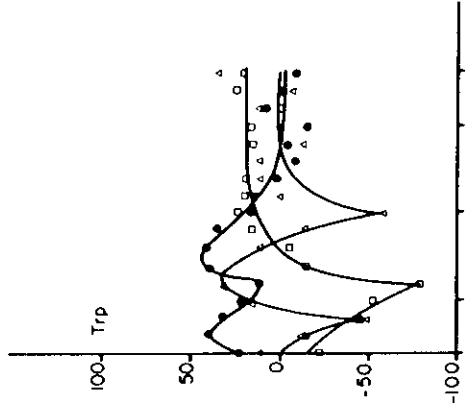
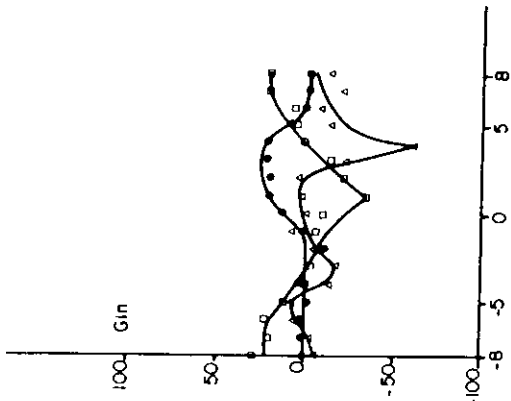
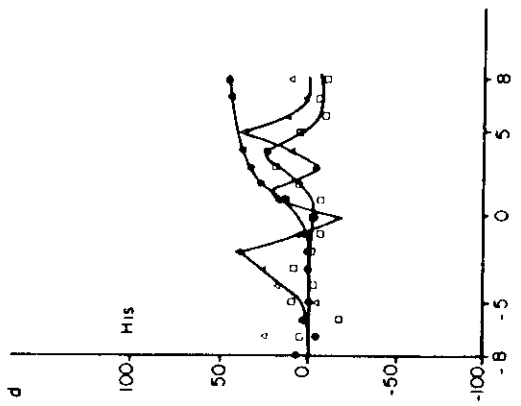
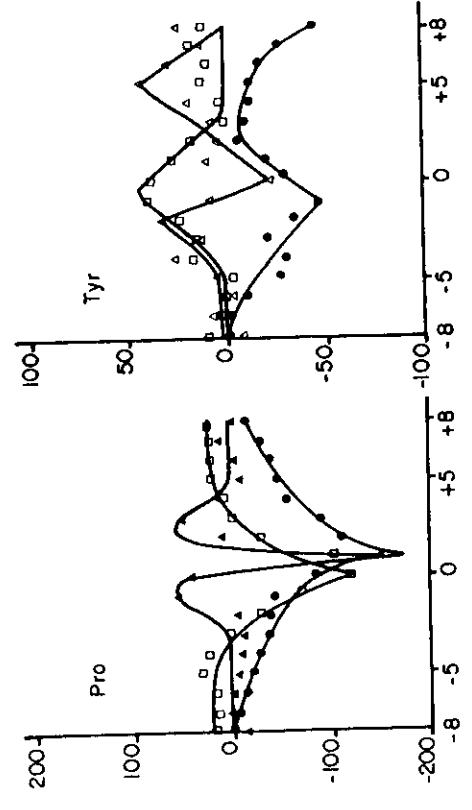
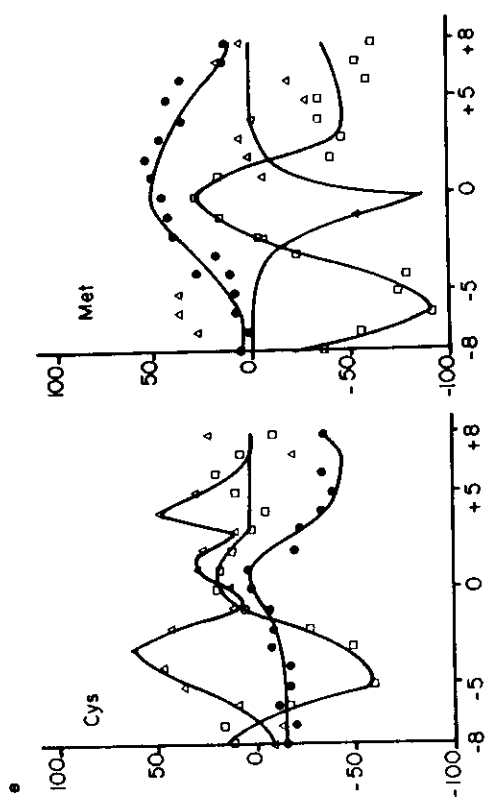


Figure 5a-e. Directional information values for the 20 amino acids, $I(S = X:i;R_j + m)$ in centinats with $m > 0$ for residue R on the C-terminal side of the position j : ●, α helix; □, β sheet; Δ or \blacktriangle , β turn. From Gibrat, 1986.



d



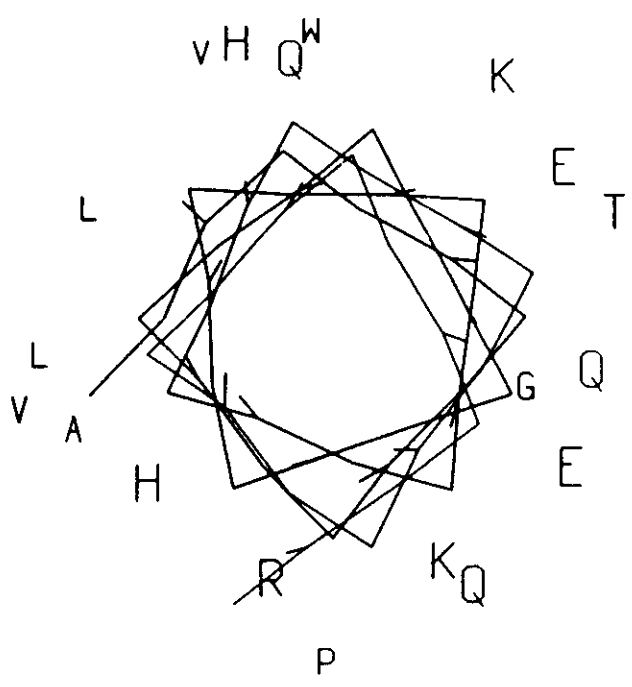


Figure 102. Example of a helix-wheel layout for an actual globular protein helix (TIM 178–196). Hydrophobic residues, mainly on the left, are in small letters, and hydrophilic residues in large letters.

The role of long-range interactions in defining the secondary structure of proteins is overestimated

András Fiser, Zsuzsanna Dosztányi and István Simon¹

Abstract

Motivation: Secondary structure predictions based on the properties of individual residues, and sometimes on local interactions, usually fail to exceed 65% efficiency. Therefore, non-local, long-range interactions seem to be a significant cause of this limitation.

Results: In this paper, we apply approaches to localize highly interacting residues and clusters of residues involved in multiple non-local interactions, and test various secondary structure predictions on this separate subset to assess the effect of long-range interactions on the prediction efficiencies. It was found that only a marginal part of the failure of secondary structure predictions results from the presence of long-range interactions. Alternative possibilities are also discussed.

Contact: fiser@enzim.hu; zsuzsa@enzim.hu; simon@enzim.hu

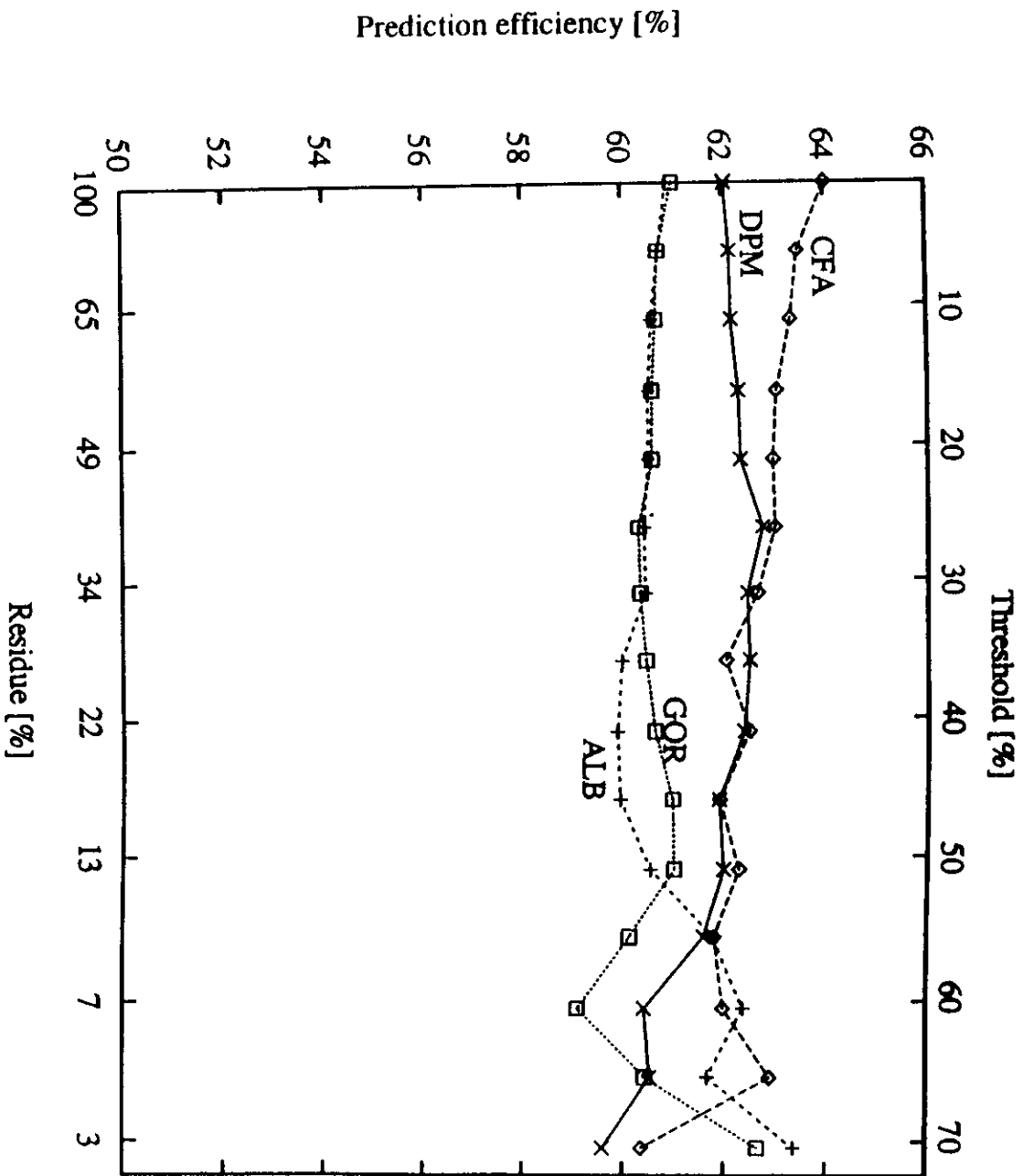


Fig. 1. The averaged efficiencies of the four different prediction methods (DPM, GOR, ALB and CFA) are shown versus the applied threshold for highly interacting residues. The threshold is indicated on the top side, while the number of remaining residues underneath is expressed in per cent. The applied threshold indicates the minimal number of relative long-range interactions to be made for a residue to consider it being a member of a stabilization centre.

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.

	A	C	D		V	W	Y
A	1.13	0.93	0.98				
C	1.02	0.79	1.05				
D	1.00	0.90	0.99				
V					0.97	0.97	0.95
W					1.04	1.02	1.09
Y					0.96	1.13	1.08

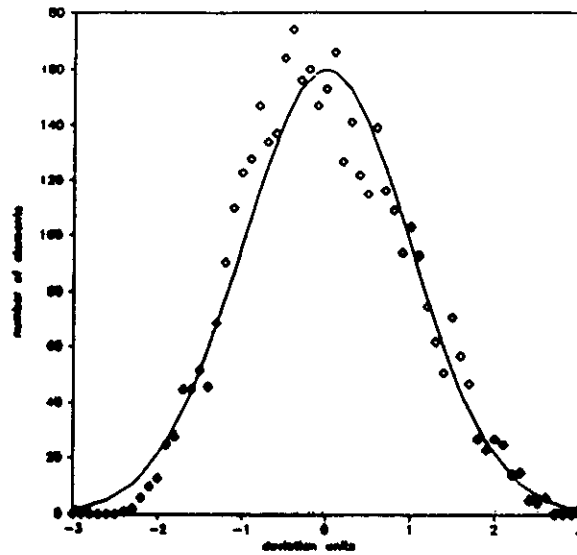


FIGURE 2
 Abundance of the various pair preference matrix elements in the 21st. . . 30th neighbour region. The solid line is the theoretical Gaussian distribution.

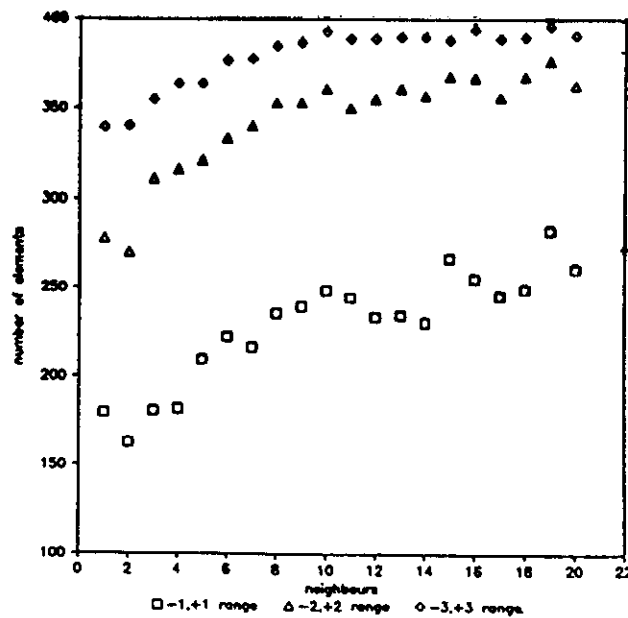


FIGURE 3
 Number of pair preference matrix elements between $\{-1, +1\}$, $\{-2, +2\}$ and $\{-3, +3\}$ deviation units versus the sequential distance. Theoretical values of random pair occurrence within the $\{-1, +1\}$, $\{-2, +2\}$ and $\{-3, +3\}$ deviation unit range are marked on the right ordinate.

Gábor E. Tusnády, Gábor Tusnády¹ and István Simon²

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

²To whom correspondence should be addressed

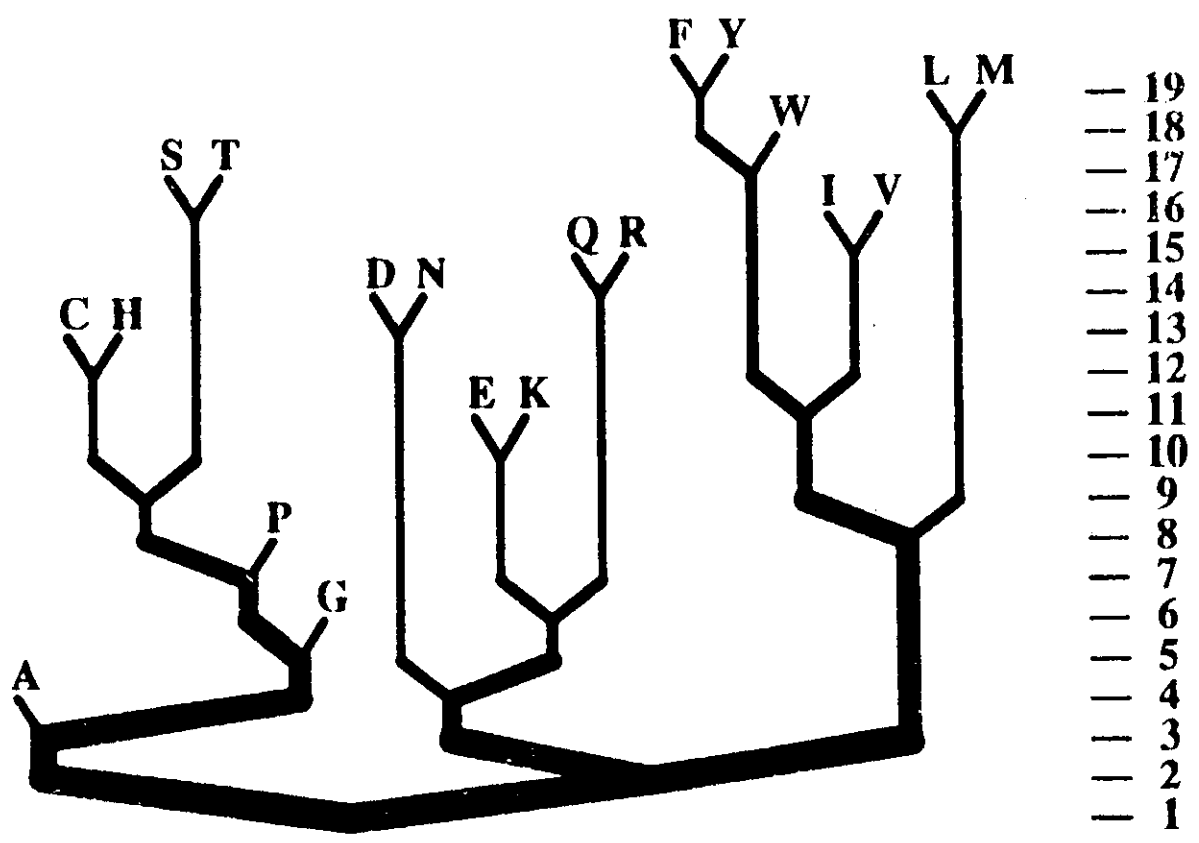
The discovery of the relationship between amino acids is important in terms of the replacement ability, as used in protein engineering homology studies, and gaining a better understanding of the roles which various properties of the residues play in the creation of a unique, stable, 3-D protein structure. 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

Juswidy, 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



COMMUNICATIONS

New Alignment Strategy for Transmembrane Proteins

M. Cserzö^{1,3}, J.-M. Bernassau², I. Simon³ and B. Maigret¹

¹*Laboratoire de Chimie Théorique
URA CNRS No. 510 Université de Nancy-1-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



Figure 11-1
 Electron micrograph of a preparation of plasma membranes from red blood cells. These membranes are seen "on edge," in cross section. [Courtesy of Dr. Vincent Marchesi.]

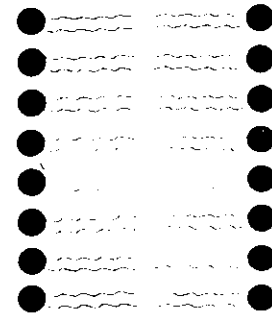


Figure 11-12
 Diagram of a section of a bilayer membrane formed from phospholipid molecules.

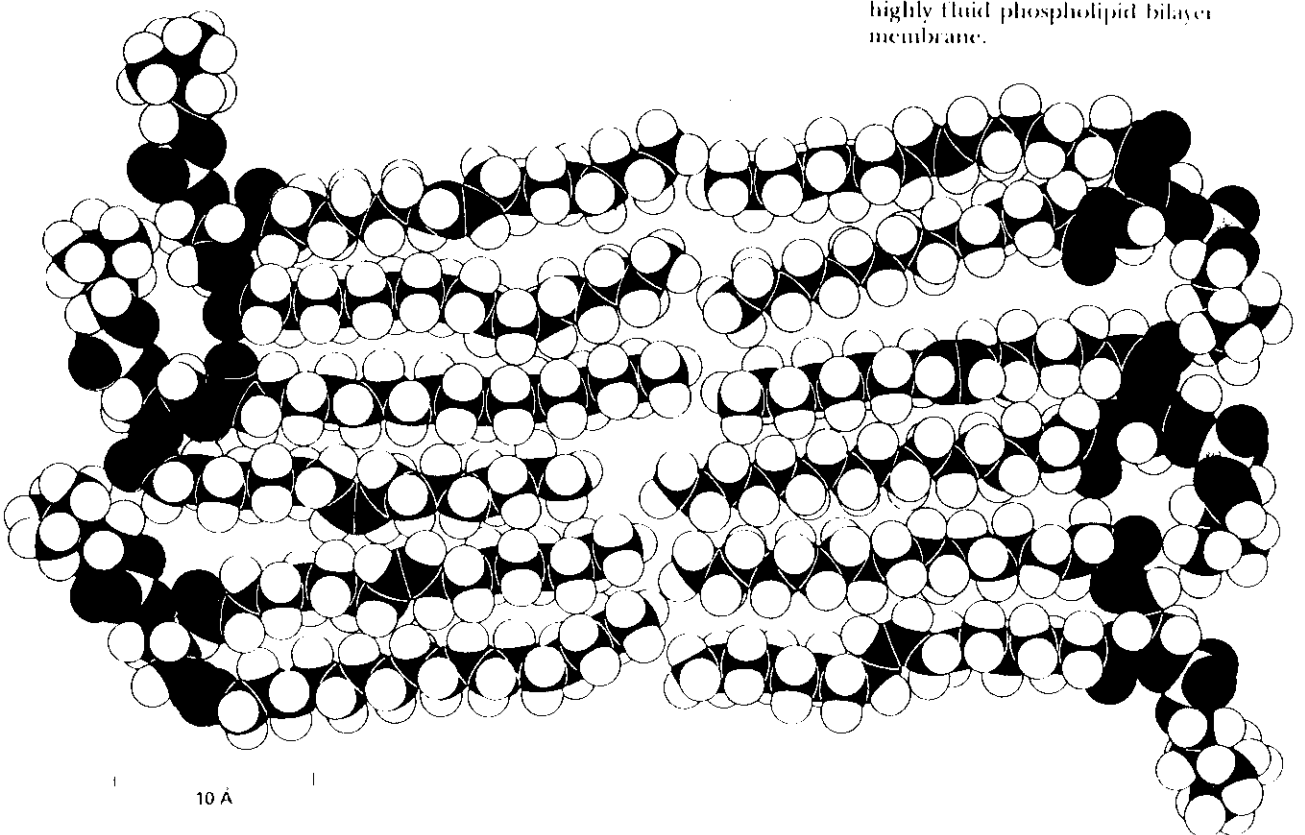


Figure 11-13
 Space-filling model of a section of a highly fluid phospholipid bilayer membrane.

10 Å

Polarity scale for identifying transmembrane helices

<i>Amino acid residue</i>	<i>Transfer free energy (kcal/mol)</i>	<i>Amino acid residue</i>	<i>Transfer free energy (kcal/mol)</i>
Phe	3.7	Ser	0.6
Met	3.4	Pro	-0.2
Ile	3.1	Tyr	-0.7
Leu	2.8	His	-3.0
Val	2.6	Gln	-4.1
Cys	2.0	Asn	-4.8
Trp	1.9	Glu	-8.2
Ala	1.6	Lys	-8.8
Thr	1.2	Asp	-9.2
Gly	1.0	Arg	-12.3

Note: The free energies are for the transfer of an amino acid residue in an α helix from the membrane interior (assumed to have a dielectric constant of 2) to water. After D.M. Engelman, T.A. Steitz, and A. Goldman. *Ann. Rev. Biophys. Biophys. Chem.* 15(1986):330.

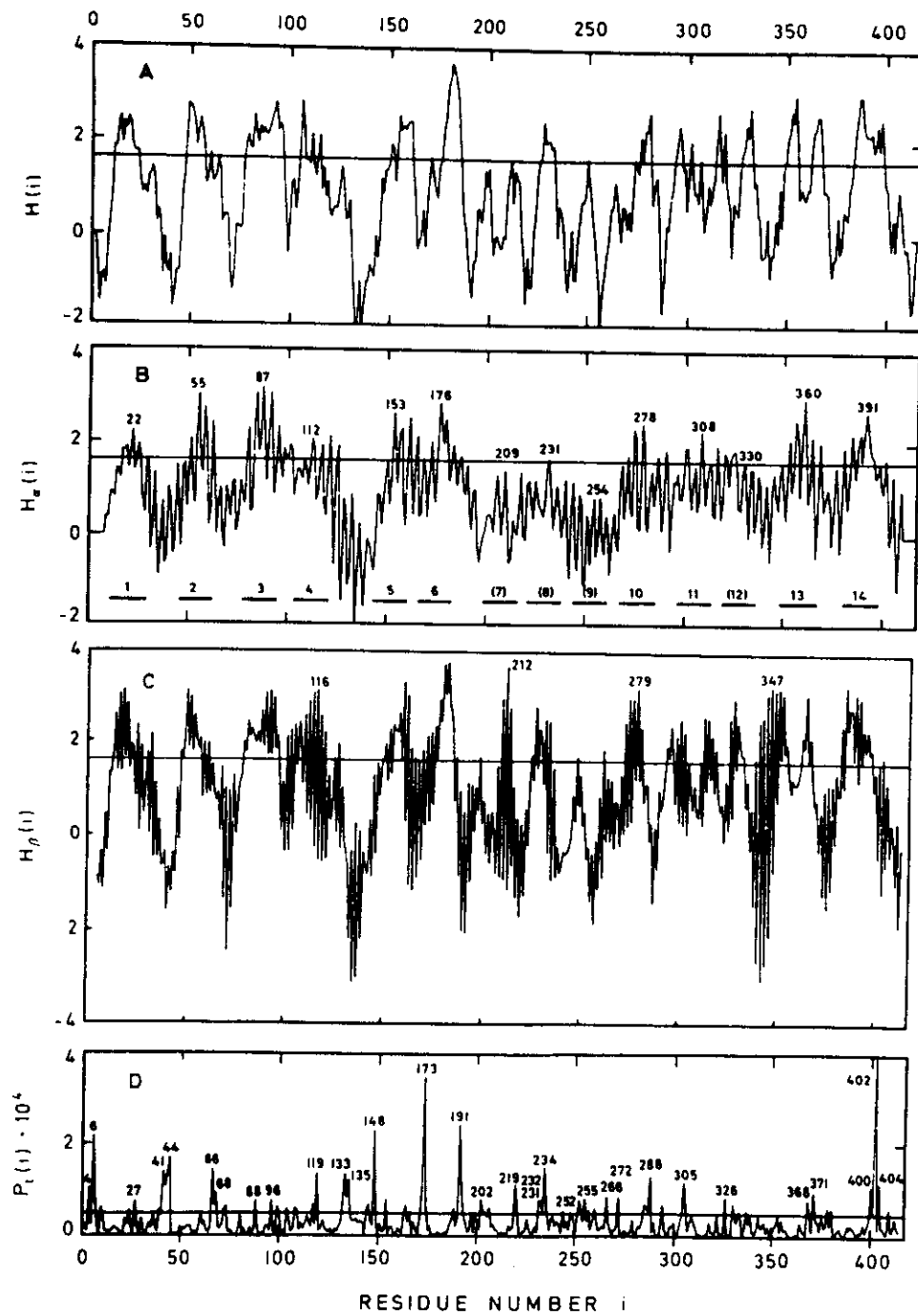


Figure 1. Structure prediction plots for the lactose permease: (A) hydrophobicity $H(i)$, (B) α -helix side hydrophobicity $H_{\alpha}(i)$, (C) β -strand side hydrophobicity, and (D) β -turn potential $P_t(i)$. Numbers in B specify the residues in the middle of the most hydrophobic sides; the bars specify the extension of the corresponding helices.

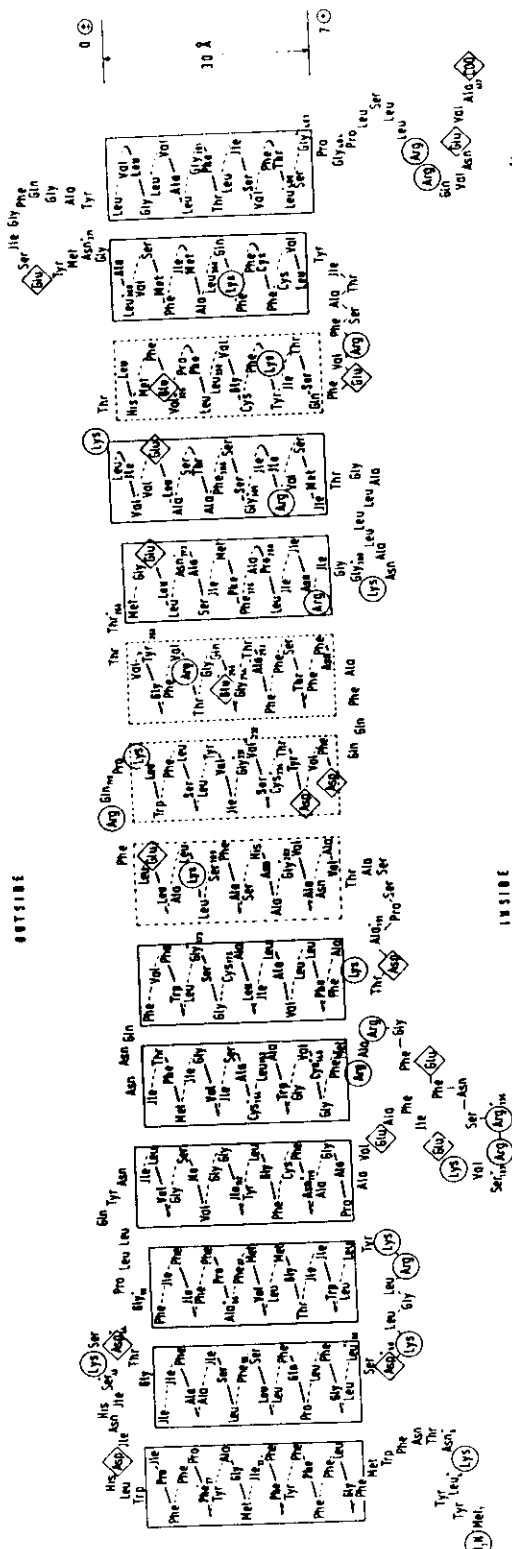


Figure 2. Folding model for the lactose permease. Solid rectangles represent predicted membrane-spanning α -helices, with boldface letters denoting the residues on the more hydrophilic sides. Dashed rectangles represent hydrophobic α -helix regions that are not predicted as membrane-spanning. Zigzags symbolize β -strands, and asterisks denote β -turns. Circles and squares indicate positively and negatively charged residues, respectively.

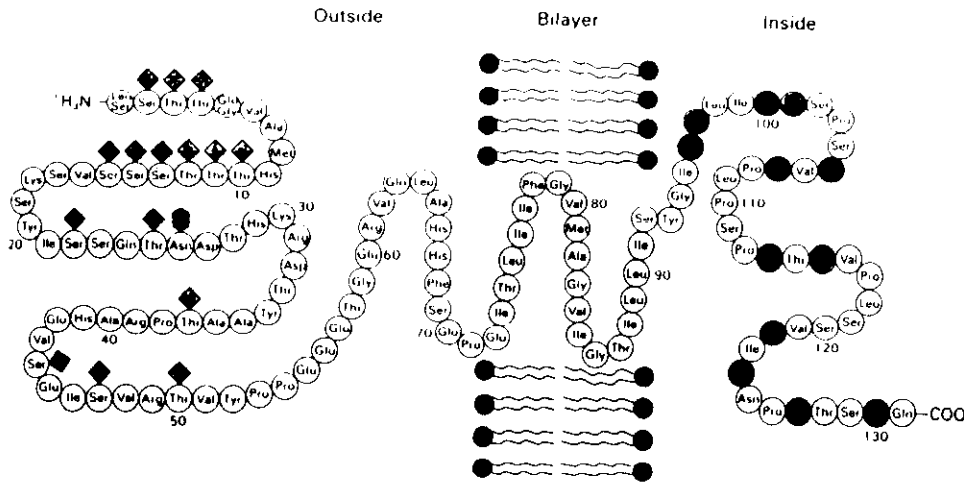


Figure 11-43
 Amino acid sequence and transmembrane disposition of glycoporphin A from the red-cell membrane. The fifteen *O*-linked carbohydrate units are shown in light green and the *N*-linked unit in dark green. The hydrophobic residues (yellow) buried in the bilayer form a transmembrane α helix. The carboxyl-terminal part of the molecule, located on the cytosolic side of the membrane, is rich in negatively charged (red) and positively charged (blue) residues. [Courtesy of Dr. Vincent Marchesi.]

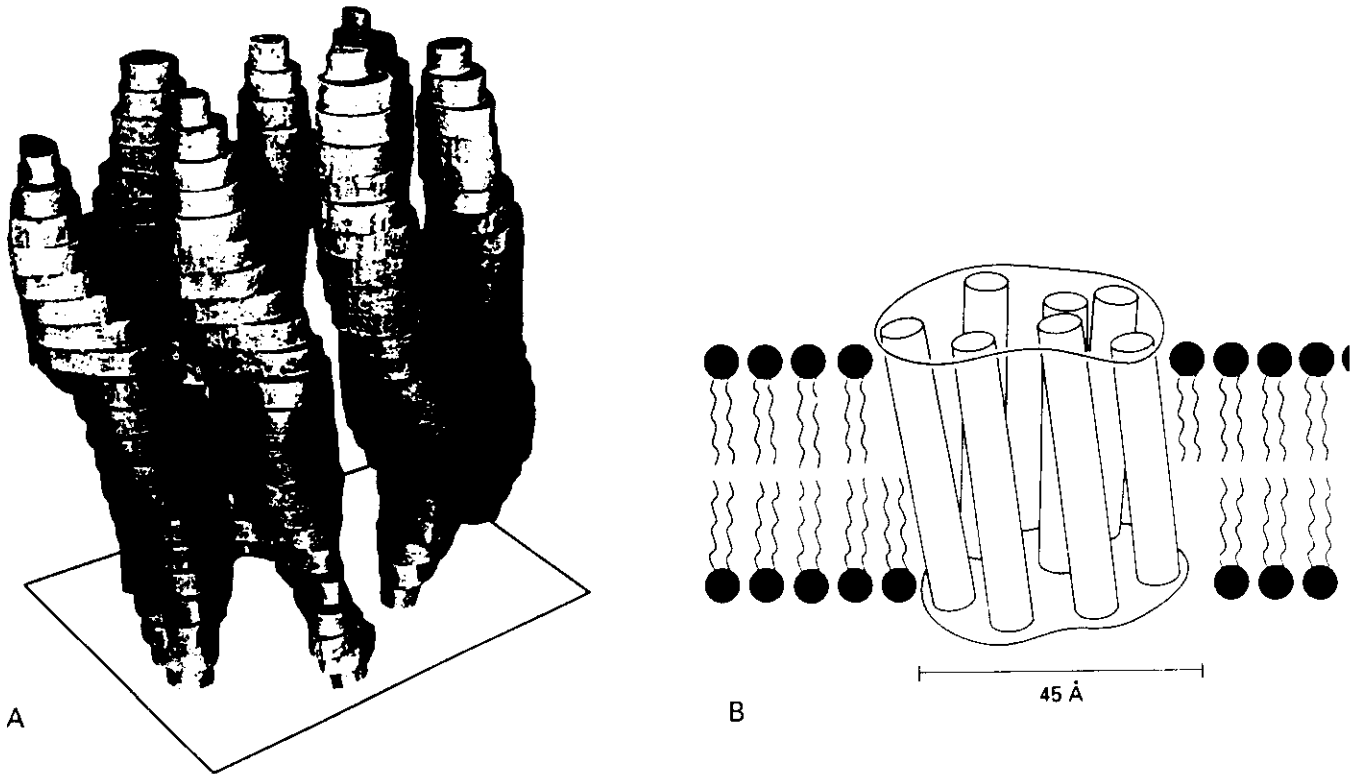
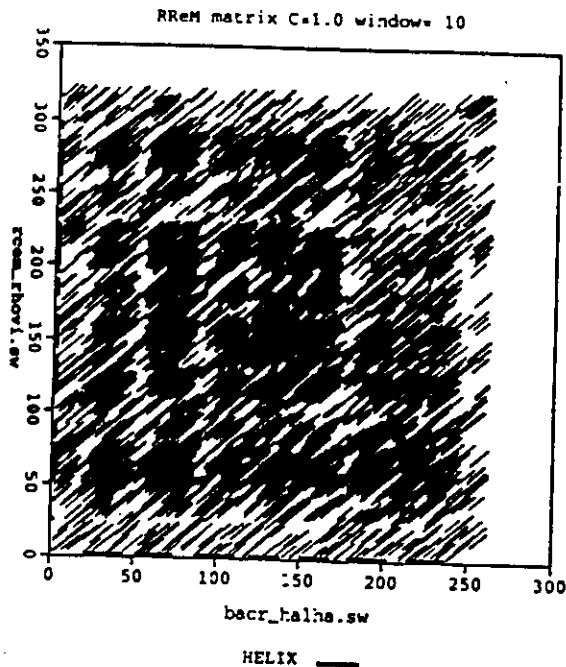
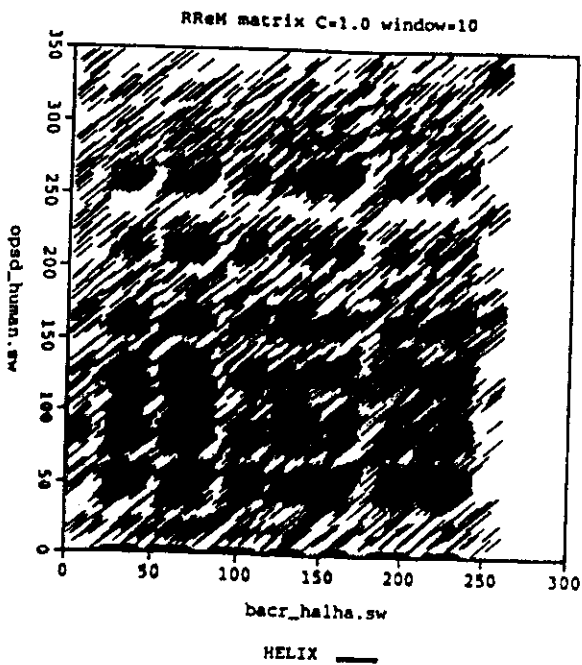


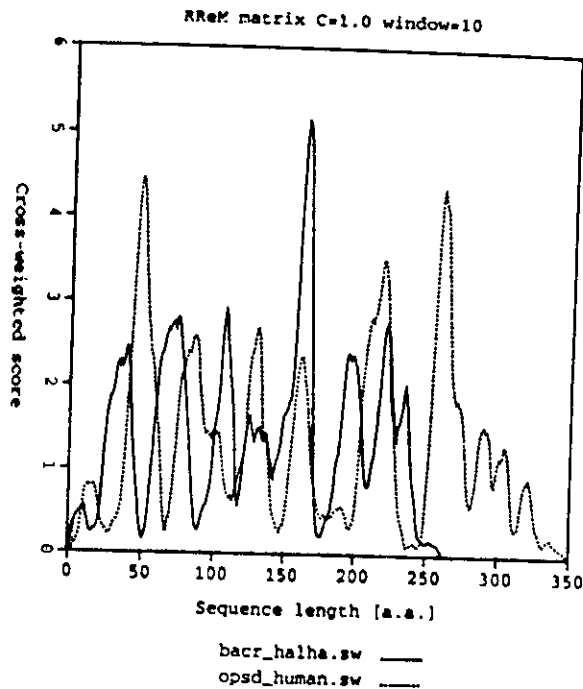
Figure 11-48
 (A) Model of bacteriorhodopsin constructed from a 7-Å three-dimensional map.
 (B) Interpretive diagram showing the arrangement of α -helical segments in the lipid bilayer. The connections between helices were not evident at this resolution. [Courtesy of Dr. Richard Henderson and Dr. Nigel Unwin.]



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



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.

X-weighted profiles were generated with:
 RReM matrix, Cutoff=1.0 Window=10
 Sequence alignment were generated with:
 Gap-penalty=1.0 Smoothing-tolerance= 3

```

seq1: m-----lel lptavegvsq aqi-----t --gr--pEWI WLALGTALMG
seq2: MNGTEGPNFY VPFSNATGVV RSPFEYPOYY LAE-PWQFSH LAAYMFLIV

seq1: LGTLTYFL-VK gmgve----- -dpdAKKFYA ITTLVPAIAF TMYLSMllgy
seq2: LGFPINFLTL YVTVQHKRLR TP----LNYI LLNLAVADLF MVLGGFTSTL

seq1: gltmvpgge QNPIYWA--- -RYADNLFTT PLLLLDLall vdadqg----
seq2: YTSLHGYPVF GPTCCNLEGF FATLGEAL WSLVLAIER YVVVCKPMSN

seq1: -----TIL ALVGADGIMI GTGLVCal-- ----- -t-kv-y---
seq2: FRFCENHAIM GVAFTVMAL ACAAPPLAGW SRYIPEGLQC SCGIDYYTLK

seq1: -----syrf vWAISTAAM LYILYVLFPO ftskkaerp eVAST-----
seq2: FEVNESFVI YM--FVVHFT IPHIIIFPCY GQLVPTVKE- -----AAAQO

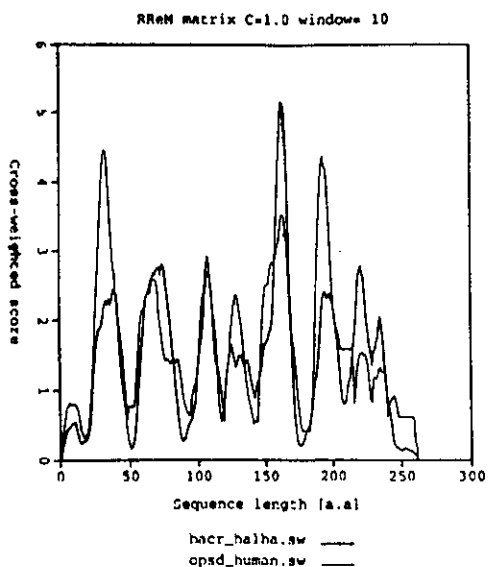
seq1: ----- --FKVLRNVT VVLSAYPVV Wli---gseg agivpln---
seq2: QESATTQKAE KEVTRMVIIM VIAFLICWVP YASVAFYI-- -----FTH

seq1: -----IETLL FMVLDVSAK- -VGFGLILlr s--r----- -aifgeseap
seq2: QCSNFGPIFM TIPAFFAKSA AIYNFVIY-- -IMNKQFRN CMLTTCGG-

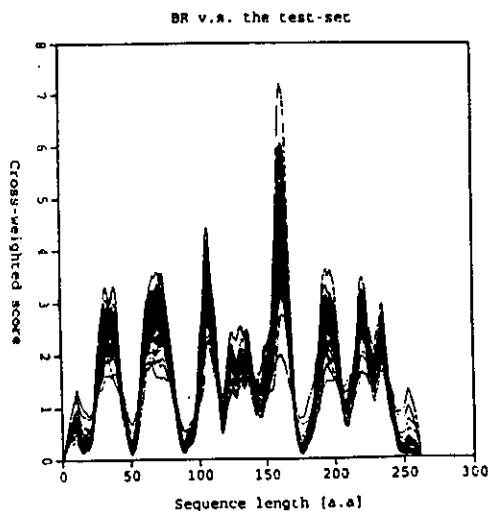
seq1: epsagdgaaa ---t----- -s----- ----d
seq2: ----- KNPLGDDEAS ATVSKTETSQ VAPA

```

Peak to peak alignment of BR versus HR transmembrane helices. For BR the residues falling into the helical regions are printed with capital letters. The residues proven to be involved in structure/function relationship are marked by *. Note that this alignment still contains the false gaps inserted into the helical regions. These should be removed manually by the user.



The crossweighted cumulative scores of BR and hR as the modified Needleman & Wunsch algorithm aligns them.



The cross weighted cumulative score profiles of BR obtained with the alignment against the 56 other proteins of the test-set. The positions of the peaks are conserved, the height of the peaks deviate considerably, but typically higher than 2.0.

Prediction of transmembrane α -helices in prokaryotic membrane proteins: the dense alignment surface method

Miklos Cserzö^{2,3}, Erik Wallin¹, Istvan Simon,
Gunnar von Heijne¹, Arne Elofsson¹

Institute of Enzymology, Biological Research Center Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary and ¹Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden
²Present address: University of Birmingham, School of Biochemistry, Edgbaston, Birmingham B15 2TT, UK

³To whom correspondence should be addressed

A new, simple method for predicting transmembrane segments in integral membrane proteins has been developed. It is based on low-stringency dot-plots of the query sequence against a collection of non-homologous membrane proteins using a previously derived scoring matrix [Cserzö *et al.*, 1994, *J. Mol. Biol.*, 243, 388-396]. This so-called dense alignment surface (DAS) method is shown to perform on par with earlier methods that require extra information in the form of multiple sequence alignments or the distribution of positively charged residues outside the transmembrane segments, and thus improves prediction abilities when only single-sequence information is available or for classes of membrane proteins that do not follow the 'positive inside' rule.

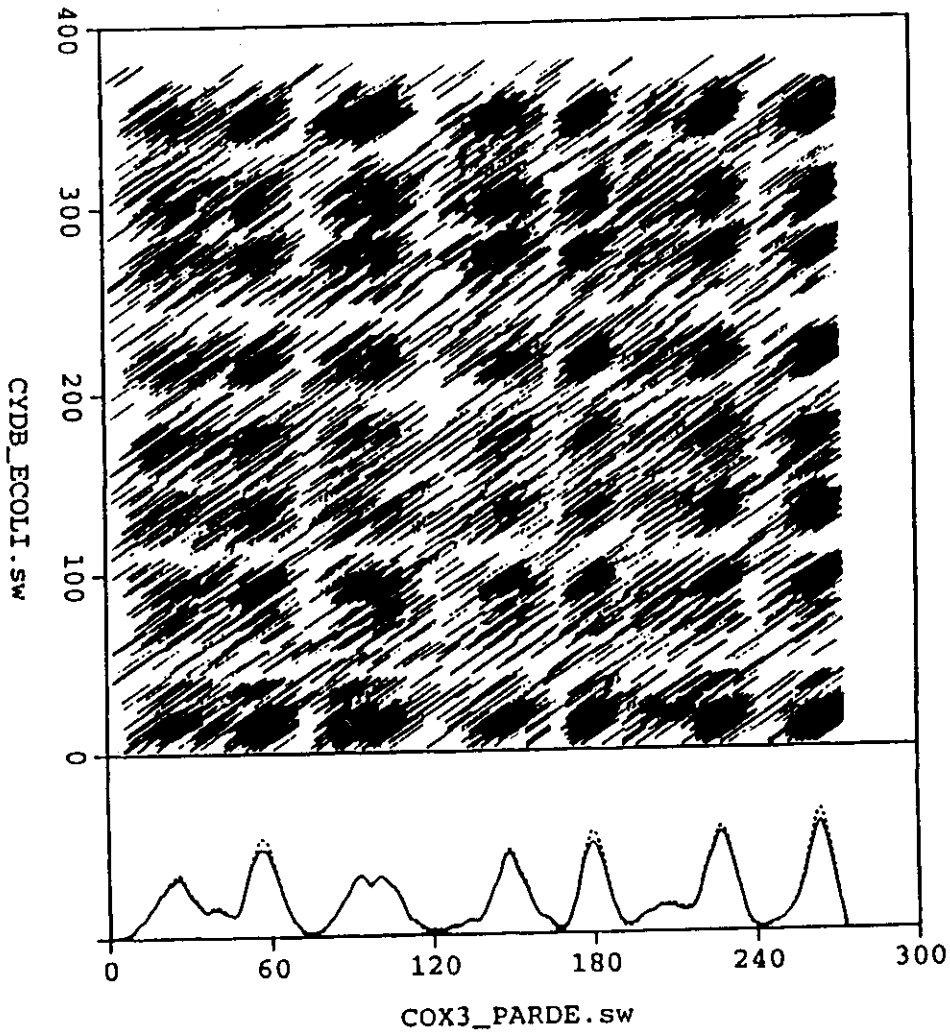


Fig. 1. DAS plot of two arbitrarily chosen proteins (COX3_PARDE versus CYDB_ECOLI). The cross weighted cumulative score profile (dotted line) and the global DAS profile (continuous line) calculated as the average of the cumulative score profiles obtained for comparisons with the other 43 proteins in the test set are also shown for COX3_PARDE. COX3_PARDE has seven and CYDB_ECOLI has eight transmembrane segments.

Hypothesis

Different sequence environments of cysteines and half cystines in proteins

Application to predict disulfide forming residues

András Fiser, Miklós Cserző, 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

Ratio of the normalized abundances of various residues in a given position in the vicinity of half cystines and 'free' cysteines

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	Cys	1	2	3	4	5	6	7	8	9	10
A	1.50	0.73	1.44	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	1.42	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	1.55	
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	0.80	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	0.95	0.72	1.18	1.03	0.87	0.59	0.99	0.89	0.53	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	0.73	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	1.00	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	1.86	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	1.65	1.28	2.05	1.55	1.11	1.03	1.37	1.25	2.73	1.81	2.21	0.89	1.81	1.60	2.22	1.63	1.36	2.50	1.59	1.03	

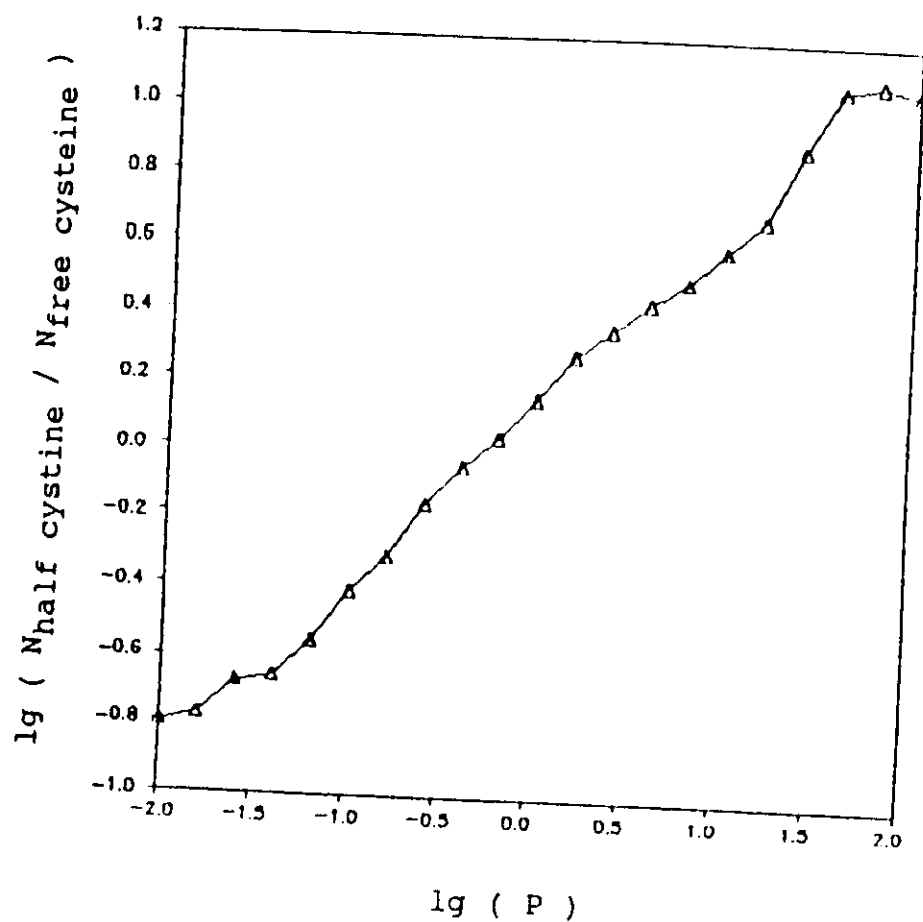


Fig. 1. The logarithm of the ratio of cysteine and 'free' cysteine abundances versus the logarithm of the disulfide forming potential (P).

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. © Munksgaard 1994.

Prediction data^a

Amino acid	Total number of occurrences	Prediction power (%)	No. of residues with extreme potential	Prediction power (%)
Ala	1447	60.61	188 (13.0%)	70.89
Cys	355	61.69	170 (48.0%)	66.08
Asp	986	60.85	185 (18.8%)	76.34
Glu	1148	50.44	190 (16.6%)	66.31
Phe	679	56.70	146 (21.5%)	68.49
Gly	1534	57.69	166 (10.8%)	63.64
His	369	63.14	153 (41.5%)	67.97
Ile	899	59.84	165 (18.4%)	69.23
Lys	1007	59.29	187 (18.6%)	67.91
Leu	1410	57.80	171 (12.1%)	72.51
Met	322	57.14	158 (49.1%)	55.70
Asn	703	56.47	102 (14.5%)	62.50
Pro	776	59.41	157 (20.2%)	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	131 (13.0%)	61.83
Val	1217	58.26	173 (14.2%)	68.97
Trp	240	54.58	127 (53.0%)	55.91
Tyr	590	54.58	149 (25.3%)	57.72

^a The first column contains the three-letter codes of the amino acids. The second gives the total number of their occurrences in the analysed database. The third column contains the 'jack-knife' test results of the prediction of long-range interaction for all residues. Column 4 shows the number of residues with extreme interacting potential values (more than 10 or less than 0.1) and the percentage of these in the total number of residues (in parentheses). The last column shows the prediction success for the residues of column 4.

Stabilization Centers in Proteins: Identification, Characterization and Predictions

Zsuzsanna Dosztányi, András Fiser and István Simon*

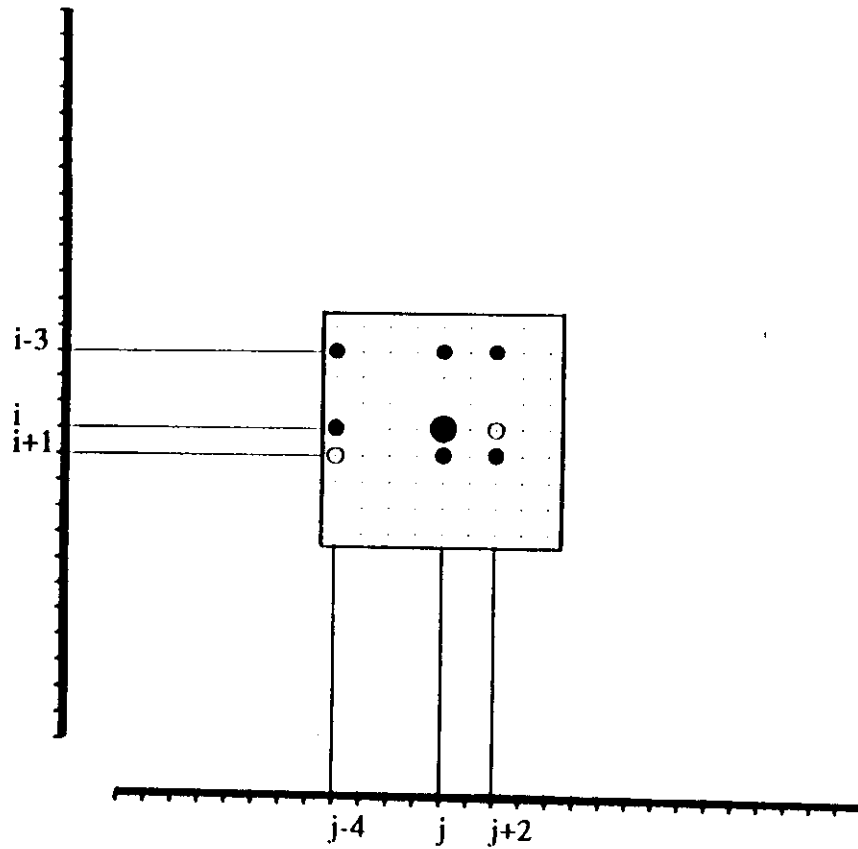
*Institute of Enzymology
Biological Research Center
Hungarian Academy of
Sciences, H-1518 Budapest
PO Box 7, Hungary*

Methods are presented to locate residues, stabilization center elements, which are expected to stabilize protein structures by preventing their decay with their cooperative long range interactions. Artificial neural network-based algorithms were developed to predict these residues from the primary structure of single proteins and from the amino acid sequences of homologous proteins. The prediction accuracy using only single sequence information is 65%, but the incorporation of evolutionary information in the form of multiple alignments and conservation scores raises the efficiency by 3%. The composition, relative accessibility, number and type of interactions, conservation and the X-ray thermal factor of the identified stabilization center residues are different, not only from the whole data set but from the rest of the long range interacting residues as well. The most frequent stabilization center residues are usually found at buried positions and have a hydrophobic or aromatic side-chain, but some polar or charged residues also play an important role in the stabilization. The stabilization centers show significant difference in the composition and in the type of linked secondary structural elements compared with the rest of the residues. The performed structural and sequential conservation analysis showed the higher conservation of stabilization centers over protein families. The relation of the proposed stabilization centers to folding nuclei is also discussed.

© 1997 Academic Press Limited

Keywords: long range interaction; stabilization center; prediction; neural network; protein stability

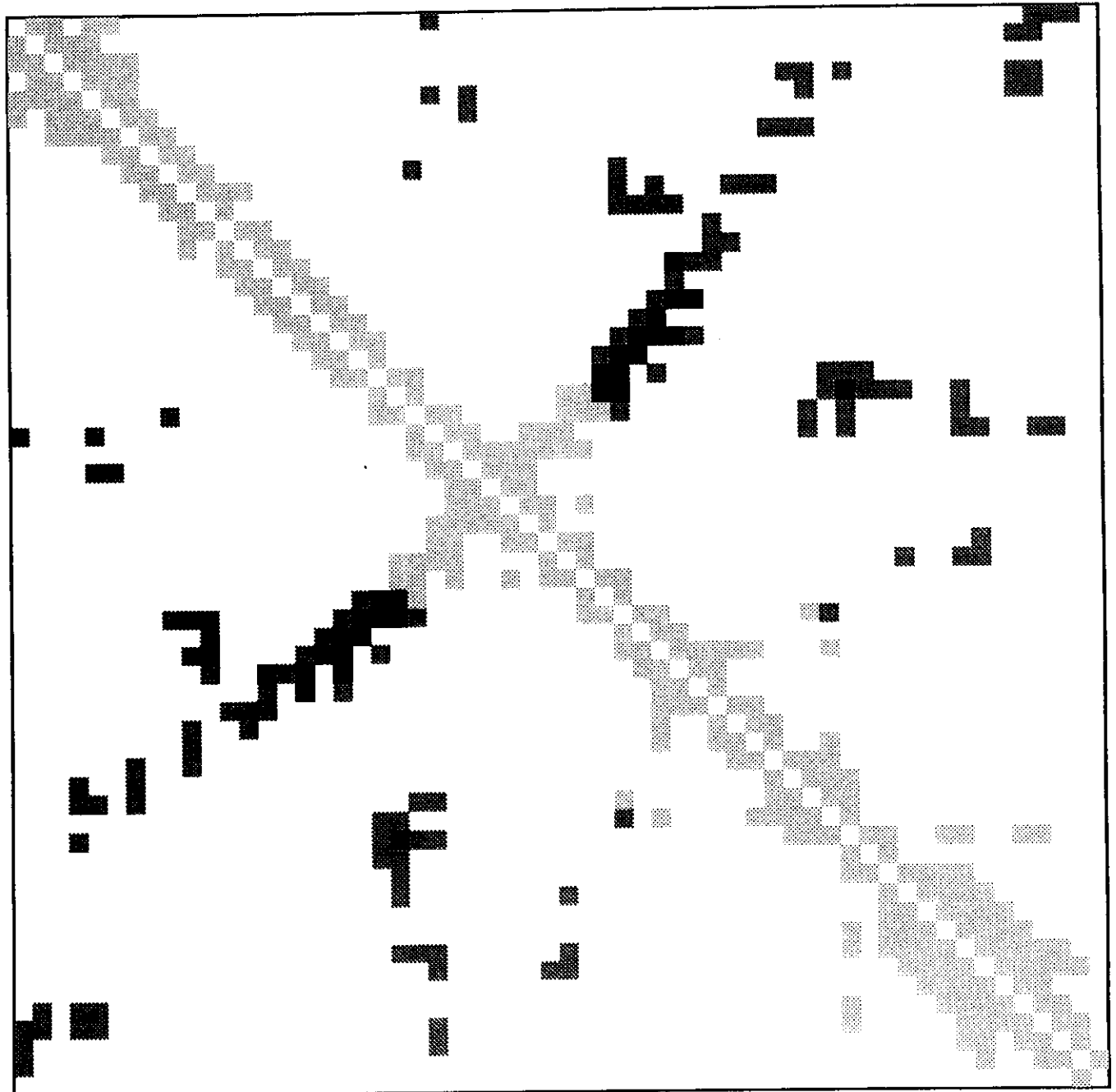
*Corresponding author



Definition of a *SC* element on a hypothetical contact map detail: two residues (i and j are considered as elements of an *SC*'s, if they are in long-range contact, and it is possible to select residues from both flanking tetrapeptides of both residues, that make at least 7 contacts between these two triplets out of the possible 9 ones (filled circles represent contacts between the central residues and the two selected ones, empty circles represent missing contacts between residues). The larger filled circle is the *SC* element. In the given example $i+1$ and $i-3$ and $j-4$ and $j+2$ are the selected residues in the flanking tetrapeptides.

5pti

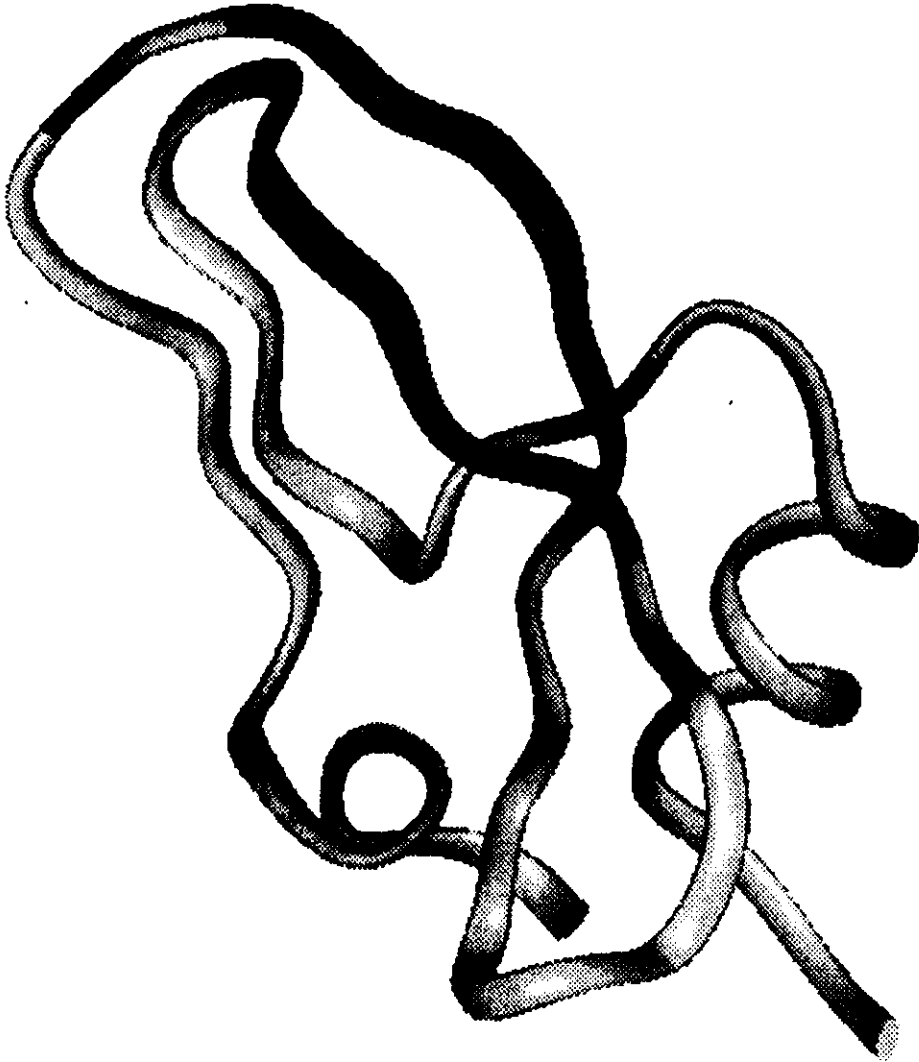
0

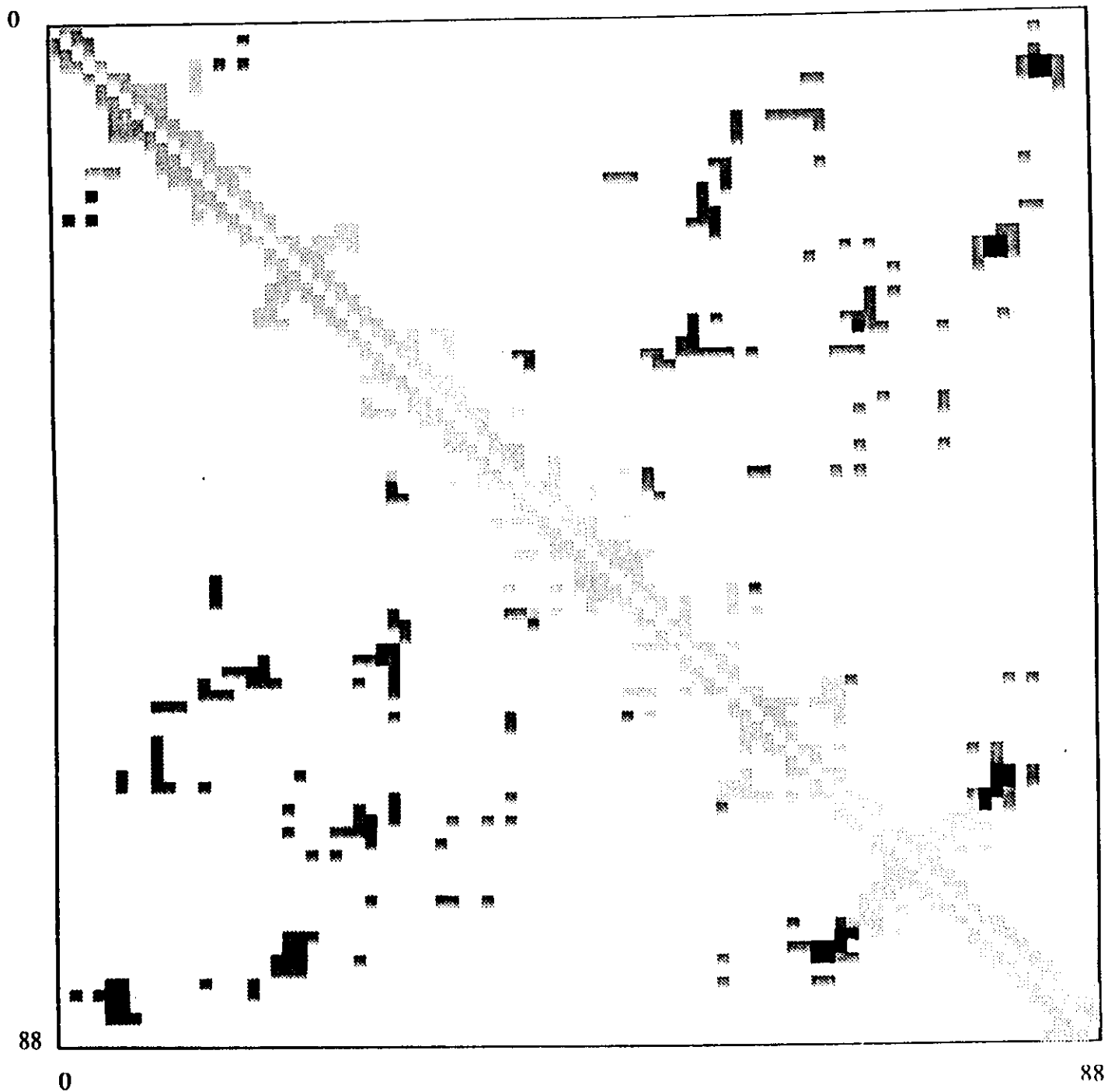


58

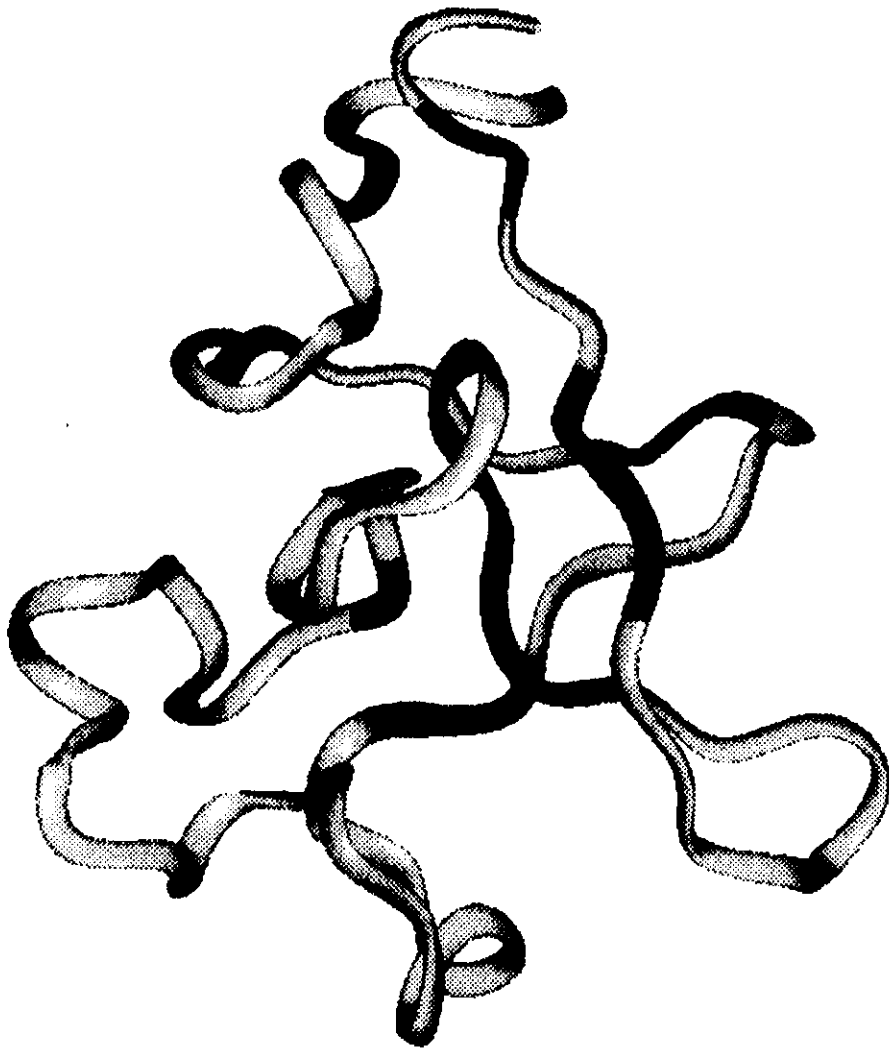
0

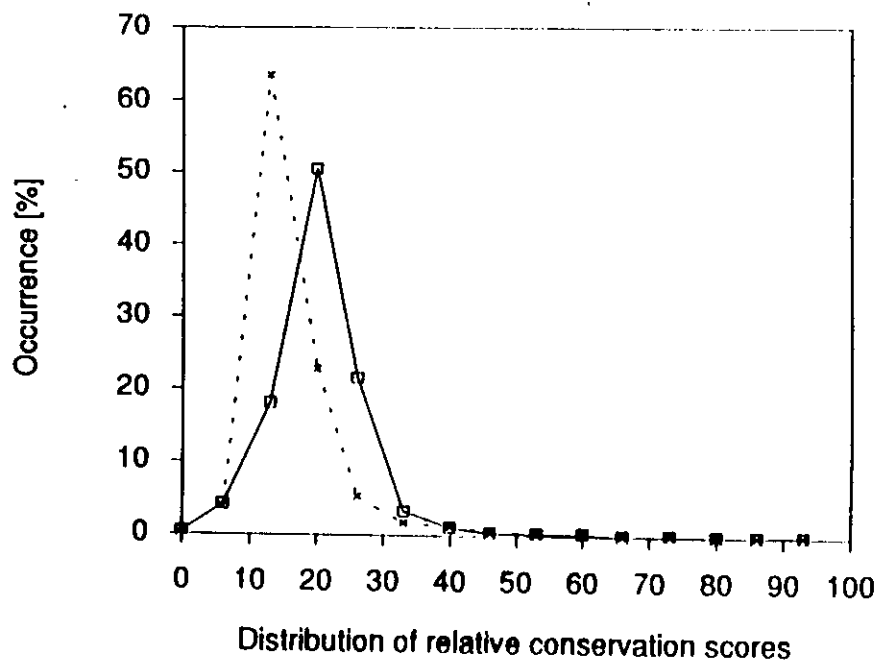
58



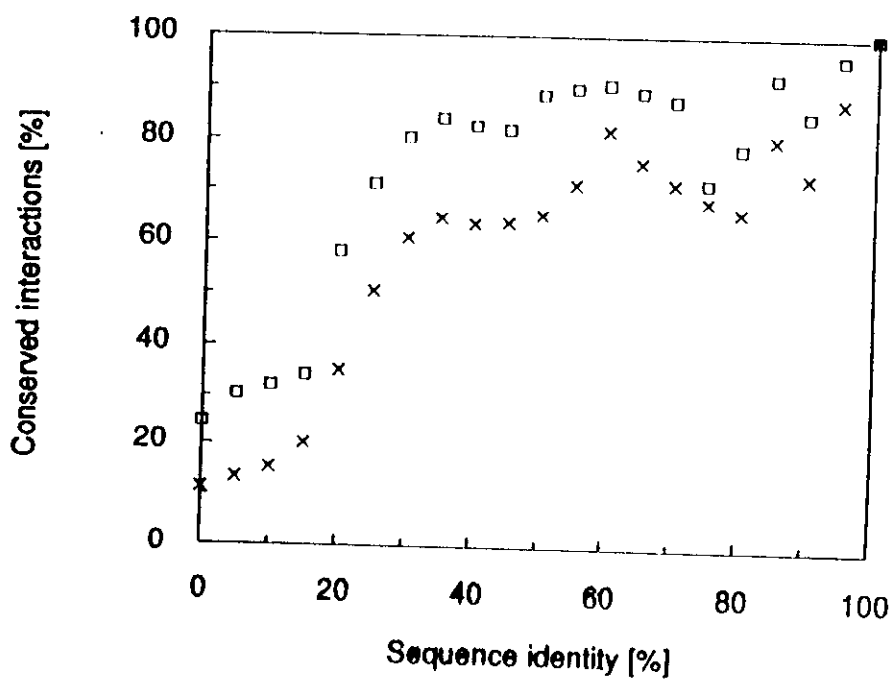


a. and b. The contact map for 1tpk (tissue plasminogen activator kringle domain) and 5pli (BPTI), respectively. Black boxes represent contacts between stabilization center elements, medium grey boxes stand for the rest of residues involved in long range contacts, light grey boxes near the diagonal represent short range interactions. c. and d. The ribbon structure of 1tpk and 5PPI. Dark color indicates residues in *SCS*.

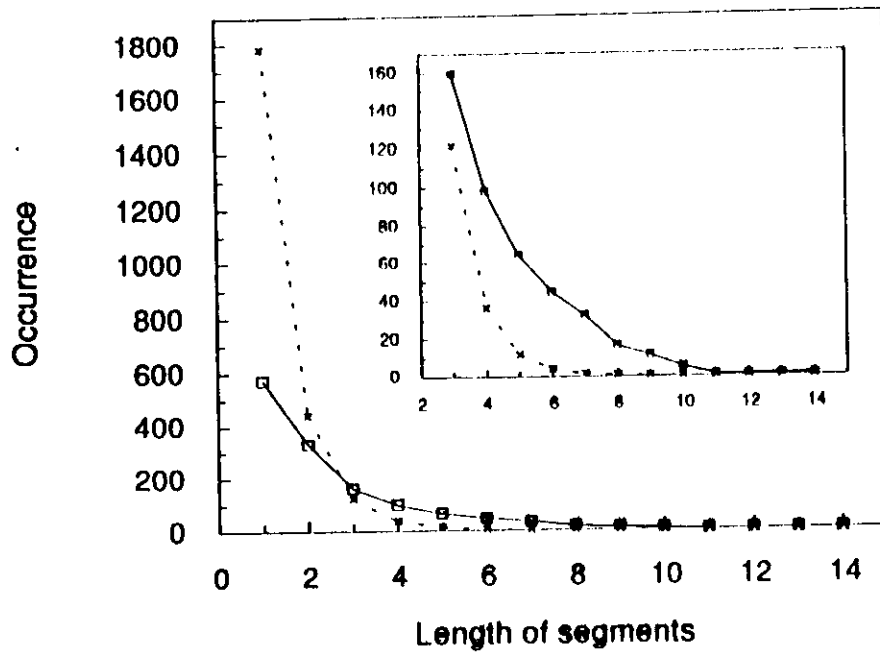




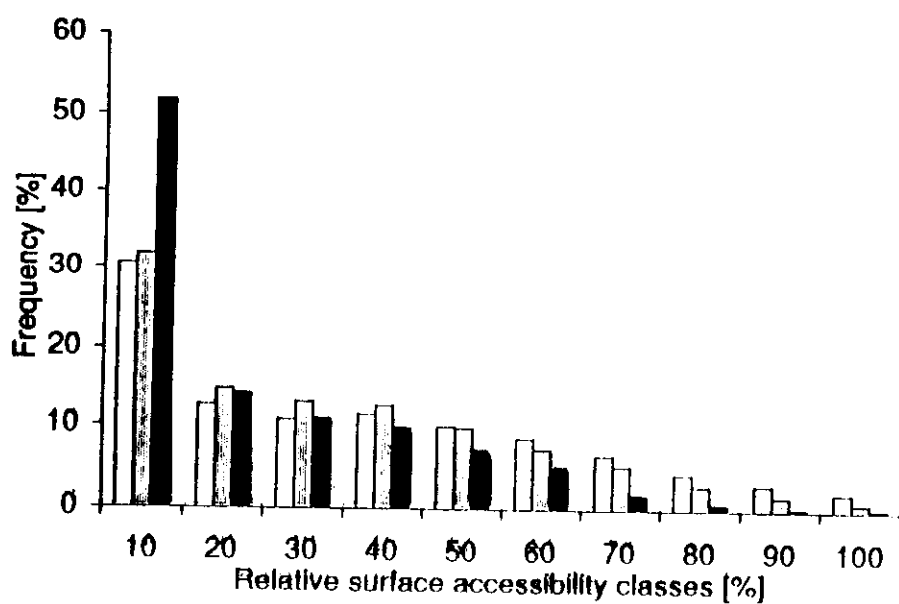
Plots of the average relative sequence conservation over protein families for *SC* residues (indicated with \square -es) and for the rest of the data set (indicated by x).



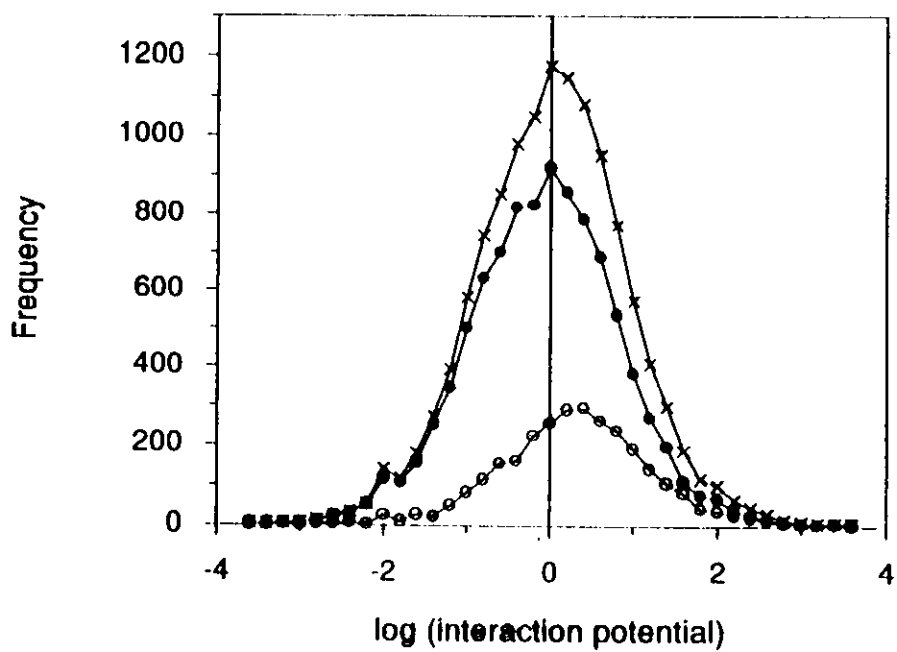
The percent of average conserved long range contact versus the sequence identity over protein families. SC elements are indicated by □ while the rest of residues making long range contacts, indicated by x's.



Length of observed segments in the sequence for *SC*'s (continuous line with \square -es) and for the generated reference set (dotted line with x -s). The insert shows the same figure on a 10-fold enlarged ordinate scale to accentuate the most relevant part of the plot.



The histogram shows the distributions of three class of residues (all the residues, residues in long range contact but not in stabilization centers and *SC* elements, indicated by white, grey and black bars, respectively) in the relative surface accessibility region 0-100%.



Plots of calculated long range interacting potentials: *x*-es represent the whole database, filled circles belongs to the long range interacting residues, while empty circles to the *SC* residues.

AA	Composition of residues					$\frac{(SC-ALL)}{\delta_{SC}}$	$\frac{(SC-LRI)}{\delta_{SC}}$
	ALL %	LRI % δ	SC % δ				
A	8.34	8.28 0.20	6.57 0.41	-4.3	-4.2		
C	2.31	2.59 0.11	3.03 0.22	3.3	2.0		
D	5.83	5.44 0.17	3.64 0.36	-6.1	-5.0		
E	5.85	5.24 0.17	3.94 0.34	-5.6	-3.8		
F	3.73	4.51 0.14	4.31 0.28	2.1	-0.7		
G	8.88	7.37 0.20	7.61 0.42	-3.0	0.6		
H	2.38	2.48 0.11	3.09 0.22	3.2	2.8		
I	5.04	5.38 0.16	7.52 0.33	7.5	6.5		
K	6.30	5.75 0.18	4.86 0.36	-4.0	-2.5		
L	7.87	9.24 0.20	9.51 0.40	4.1	0.7		
M	2.08	2.41 0.11	2.38 0.21	1.4	-0.1		
N	4.46	4.51 0.16	3.42 0.32	-3.2	-3.4		
P	4.56	4.67 0.15	3.45 0.31	-3.6	-3.9		
Q	3.79	3.68 0.14	2.81 0.30	-3.3	-2.9		
R	4.14	4.36 0.15	4.25 0.30	0.4	-0.4		
S	6.53	5.90 0.18	5.81 0.38	-1.9	-0.2		
T	5.80	5.24 0.17	6.36 0.35	1.6	3.2		
V	7.10	7.24 0.19	10.70 0.38	9.5	9.1		
W	1.32	1.58 0.08	1.59 0.17	1.6	0.1		
Y	3.68	4.12 0.14	5.14 0.28	5.2	3.6		
total	14 139=100%	7986=100%	3271=100%				

Table II. The first column shows the one letter code of amino acids (AA), the second column shows the composition of residues on the whole dataset (ALL), the third and fourth column show the residue composition with the standard deviations for those residues which participate in long range interactions (LRI) but not members of stabilization centers and for those residues which are in stabilization centers (SC), respectively. In the last two columns the observed composition differences are shown between the stabilization center residues and the whole dataset and between the stabilization center residues and the long range interacting residues, respectively, in deviation units.

A	1.03																											
C	0.37	2.43																										
D	0.85	0.67	0.84																									
E	0.86	1.09	0.52	0.48																								
F	1.10	0.99	0.23	0.87	1.57																							
G	1.15	1.29	1.27	1.05	0.79	0.90																						
H	0.46	1.59	0.66	1.07	0.83	1.03	1.17																					
I	1.54	0.57	0.61	0.94	1.02	0.81	0.80	1.57																				
K	0.70	1.51	1.26	2.03	0.88	0.85	0.74	1.07	0.31																			
L	1.04	0.77	1.07	0.69	0.99	0.89	0.88	1.37	0.52	1.67																		
M	0.83	0.77	0.86	0.79	1.79	1.63	0.50	0.83	0.64	0.82	0.65																	
N	0.41	0.53	1.77	0.41	0.74	1.41	1.04	0.43	0.88	0.56	1.13	1.55																
P	1.06	0.89	0.15	0.95	1.11	0.92	1.22	0.29	0.77	0.79	0.90	1.55	0.93															
Q	0.70	1.31	1.99	0.67	1.22	0.95	0.85	0.88	0.95	0.55	0.28	0.76	0.38	1.41														
R	1.39	0.29	1.44	1.66	1.11	0.52	1.98	0.52	0.72	0.37	0.55	1.26	1.01	1.40	1.03													
S	1.31	1.05	0.97	0.97	0.59	1.09	1.45	0.59	0.85	0.90	1.21	1.11	1.39	1.25	1.50	1.21												
T	1.05	0.67	1.11	0.81	1.20	1.14	1.03	1.38	0.65	1.03	1.09	1.34	0.75	0.82	1.22	0.94	0.72											
V	0.89	0.80	0.52	1.23	1.15	1.09	0.78	1.24	1.32	1.07	1.09	0.75	0.65	1.04	1.06	0.95	0.78	1.48										
W	0.71	1.93	--	0.59	0.81	1.07	1.89	1.86	1.44	1.10	0.98	2.70	--	1.24	1.92	1.21	0.55	0.43	--									
Y	0.93	0.60	1.39	1.37	1.00	1.23	1.52	1.01	1.49	1.02	1.21	1.46	1.78	1.03	1.10	1.43	0.84	0.94	0.91	0.84								
A		C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y								

Table III. The half matrix represents relative frequencies of long range interactions in stabilization centers normalized by the product of the relative occurrences of the two amino acids for a given pair.

AA	LRI	SC
A	13.2	27.0
C	21.5	32.2
D	18.0	30.5
E	17.3	36.0
F	39.9	61.5
G	12.1	24.2
H	26.1	49.1
I	26.0	44.7
K	19.0	38.6
L	23.1	41.2
M	23.4	41.2
N	19.8	34.2
P	17.5	30.1
Q	20.9	39.1
R	31.7	49.9
S	15.5	31.2
T	18.3	32.9
V	22.0	38.0
W	52.8	80.8
Y	39.9	58.1

Table IV. The first column shows the one letter code of amino acids (AA), the second column shows the average number of long range atom contacts made by the residues which participate in long range interactions but not members of stabilization centers (LRI), the third column belongs to residues in stabilization centers (SC).

		Linked structures				Composition
		Helix	Sheet	Turn	Coil	
LRI (20190 links)	Helix	12.12%				33.05%
	Sheet	9.11%	17.02%			14.10%
	Turn	4.17%	3.41%	0.88%		12.43%
	Coil	13.33%	18.48%	7.23%	14.04 %	40.43%
SC (2702 links)	Helix	5.44%				13.27%
	Sheet	1.44%	48.08%			51.33%
	Turn	1.55%	1.44%	0.56%		5.23%
	Coil	5.92%	16.58%	4.52%	14.47%	30.17%

Table V. The table shows the normalized frequency of the observed number of links among the secondary structural elements in the case of those long range interactions which are not members of *SC*-s (LRI) and in the case of stabilization centers (SC). The right-hand column shows the distribution of residues among the secondary structural elements.

Units	Accuracy of prediction	
	on single sequences	on multiple alignments
0	65%	68%
1	64%	57%
2	64%	68%
3	65%	67%
4	65%	68%
5	65%	67%
6	65%	67%
7	65%	68%
8	65%	68%
9	65%	67%

Table VI. The first column shows the number of applied hidden units in the neural network. The second column shows the prediction efficiencies using only single sequence information. In the third column the best achieved prediction efficiencies are shown using multiple sequence alignment.

'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 1978¹, 1984², and 1986³, 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 structures⁶, 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	4.29	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

^aDerived 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.

of multidomain proteins⁷, 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 prediction⁸⁻¹⁰

References

- 1 Vonderviszt, F., Mátrai, 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 Cserző, 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 Vonderviszt, 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. CSERZŐ

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

Table: Amino acid composition of proteins derived from data bases which were available in 1978 (I), in Jan. 1984 (II), and in Dec. 1986 (III) and from ORF proteins appearing in the 1986 data base (IV), in %.

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	4.29	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

Interresidue Interactions in Protein Classes

Z. Gugolya,¹ Zs. Dosztányi,² and I. Simon^{2*}

¹Department of Physics, University of Veszprém, Veszprém, Hungary

²Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

ABSTRACT The free energy difference between folded and unfolded state is about the same for most proteins and it is not more than the energy of a few noncovalent interactions. In addition to the numerous noncovalent interactions, some proteins 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 interactions per residue is almost the same for all protein. Also the number of long-range interactions per residue is the same in all proteins. Proteins with S—S bond(s) (extracellular proteins) have more medium-range and fewer 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 noncovalent 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. *Proteins* 27:360–366, 1997. © 1997 Wiley-Liss, Inc.

TABLE 1 List of proteins*

PDB name	number of	protein	number of	number of interactions			
	SS bonds	class [†]	residues	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

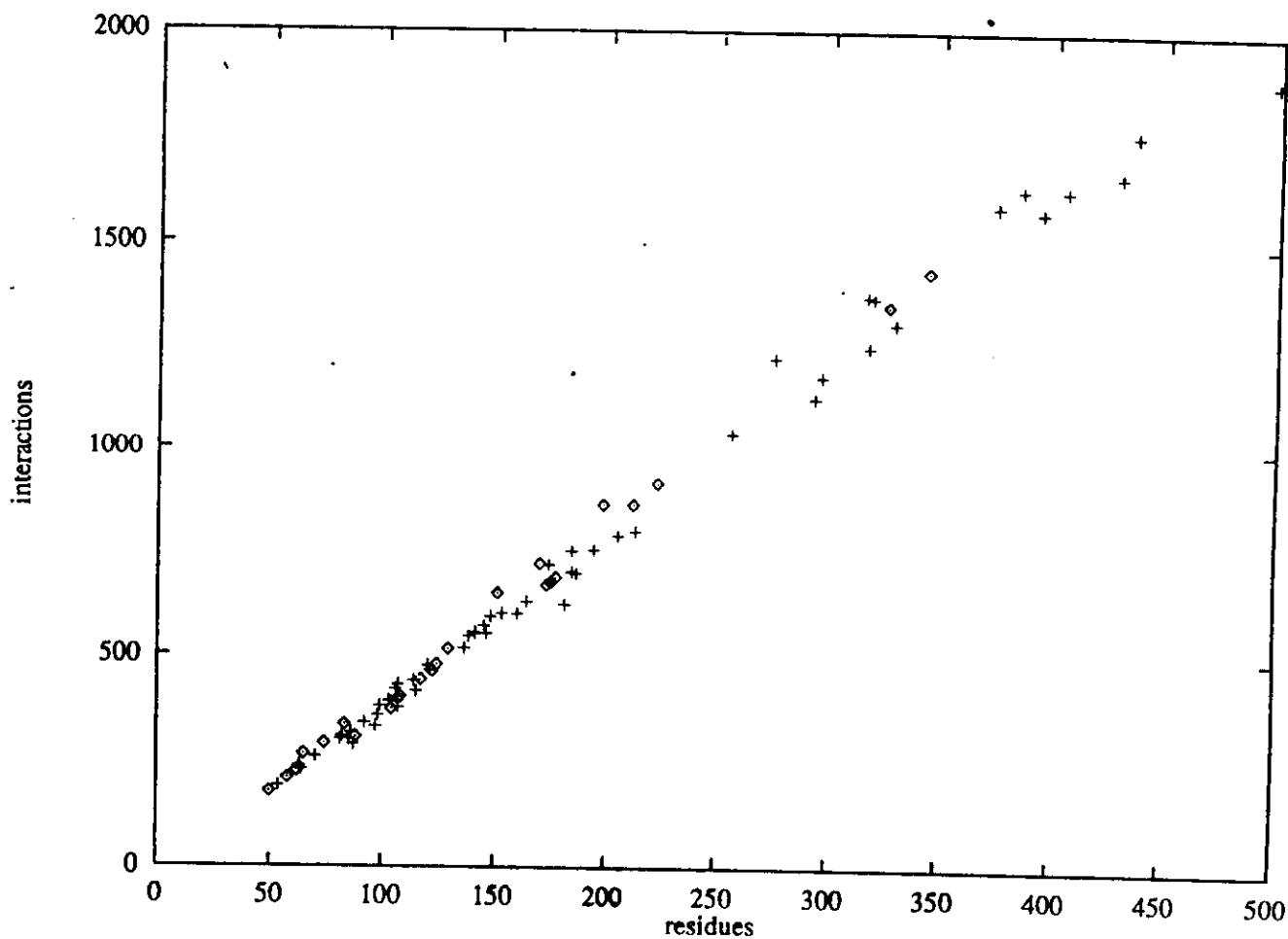


FIGURE 1 The total number of interactions as a function of the number of residues for proteins with disulfide bonds (\diamond), and without disulfide bonds (+).

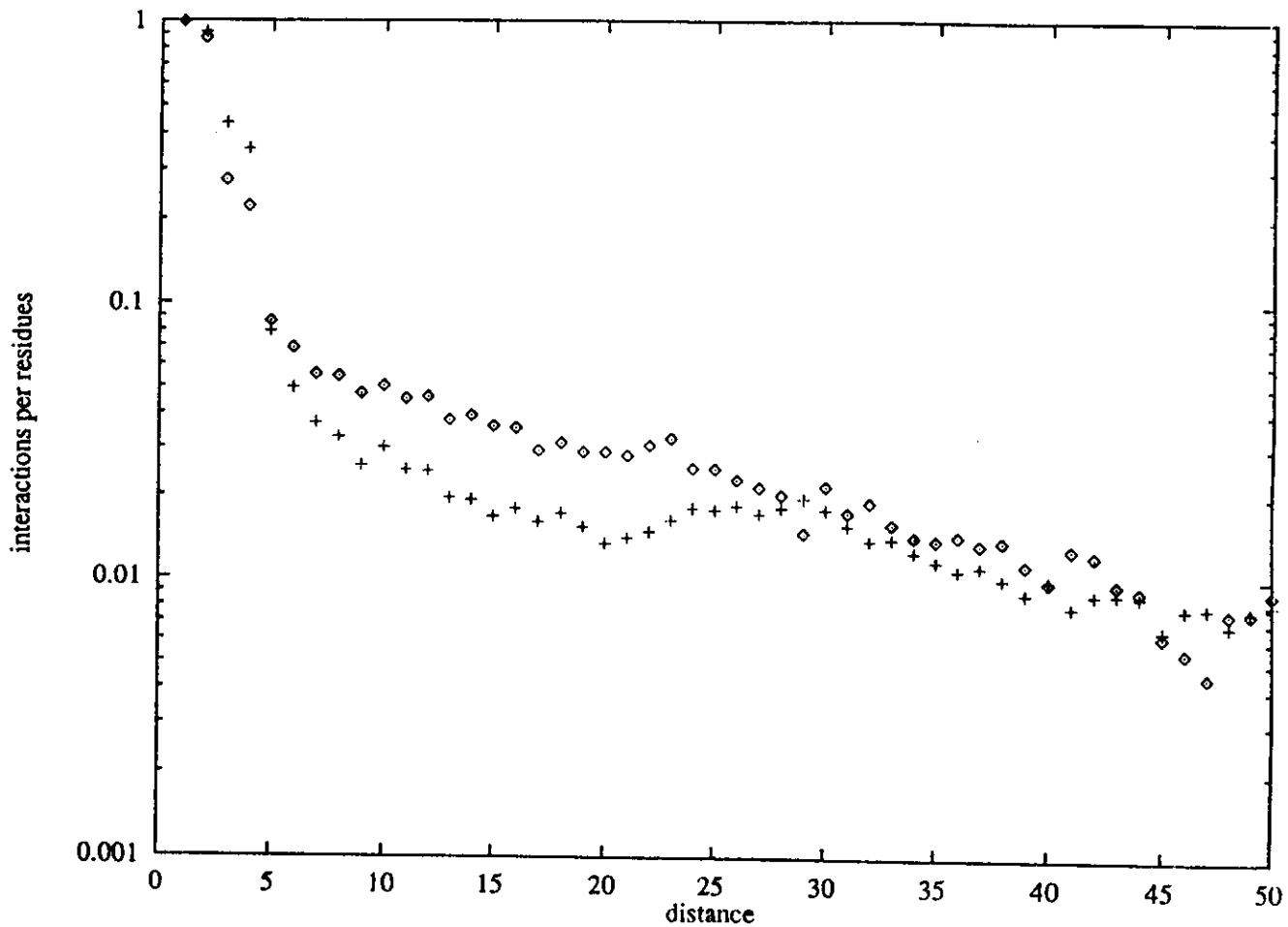


FIGURE 2 The average of the number of interactions per residues as a function of sequential distance, on semi-logarithmic scale, for proteins with (◊) and without (+) disulfide bonds.

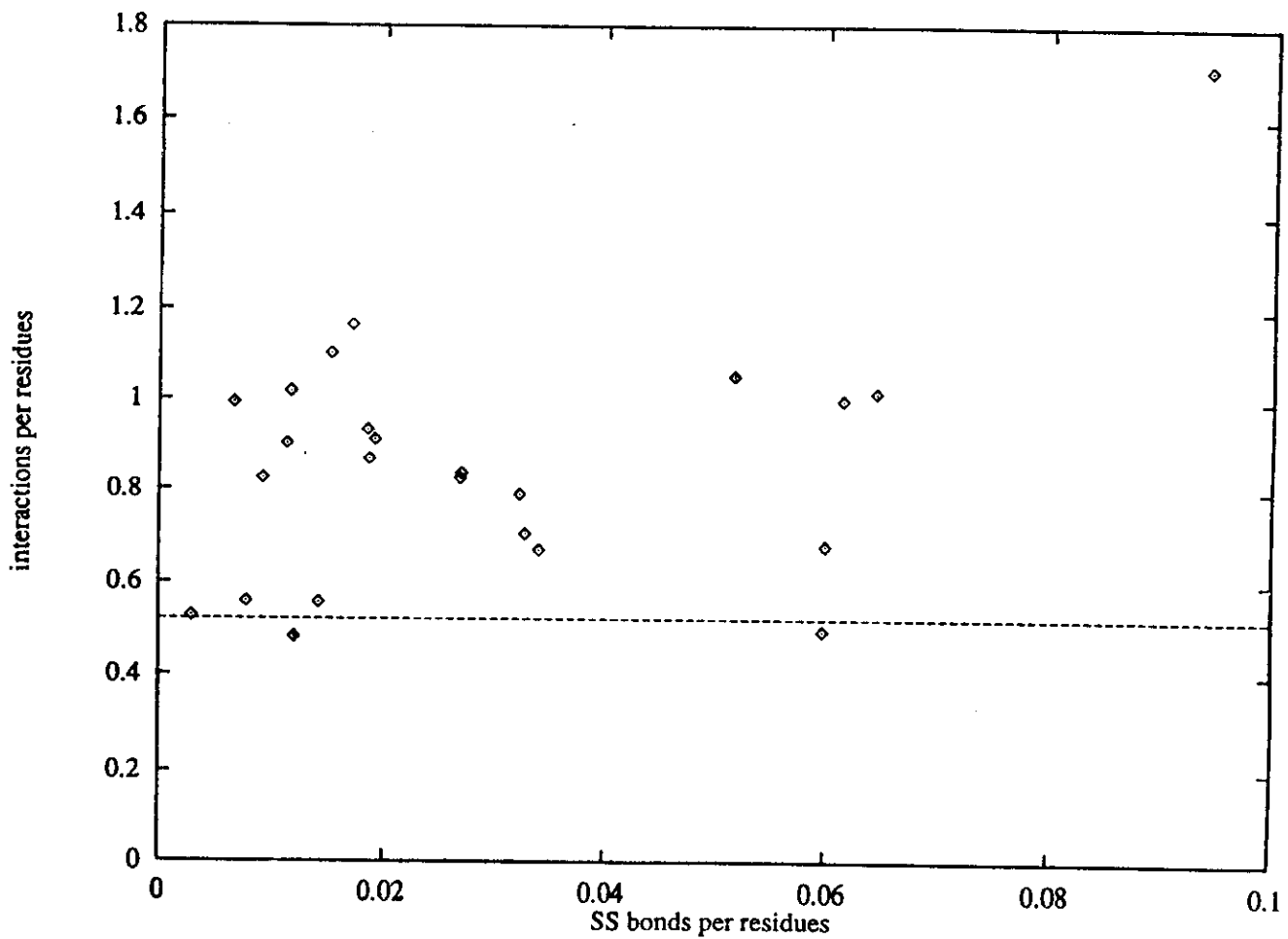


FIGURE 3 The number of interactions per residues for proteins with disulfide bonds as a function of the number of SS bonds per residues. The horizontal line marks the average number of interactions per residue for disulfide free proteins.

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

	proteins with SS bonds		disulfide free	
	intact	"modified" [§]	proteins	Δ^{\dagger}
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	0.708	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.

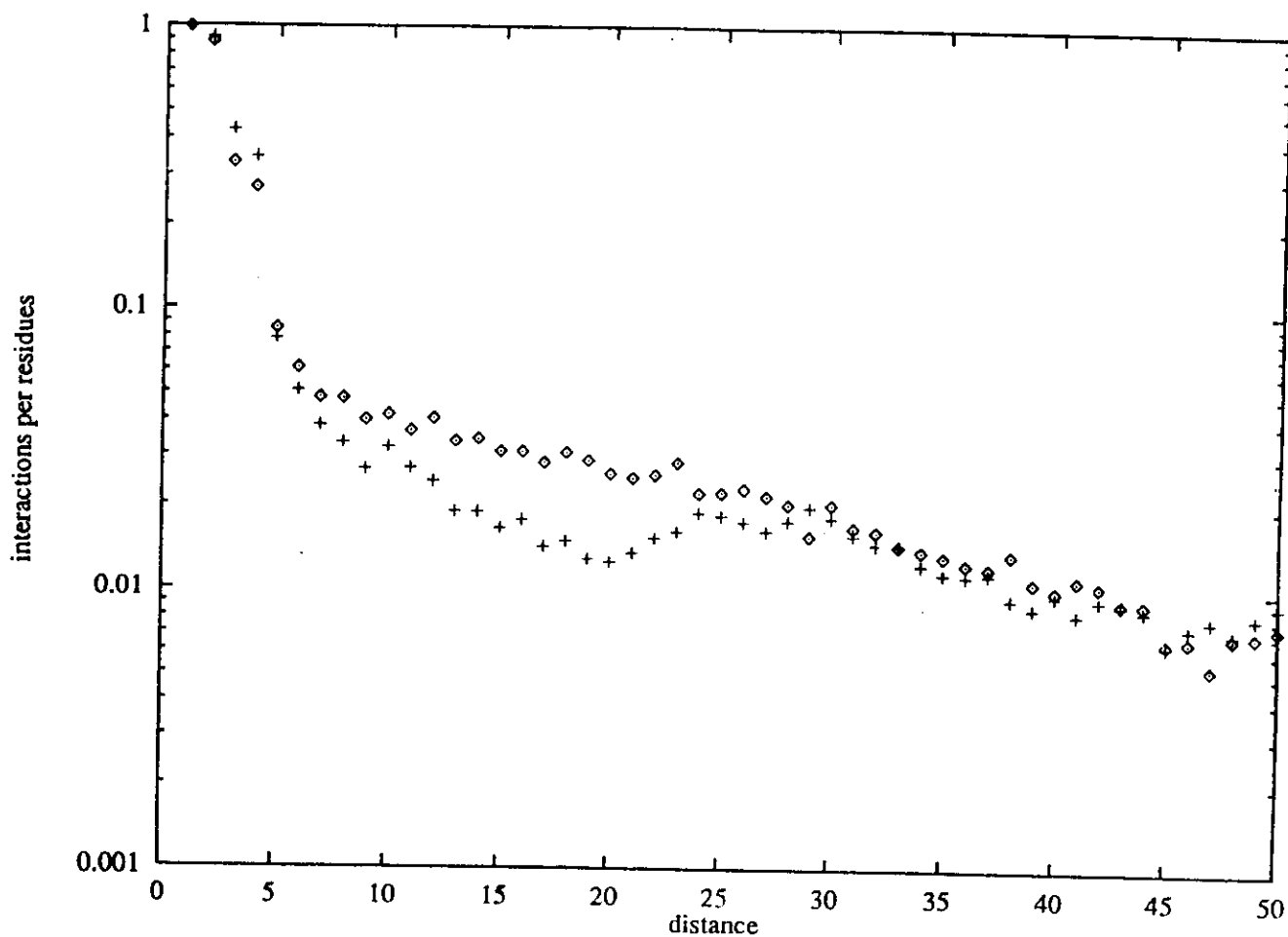


FIGURE 4 The average of the number of interactions per residues as a function of sequential distance, on semi-logarithmic scale, for extracellular (\diamond) and intracellular (+) proteins.

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.

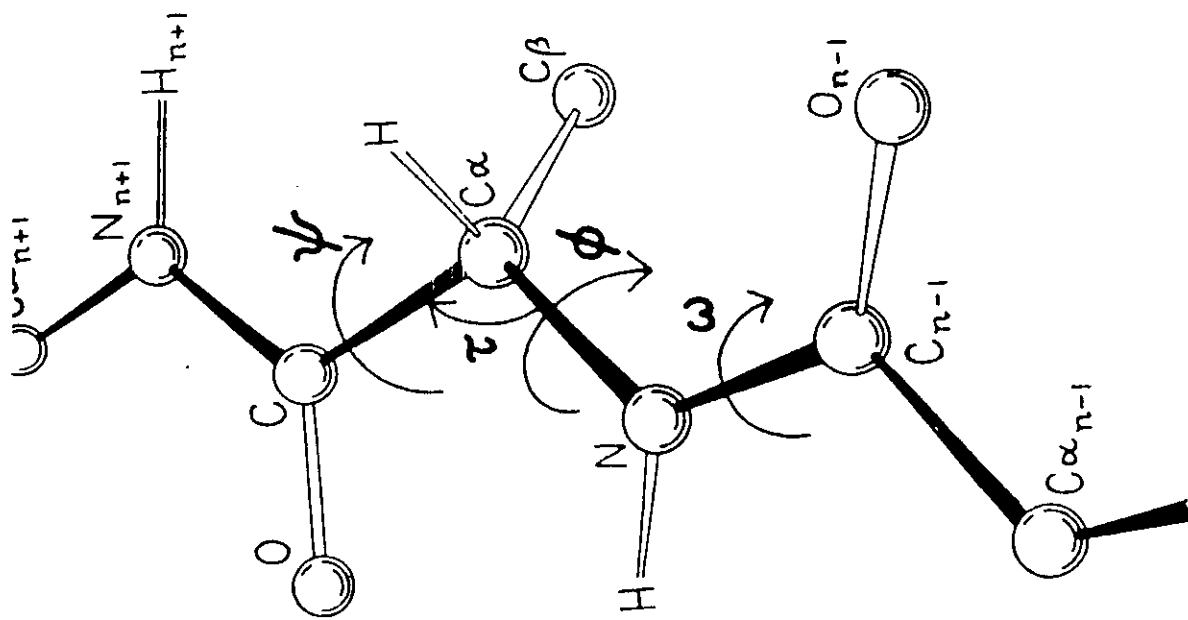


Figure 3. A key to nomenclature for the atoms of the polypeptide chain, the tetrahedral bond angle τ , and the backbone dihedral angles ϕ , ψ , and ω .

$$E_{pot} = \sum_b D_b \left[1 - e^{-\alpha(b-b_0)^2} \right] + \frac{1}{2} \sum_{\theta} H_{\theta}(\theta - \theta_0)^2 + \frac{1}{2} \sum_{\phi} H_{\phi} [1 + s \cos(n\phi)] + \quad (1)$$

$$\quad (2) \quad (3)$$

$$+ \frac{1}{2} \sum_x H_x \chi^2 + \sum_b \sum_{b'} F_{bb'}(b - b_0)(b' - b'_0) + \sum_{\theta} \sum_{\theta'} F_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta'_0) + \quad (4)$$

$$\quad (5) \quad (6)$$

$$+ \sum_b \sum_{\theta} F_{b\theta}(b - b_0)(\theta - \theta_0) + \sum_{\phi} F_{\phi\theta\theta'} \cos\phi(\theta - \theta_0)(\theta' - \theta'_0) + \sum_x \sum_{x'} F_{xx'} \chi \chi \quad (7)$$

$$\quad (8) \quad (9)$$

$$+ \sum \varepsilon \left[\left(\frac{r^*}{r} \right)^{12} - 2 \left(\frac{r^*}{r} \right)^6 \right] + \sum \frac{q_i q_j}{\varepsilon r_{ij}} \quad (10)$$

$$\quad (11)$$

Conformational energy terms

Coulomb term:

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

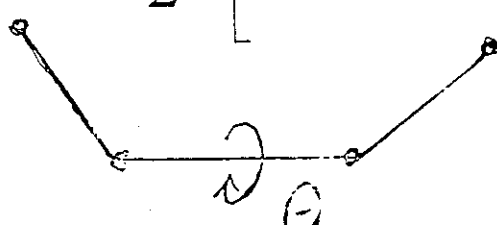
Lennard-Jones 6-12 potential:

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

Hydrogen-bonding potential:

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

Intrinsic torsional potential:

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


The diagram illustrates a torsional potential for a four-atom chain. It shows a central bond between two atoms, with two other atoms attached to them, forming a zig-zag shape. The angle between the two outer bonds is labeled as θ . A curved arrow indicates the rotation around the central bond.

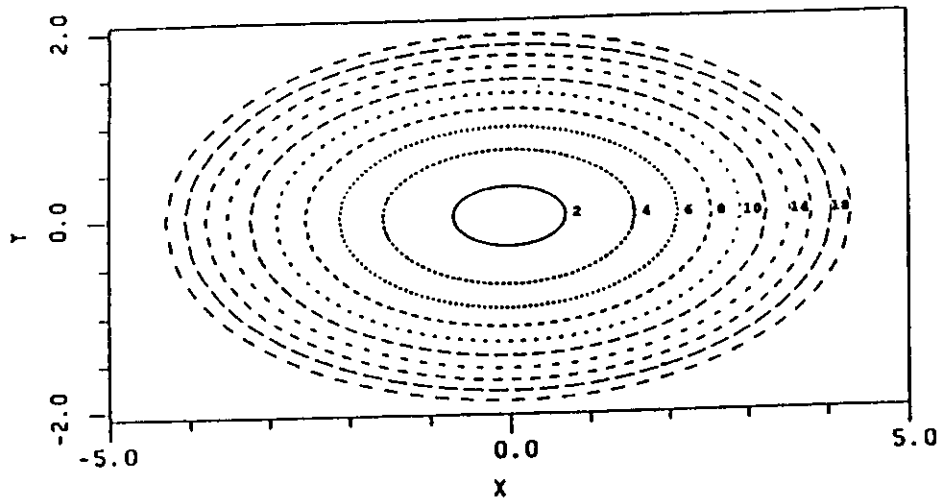


Figure 1. An energy contour surface for the function $x^2 + 5y^2$. Each contour represents an increase of two arbitrary energy units.

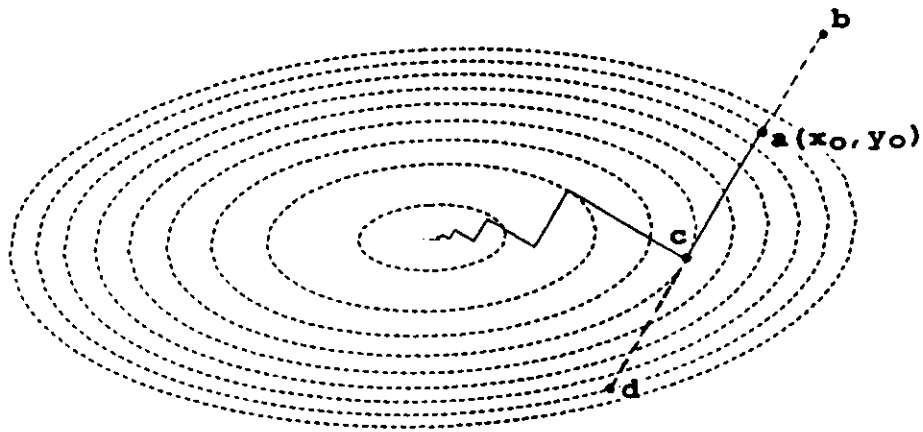


Figure 2. An energy surface for equation 1 with the gradient from the initial point $a(x_0, y_0)$ defining the line search direction. Note that the gradient does not point directly to the minimum. Compare this representation to that of Fig. 3, where the line $(b-a-c-d)$ is searched in one dimension for the minimum. Note that the minimum (point c) occurs precisely at the point where the gradient is tangent to the energy contours, thus implying that the subsequent gradient will be orthogonal to the previous gradient.

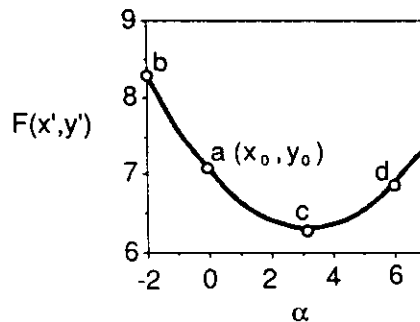


Figure 3. The cross section of the energy surface defined by the intersection of the gradient in Fig. 2 with the energy surface. The independent variable, α , is a one-dimensional parameter that is adjusted to minimize the value of the function $E(x', y')$, where x' and y' are parameterized in terms of α in equation 2. The point a corresponds to the initial point (when α is 0), and point c is the local one-dimensional minimum. Points b and d along with a bound the minimum and form the basis for an iterative search for the minimum.

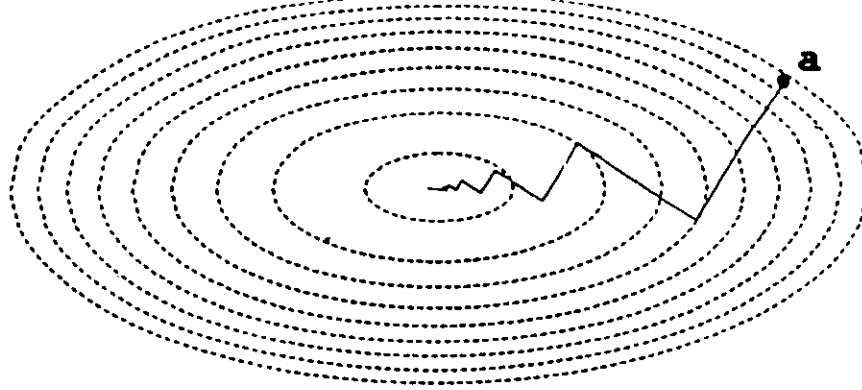


Figure 4. Minimization path following a steepest descent path using complete line searches starting from point a and converging on the minimum in about 12 iterations. In this case, where a rigorous line search is carried out, approximately eight function evaluations were needed for each line search using a quadratic interpolation scheme. Note how steepest descents consistently overshoots the best path to the minimum, resulting in an inefficient oscillating trajectory.

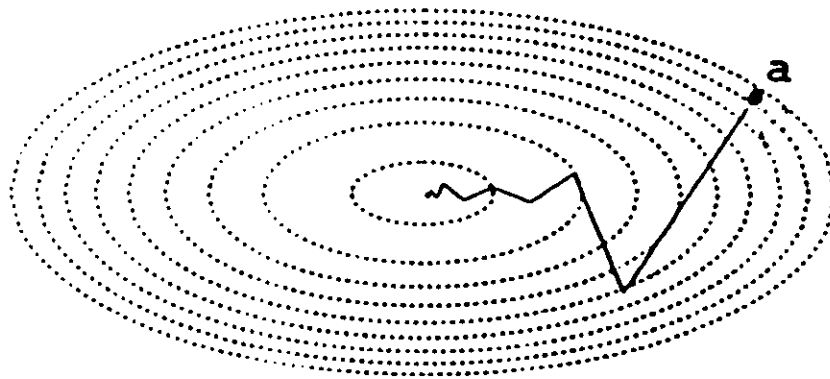


Figure 5. Minimization path following a steepest descent path with no line searches starting from point a and converging on the minimum in about 12 iterations. Although the number of iterations is comparable to a steepest descent path with line searches (Fig. 4), the total minimization was five times faster since, on the average, each iteration used only 1.3 function evaluations. In most applications to biological systems, the function evaluation is the most time-consuming portion of the calculation.

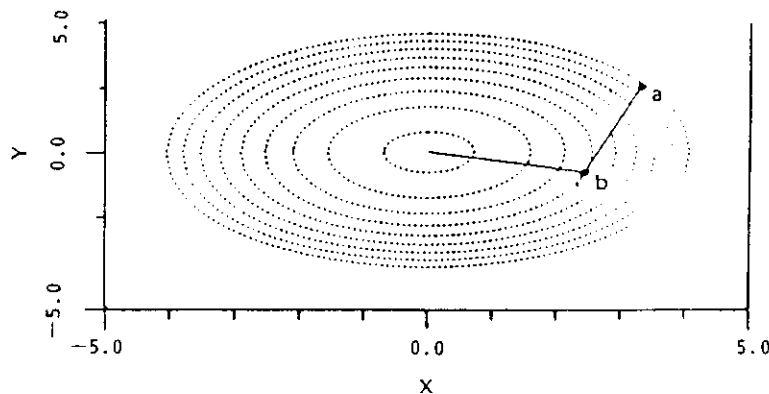


Figure 6. Minimization path following a conjugate gradients path with line searches starting from point a and converging on the minimum in two iterations. The total number of function evaluations needed was approximately half those needed for steepest descents without line searches and only 10% compared to steepest descents with line searches. As in steepest descents, successive line searches result in a set of mutually orthogonal gradients. Unlike steepest descents, however, successive directions are not orthogonal but rather conjugate. Constraining the second direction to be conjugate to the first results in a vector that passes through the minimum. The two directions used are thus a complete set of mutually conjugate directions for this two-dimensional system. Conjugate gradients will converge in N iterations for a harmonic system (where N is the dimensionality of the system) if the line search is exact. Anharmonic systems may require several passes of N steps each.

Figure 7. The number of nonbonded pairwise interactions (in millions) expected for a 5000-atom system as a function of a cutoff distance. The time required to evaluate the total energy of this system is approximately proportional to the number of nonbonded interactions.

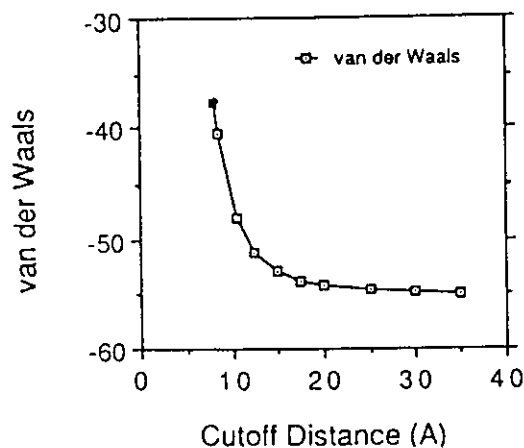
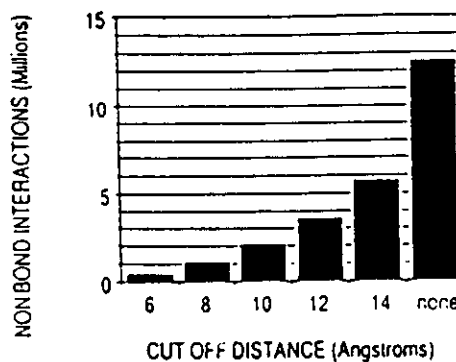


Figure 8. The van der Waals energy for the hexapeptide crystal [Ala-Pro-D-Phe]₂ as a function of energy cutoff distance. Note that the van der Waals energy does not converge until ≈ 20 Å.

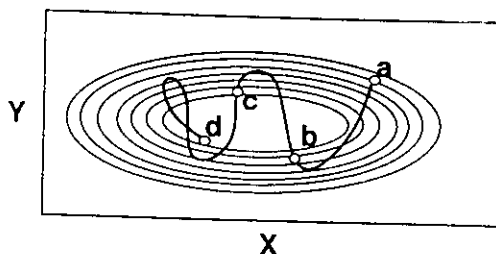
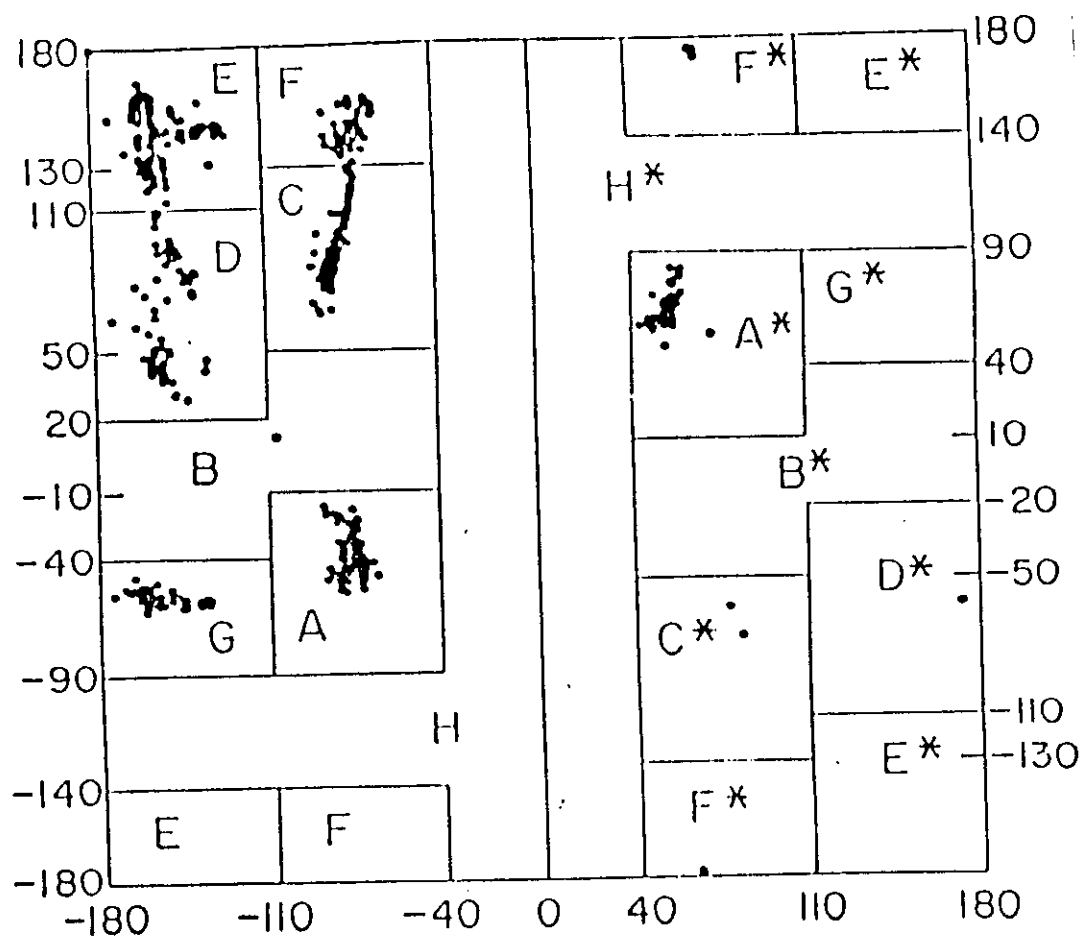


Figure 11. A dynamics trajectory for a particle beginning at rest at point a and continuing with constant total energy (kinetic plus potential) until point d. On this elliptic energy surface, minimization from any of the points, a, b, c, or d, will converge to the same point. Thus, minima can be used to characterize many nearby points sampled during dynamics. In molecules in which there are many more degrees of freedom, periodic minimization during a dynamic calculation will result in relatively fewer (compared to the number of discrete dynamic structures) structures for use in detailed structural and energetic studies.

ψ



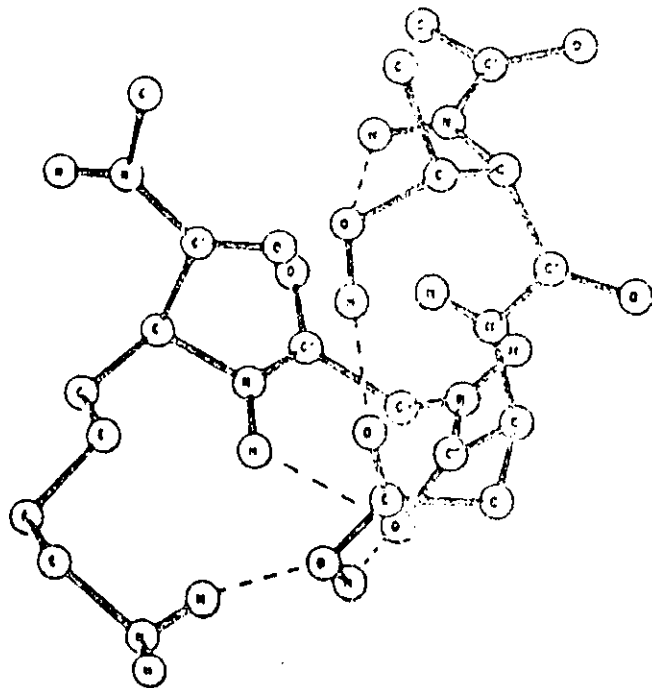
ϕ

Conformational Energy Calculations of the Effects of Sequence Variations on the Conformations of Two Tetrapeptides¹

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.



Lowest energy structure of Thr-Asp-Gly-Lys

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

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.

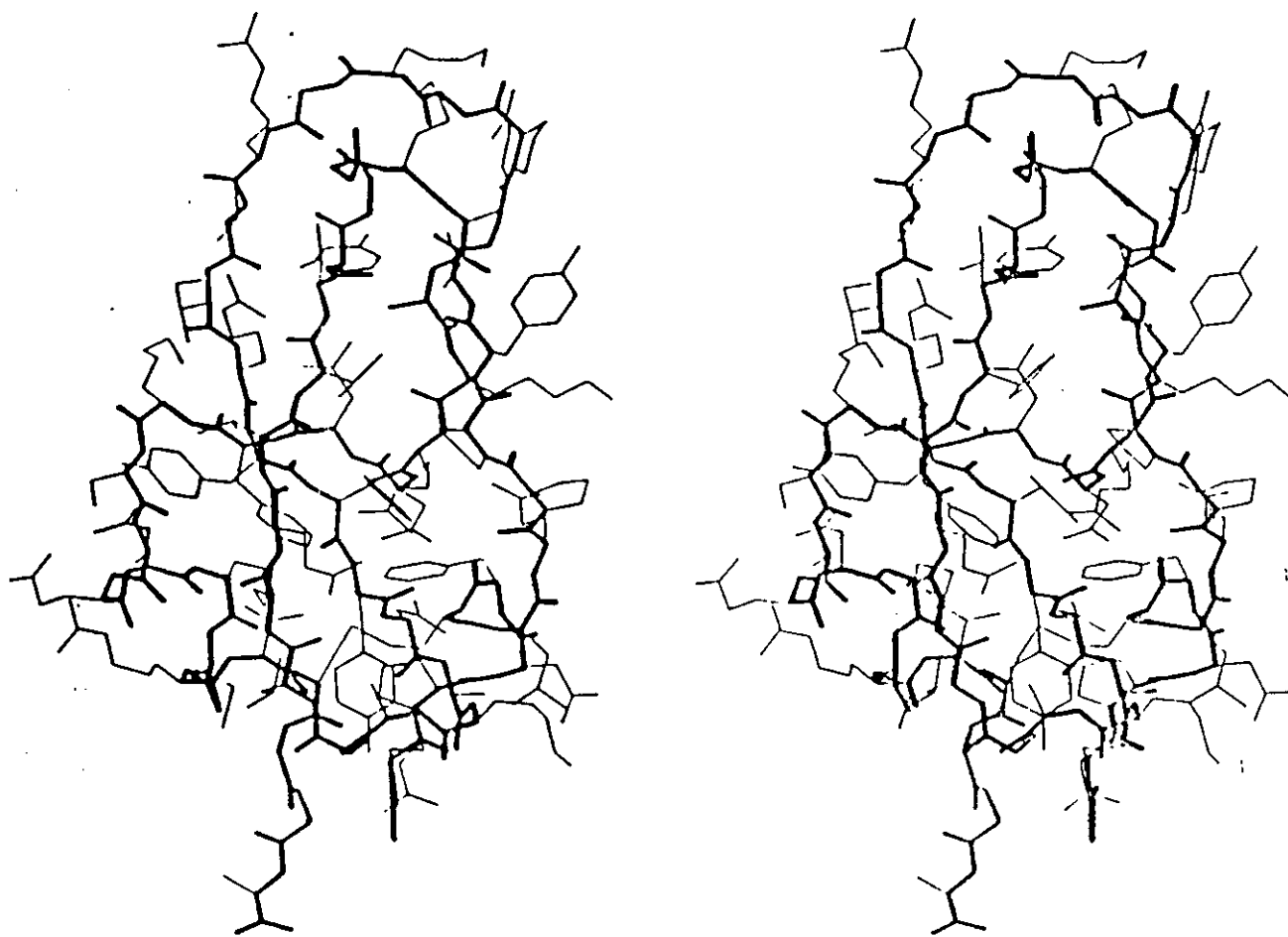
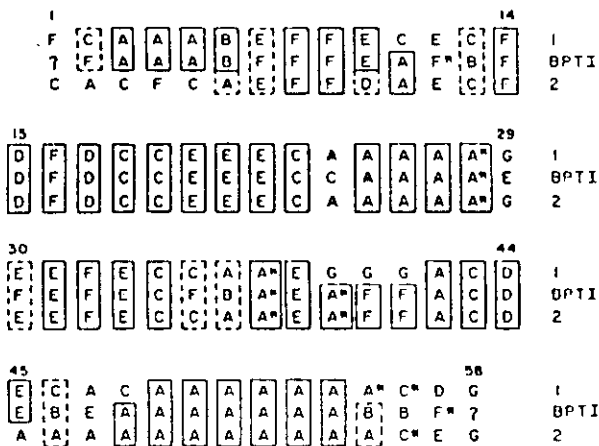
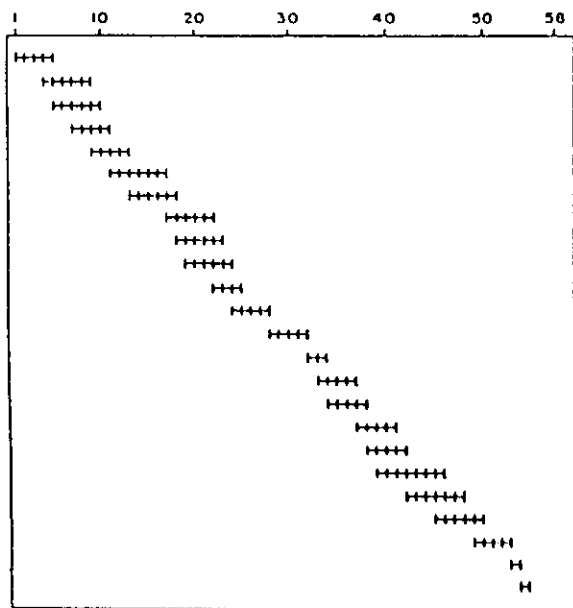


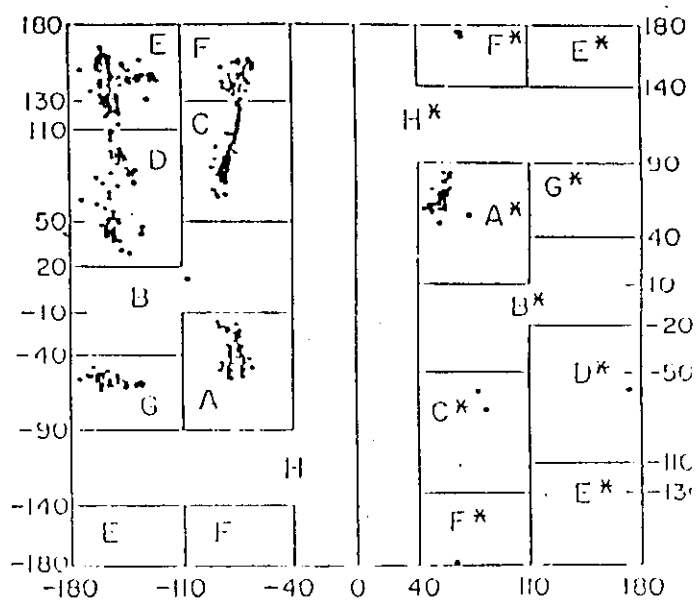
FIG. 2. Stereo drawing of all nonhydrogen atoms of basic pancreatic trypsin inhibitor. The main chain is shown with heavy lines and side chains with thin lines.

Simon, J. et al. (1991) PNAS 88 3661

Calculation of protein conformation as an assembly of stable overlapping segments (8)



ψ



ψ

