

This material may be protected by copyright law (Title 17 US Code)

DD704731

CISTI ICIST

CI-05882790-0

Document Delivery Service
in partnership with the **Canadian Agriculture Library**Service de fourniture de Documents
en collaboration avec la **Bibliothèque canadienne de l'agriculture****THIS IS NOT AN INVOICE / CECI N'EST PAS UNE FACTURE**ANTHONY ARTALE
MED LIB NATHAN CUMMINGS CTR (S-46)
MEMORIAL SLOAN KETTERING CANCER CTR
1275 YORK AVENUE
NEW YORK, NY 10021
UNITED STATES

ORDER NUMBER:	CI-05882790-0
Account Number:	DD704731
Delivery Mode:	ARI
Delivery Address:	arielsf.infotrieve.com/140.16 3.217.217
Submitted:	2005/11/28 13:55:31
Received:	2005/11/28 13:55:31
Printed:	2005/11/28 18:22:38

Direct	Book	OPENURLOPAC	UNITED STATES
---------------	-------------	--------------------	----------------------

Client Number: DDS36858

Title: PROTEIN FOLDING : PROCEEDINGS OF THE 28TH CONFERENCE OF THE GERMAN
BIOCHEMICAL SOCIETY, HELD AT THE UNIVERSITY OF REGENSBURG, REGENSBURG,
WEST GERMANY, SEPTEMBER 10-12, 1979

Author: JAENICKE, R.

DB Ref. No.: IRN11542883

ISBN: ISBN0444801979

Date: 1980

Pages: 289-316

Article Title: COMPOSITION, COOPERACTIVITY AND RECOGNITION

Article Author: LIFSON, S

Report Number: IRN11542883

Publisher: ELSEVIER/NORTH HOLLAND BIOMEDICAL PRESS,

**Estimated cost for this 28 page document: \$10.2 document supply fee +
\$33.6 copyright = \$43.8**

The attached document has been copied under license from Access Copyright/COPIBEC or other rights holders through direct agreements. Further reproduction, electronic storage or electronic transmission, even for internal purposes, is prohibited unless you are independently licensed to do so by the rights holder.

Phone/Téléphone: 1-800-668-1222 (Canada - U.S./E.-U.) (613) 998-8544 (International)
www.nrc.ca/cisti Fax/Télécopieur: (613) 993-7619 www.cnrc.ca/icist
info.cisti@nrc.ca info.icist@nrc.ca

National Research
Council CanadaConseil national
de recherches Canada

Page

1 / 1

COMPOSITION, COOPERATIVITY AND RECOGNITION IN PROTEINS

S. Lifson and C. Sander

Chemical Physics Department, The Weizmann Institute of Science, Rehovot, Israel

Abstract. The specific pairing of residues across strands in β -sheets is treated as a mutual recognition of residues and of strands. Residue pair correlations are obtained by statistical analysis of available atomic coordinates of proteins. Distinction is made between parallel and antiparallel strands. Solutions to the problem of deriving relevant statistical inferences from the limited data base available are suggested. Particular examples of cooperative tertiary structure contacts are given in stereo views.

Introduction

It is by now generally agreed that the uniqueness of the native conformation of proteins implies the existence of a path, or a set of paths, determined for each protein by its unique composition and sequence of amino acid residues, along which the free energy of the protein together with its surroundings decreases down to its equilibrium value. Along this path each residue is attracted and repelled by its momentary neighbours, and it is intuitively appealing to picture this interaction as a probing process in which residues are searching for their ultimate neighbours; and once all neighbours "recognize" each other, all are satisfied that the native conformation has been reached. Recognition is in this sense only a simple semantic device, a *façon de parler*, for discussing the trend of proteins to reach their equilibrium free energy in phenomenological terms and ignoring the details of the energy as a function of molecular geometry. In this lecture we examine the possibility of obtaining quantitative information about recognition in proteins from a statistical analysis of the available crystallographic data on protein conformations and their corresponding amino acid residue sequences. The main substance of the present study has been either published⁽¹⁾ or is in press⁽²⁾ or submitted⁽³⁾. We shall review these papers in brief, comment on them and present some further developments of significance.

Why β -Sheets?

β -sheets are protein substructures which are particularly suitable for the statistical analysis of recognition. They are the most ubiquitous among the elementary forms of tertiary structure, and their secondary and tertiary structures

are closely linked. A major factor in the thermodynamic stability of β -sheets are the hydrogen bonds between the backbones of adjacent β -strands, but these cannot account for the specificity of the β -sheet structures. It is, therefore, natural to conjecture that the matching of the side groups of neighbouring residues on both sides of the β -sheets may contribute significantly to both the specificity and stability of β -sheets. The statistical analysis of the pair distributions and pair correlations of neighbouring residues on adjacent β -strands, and the corresponding pair correlations of neighbouring β -strands is used in this study as a quantitative estimate of the role of residue-pair recognition and strand-pair recognition in the specificity of β -sheet formation.

The Data Base

Our data base comprized the β -sheets of 30 proteins whose amino acid sequences and atomic coordinates were fully determined. They were selected from data sets of atomic coordinates generously supplied to us by Richard Feldmann⁽⁴⁾. Their complete list is given in Refs. 1,3 and the detailed references are included in the AMSOM Atlas of Molecular Structures⁽⁴⁾. A computer program was used to select the data relevant to our model of residue-residue recognition in β -sheets.

The Model

Our criteria of identifying strand-strand recognition in β -sheets were based solely on inter-strand geometric relations which defined neighbouring residue pairs. Two stretches of the polypeptide chain were considered as recognizing each other if they contained at least three consecutive pairs of neighbouring residues, which fulfil the criteria for contacts between residue pairs. These criteria were three, two of them common to antiparallel and parallel strands, namely, (a) the distance between the C_α carbons is less than 7Å, (b) the bond vectors $C_\beta-C_\alpha$ point in about the same direction (angle between their directions $<90^\circ$). The third criterion required an appropriate juxtaposition of the backbone NH and CO bonds of the two residues⁽³⁾, typical for either

antiparallel (β_A) or parallel (β_P) strands.

According to our model, strand-strand recognition is measured by the strand-strand correlation, and this is determined by residue-pair correlation. The precise definition of these quantities follows.

Single Residue Contact Counts, Residue Pair Contact Counts, Residue-Pair Correlation, Strand Pair Correlation

The basic "events" for our statistical analysis were the residue pair counts N_{XY} , where X, Y are running indices for the 20 different amino acids. We counted the pairs separately for antiparallel (β_A) and parallel (β_P) β -strands, and obtained 788 β_A pairs and 263 β_P pairs on 170 β_A strand-pairs and 56 β_P strand-pairs, with an average 4.6-4.7 residues/strand. Since each pair is formed by 2 residues contacting each other, it is useful to define the single residue contact counts N_X as the number of contacts made by residues of type X with all other residues. The total number of contacts $N = \sum_X N_X$ is thus twice the total number of residue pairs. The relation between the number of the single residue contacts and the number of the residues themselves requires some comments. A residue inside a β -sheet makes two contacts, while on an outside strand it makes only one contact. The number of contacts N_X is an appropriate variable for the statistics of recognition, being naturally related to the number of residue pairs N_{XY} by $N_X = \sum_Y N_{XY}$, which is not the case for the number of residues of type X . Furthermore, the concept of residue contacts is indispensable when one distinguishes between parallel and antiparallel strand recognition, since in mixed β -sheets, for example, there are residues which make one β_A contact and one β_P contact.

If there were no recognition, no correlation, between residue pairs in β -sheets, then the expected number of XY pairs would be

$$E_{XY} = N_X N_Y / N,$$

by multiplication of independent probabilities. If N_{XY} differs significantly from E_{XY} , namely to an extent unlikely due to statistical random fluctuations, then we may consider the pair correlation

$$g_{XY} = N_{XY}/E_{XY}$$

as a measure of non-randomness of the pair-occurrence, namely as a measure of pair recognition; $g_{XY} > 1$ is a favourable correlation and $g_{XY} < 1$ is an unfavourable one. Thus, residue pair recognition, introduced first in an intuitive way, is henceforth a well defined statistical function, namely g_{XY} .

Cooperative strand-strand correlation is directly obtained from g_{XY} , with the assumption that two strands are correlated to each other solely by the cooperative effect of the correlations between their corresponding neighbouring pairs (we shall later reexamine this assumption): the strand-strand correlations, g_s , is simply the product of the residue pair correlations

$$g_s = g_{X_1 Y_1} g_{X_2 Y_2} \cdots g_{X_s Y_s}$$

where $g_{X_1 Y_1}$ is the correlation of the first residue pair, etc., and s is the length of the strand.

Statistical Significance

The number of single residue contacts in antiparallel (β_A) strands ranges from 188 for Ala to 29 for Pro, which is a fair basis for a statistical analysis of the distribution of single residue contacts. However, in β_P the smallest single residue contact counts are as low as 5 (Gln and His). This is a rather low count, and it is essential to consider the effect of random fluctuations which may misleadingly look like significant deviations from a random distribution without actually being related to recognition. The situation is even much worse in the analysis of distribution of residue pairs. In β_A we have 788 pairs

significantly
fluctuations,

distributed over 210 types of pairs (XY) with N_{XY} ranging between 26 and 0. In β_p it is 263 pairs for the same number of types. We applied therefore various devices to cope with the problem of statistical significance.

a measure of
an unfavour-
intuitive way,

First, we produced, by a random generator, a number (10) of sets of random pairs all having exactly the same number of single residue contacts N_X for all X, but with random pairing. In these sets, N'_{XY} deviated from E_{XY} only due to random fluctuations. We then examined the extent to which the distribution of the actual strand-pair correlations deviated from the corresponding average distribution of the random strand-pair 'correlations'.

from g_{XY} with
by the co-
distribution

Second, we compared the multiplicative average of g_{XY} over all N_{XY} pairs

$$\langle g \rangle = \left(\prod_{XY} g_{XY}^{N_{XY}} \right)^{1/N}$$

correlations,

with the corresponding average of the random pairing sets. An infinite random set should have $\langle g \rangle = 1$, but a finite random set has always $\langle g \rangle > 1$. The log of $\langle g \rangle$ is an information or entropy measure.

is the

Third, we focussed attention on the larger values of the N_X and N_{XY} , which are of sufficient size to be amenable to χ^2 analysis of significance as, for example, the set of individual correlations among the hydrophobic residues in antiparallel (β_A) strands.

from
of
single

Here we further pursue the question of statistical significance and physical significance in several ways. First, led by the results of Refs. 1-3 we prepared a set of stereo-pictures of a selected number of β -sheets, to obtain a visual demonstration of some particular correlations. These pictures brought out some interesting observations which we shall discuss below.

then

on

large

Second, we grouped together residues which have, on the one hand, a high resemblance in size, structure, polarity and genetic exchangeability, and on the other hand, a relatively low count of N_X and N_{XY} : Phe-Trp, Lys-Arg, Glu-Gln, Asp-Asn, Met-Cys. In this way we increased the expected value E_{XY} of most

"events" to above a threshold level of 5 (commonly required as a minimum for χ^2 analysis), while sacrificing minimally with respect to their individuality.

Third, we ordered the residues according to a decreasing order of their statistical significance (with minor alterations to bring residues of similar polarity nearer each other). This facilitates the examination of the results, drawing attention to entries closer to the top-left of the Tables.

Residue Preferences in Antiparallel (β_A) and Parallel (β_P) Strands

Table 1 was derived from Table 1 in Ref. 2 by grouping together similar residues of lower contact counts, and reordering the results in decreasing order of statistical significance. It contains the single residue contact counts in all β -sheets (N_X^{all}), in antiparallel β_A strands (N_X^A) and in parallel β_P strands (N_X^P). Other columns show the corresponding frequencies (in %, i.e. $f_X = 100 N_X/N$); the global frequency f_X^{Glob} of residues in all protein substructures was taken from the AMSOM Atlas⁽⁴⁾; and the corresponding conformational preferences $g_X = f_X/f_X^{Glob}$. Conformational preferences are related to the intrinsic preference of residues for the β -type backbone conformation, as well as to the role of residues as partners in recognition of residues on neighbor strands. The latter role is clearly demonstrated by the significant differences in conformational preferences between antiparallel and parallel strands. Our g_X^{all} agree in general with conformational preferences of residues obtained by other authors; for details see Ref. 2.

In both β_A and β_P the predominance of the branched aliphatic residues Val, Leu and Ile is most significant. However in parallel strands β_P the preferences of Val and Ile are much higher, while the preferences of many other residues are much lower than in β_A . The distinction between β_P and β_A is clearly demonstrated by the column g_X^P/g_X^A . Among the residues of high counts, Val and Ile have the largest ratios, while particularly low values (high relative preference in β_A) are observed for Thr and Tyr.

The conformational preferences in antiparallel and parallel β -strand

Table 1. Amino Acid Residue Contacts, Frequencies and Preferences in Antiparallel (β_A) and Parallel (β_P) Strands

	Contact Counts			Frequencies (%)			β -sheet preferences			
	N_X^{all}	N_X^A	N_X^P	f_X^{all}	f_X^A	f_X^P	f_X^{glob}	g_X^{all}	g_X^A	g_X^P
1 VAL	296	188	108	14.1	11.9	20.5	7.8	1.81	1.53	2.63
2 LEU	194	141	53	9.2	8.9	10.1	7.1	1.30	1.26	1.42
3 ILE	175	112	63	8.3	7.1	12.0	4.6	1.81	1.54	2.60
4 ALA	163	119	44	7.8	7.6	8.4	8.4	0.92	0.90	1.00
5 THR	160	139	21	7.6	8.8	4.0	6.8	1.12	1.30	0.59
6 SER	156	123	33	7.4	7.8	6.3	9.0	0.82	0.87	0.70
7 PHE+TRP	147	115	32	7.0	7.3	6.1	5.2	1.35	1.40	1.17
8 TYR	119	98	21	5.7	6.2	4.0	3.7	1.53	1.68	1.08
9 LYS+ARG	160	128	32	7.6	8.2	6.1	9.9	0.77	0.82	0.62
10 GLU+GLN	130	110	20	6.1	7.0	3.9	8.2	0.74	0.85	0.48
11 ASP+ASN	106	80	26	5.0	5.1	5.0	9.5	0.53	0.54	0.53
12 GLY	119	81	38	5.7	5.1	7.2	9.2	0.61	0.56	0.79
13 MET+CYS	91	69	22	4.4	4.4	4.2	3.7	1.19	1.19	1.13
14 HIS	49	44	5	2.3	2.8	1.0	2.5	0.93	1.12	0.38
15 PRO	37	29	8	1.8	1.8	1.5	4.4	0.40	0.42	0.35

pairs are thus seen to be real and interesting. They may have implications for prediction schemes of protein folding, and are obviously a basis of recognition among residue-pairs in β -sheets.

Pair Correlations in β_A

Table 2 shows the residue pair counts N_{XY} (upper entries, with their corresponding expectation values $E_{XY}=N_X N_Y/N$ (lower entries, in italics), for antiparallel β_A -strands. The counts for all 210 values of N_{XY} are given in Ref. 1. Here we grouped some of the residues of lower counts, thus reducing the number of types of pairs from 210 to 91, such that all counts are above a formal significance level of 5, except Gly-Gly ($E_{\text{Gly Gly}}=4.2$, $N_{\text{Gly Gly}}=0$). Obviously for many of these entries the significance level is still low, and the main attention should be directed to the upper-left part of Table 2. Nevertheless, extreme deviations from relatively low E_{XY} are interesting to notice, and may prompt a closer examination of the reasons for their occurrence.

In Table 3 we present the pair correlations $g_{XY}=N_{XY}/E_{XY}$ for antiparallel (β_A) strands. This Table indicates, on the one hand, the general trend of higher correlations among groups of similar polarities (hydrophobic-hydrophobic, polar-polar) which was already reported by von Heijne and Blomberg^(5,6). Indeed, a χ^2 test of the significance of correlations between hydrophobic, neutral and polar groups in our data base confirmed their conclusion, giving a 99.9% confidence level for rejecting the hypothesis of random correlations⁽³⁾.

On the other hand Table 3 indicates large variations within each group, which point toward more detailed, more individual correlations. Note, for example, the Leu-Leu correlation 0.6 compared to Val-Leu 1.5 and Val-Ile 1.7. To further examine this point we performed a χ^2 analysis on the hydrophobic residues only, against the "null hypothesis" that all hydrophobic residues are equally preferred. This hypothesis was rejected with a confidence level of 97% in favour of individual correlations within the group of hydrophobic residues. For example the pair Val-Ile has a correlation $g_{\text{Val/Ile}}=1.5$ within the set of

Table 2. Residue Pair Contacts N_{XY} in Antiparallel β_A -Strands and Their Random Expectation Values E_{XY} .
 N_{XY} - upper entries, E_{XY} - lower entries, in italics, $N_X = \sum_Y N_{XY}$

	VAL	LEU	THR	SER	ALA	ILE	PHE, TRP	TYR	LYS, ARG	ASP, ASN, HIS	GLU, GLN	MET, CYS, PRO	GLY	N_X
1 VAL	26 22.4	25 16.8	6 16.6	19 14.6	10 14.2	23 13.3	18 13.7	11 11.7	13 15.3	8 14.8	6 13.1	11 11.7	12 9.7	188
2 LEU		8 12.6	12 12.4	5 11.0	10 10.6	12 10.0	14 10.3	12 8.8	5 11.5	10 11.1	4 9.8	14 8.8	10 7.2	141
3 THR			18 12.3	21 10.8	13 10.5	9 9.9	3 10.1	9 8.6	13 11.3	7 10.9	14 9.7	10 8.6	4 7.1	139
4 SER				14 9.6	11 9.3	4 8.7	4 9.0	1 7.6	11 10.0	13 9.7	7 8.6	8 7.6	5 6.3	123
5 ALA					4 9.0	18 8.5	8 8.7	7 7.4	8 9.7	7 9.4	7 8.3	9 7.4	7 6.1	119
6 ILE						8 8.0	9 8.2	5 7.0	8 9.1	9 8.8	3 7.8	2 7.0	2 5.8	112
7 PHE, TRP							8 8.4	8 7.2	12 9.3	10 9.0	7 8.0	7 7.2	7 5.9	115
8 TYR								12 6.1	7 8.0	6 7.7	8 6.8	6 6.1	6 5.0	98
9 LYS, ARG									12 10.4	11 10.1	16 8.9	5 6.8	7 6.6	128
10 ASP, ASN, HIS										20 9.8	12 8.7	2 7.7	9 6.4	124
11 GLU, GLN											10 7.7	8 6.8	8 5.7	110
12 MET, CYS, PRO												12 6.1	4 5.0	98
13 GLY													0 4.2	81
N_X	188	141	139	123	119	112	115	98	128	124	110	98	81	1576

Table 3. Residue Pair Correlations g_{xy} in Antiparallel β_A -Strands

	1	2	3	4	5	6	7	8	9	10	11	12	13
	VAL	LEU	THR	SER	ALA	ILE	PHE, TRP	TYR	LYS, ARG	ASP, ASN	GLU, GLN	MET, CYS	GLY
1 VAL	1.2	1.5	.36	1.3	0.7	1.7	1.3	.94	.85	.54	.46	.94	1.2
2 LEU		.63	.96	.45	.94	1.2	1.4	1.4	.44	.90	.41	1.6	1.4
3 THR			1.5	1.9	1.2	.91	.30	1.04	1.2	.64	1.4	1.2	.56
4 SER				1.5	1.2	.45	.45	.13	1.1	1.3	.82	1.05	.79
5 ALA					.44	2.1	.92	.95	.83	.74	.84	1.2	1.1
6 ILE						1.0	1.1	.72	.88	1.00	.38	.29	.35
7 PHE TRP							.95	1.1	1.3	1.1	.87	.98	1.2
8 TYR								2.0	.88	.78	1.2	.98	1.2
9 LYS, ARG									1.2	1.09	1.8	.73	1.06
10 ASP, ASN, HIS										1.7	1.4	.26	1.4
11 GLU, GLN											1.3	1.2	1.4
12 MET, CYS, PRO												1.8	.79
13 GLY													.00
N_X	188	141	139	123	119	112	115	98	128	124	110	98	81

hydrophobic residues, which is only slightly less than its 1.7 correlation within the set of all amino acid residues. A similarly high individual correlation is found for Ser and Thr, going far beyond the average correlation among the "neutral" (partly polar) class.

The highest favourable correlations found hardly leave any doubt as to the individual specificity of the recognition. However, the physical nature of the recognition, whether mediated by direct interaction or indirect interaction via a third partner, has to be investigated by other means.

Strand Recognition in β_A

The distribution of strand correlations in β_A is given in Fig. 1 by a histogram of the number of strands with a strand correlations g_s vs. $\log g_s$. If there were no correlation, i.e. no recognition, and if the sample were very large, the histogram would be distributed symmetrically with respect to $g_s=1$, like a Gaussian distribution. A finite random distribution exhibits a shift towards larger g_s due to fluctuations which divert the system from its ideally random distribution. We have included, therefore, in Fig. 1 a reference distribution, composed of an average of 10 equivalent sets of random pairs whose single residue contact distribution is exactly the same as the observed distribution N_X (see above and Ref. 3). Both the average random distribution and the actually observed distributions are shifted towards larger g_s , but there is still a difference which measures the real strand correlation. It is small, however, relative to the width of the distributions, and relative to the shift of both from an ideal (infinitely large) random distribution. This indicates that the strand recognition due to nearest neighbour pair correlations is only marginal. Possible reasons for this property are discussed below.

An overall measure of cooperative β -strand recognition is given by the average correlation per residue pair $\langle g \rangle$ (see above and Ref. 3). Its value for β_A was found to be $\langle g \rangle = 1.20$. The random distributions have $\langle g \rangle = 1.146 \pm 0.014$. These numbers support the qualitative conclusions based on inspection of the

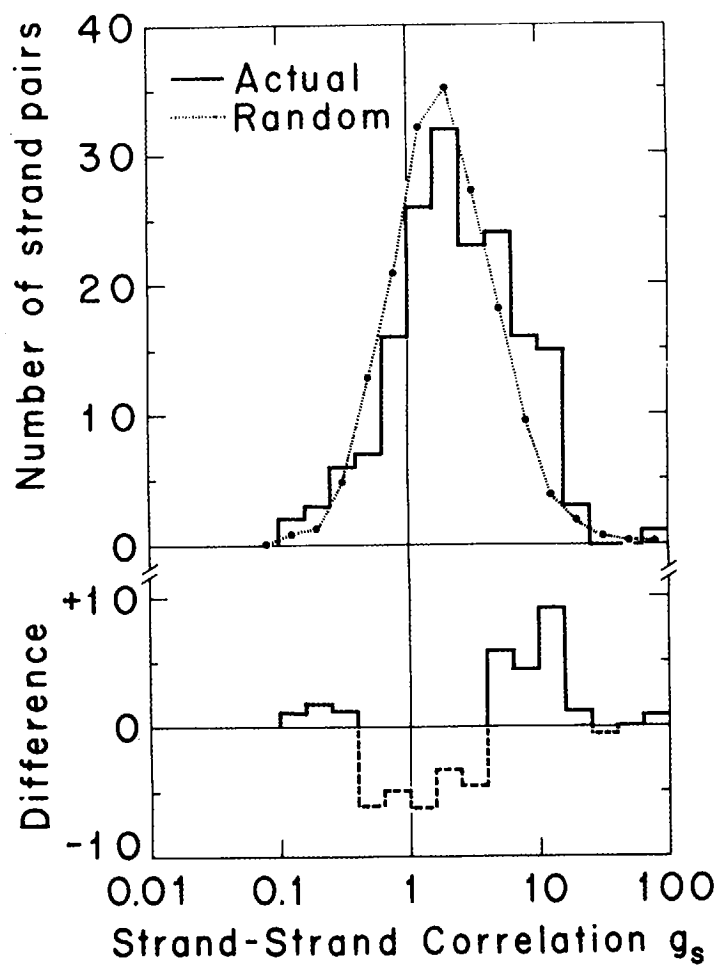


Fig. 1. Distribution of strand-strand correlations g_s for the 170 antiparallel strand pairs; solid line: *actually observed* strands, dashed line: *randomized* strands with the same amino acid content. The shift of the actual relative to the random distribution is a measure of cooperative recognition. The difference histogram (bottom) brings out the nature of the shift.

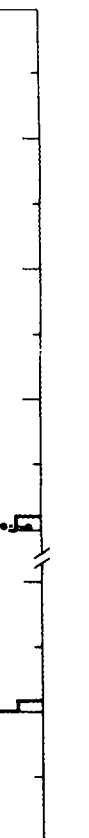
strand correlation histogram. The average of random distributions is shifted by ~ 0.15 from the ideal random distribution ($\langle g \rangle = 1$) as expected due to fluctuations in small samples. The shift of the real average correlation is 0.20, it is "marginally" larger. Yet it is significant, since the difference 0.05 is still about three times larger than the standard deviation (0.014) of the random average correlations.

Why is strand correlation only marginal? Is this consistent with the uniqueness of protein folding which seems to require high correlations? Three reasons come to mind as possible solutions of this puzzle.

First, pair correlations of nearest neighbours are only one single aspect of the recognition process. Interactions among residues of the sheet which are not immediate neighbour pairs are also significant, as we shall see below (section on Recognition in Stereo-Pictures). Furthermore, interaction between sheet residues and other neighbouring parts of the protein also contribute to recognition. Indeed the strands forming a given β -sheet may in general not form uniquely the same sheet without the cooperative effect of other parts of the whole protein.

Second, pair correlations are statistical averages while each sheet has its own character, both as a geometric structure and as a functional element in the biological role of the protein as a whole.

Third, recognition and uniqueness of the specific structure of the β -sheet as well as the protein to which it belongs are the result of natural selection, therefore correlations need not be higher than required by evolution. Variations of sequence are permitted as long as the biological function of the protein is secured, as is manifestly proven by the sequence variability of most proteins, among species and even within the same species. Thus recognition need not be perfect. It can be and is marginal in the sense that it only has to secure a marginal stability of the native structure.



100
n g_s

the 170 antiparallel
ed line: *randomized*
e actual relative to
ition. The difference

Table 4. Residue Pair Counts N_{XY} in Parallel β -Strands, Their Random Expectation Values E_{XY} and Pair Correlations g_{XY}
 $N_X = \sum_Y N_{XY}$, $N = \sum_X N_X = 526$. The common one-letter notations are used for the residues in groups 4, 5 and 7. They are:
 4 - Ala, Gly, Met, Cys, Pro; 5 - Asp, Glu, Asn, Gln, Lys, Arg, His; 7 - Phe, Tyr, Trp.

	1 VAL	2 ILE	3 LEU	4 A,G,M,C,P	5 D,E,N,Q,K,R,H	6 SER,THR	7 F,Y,W,	N_X
1 VAL	34	1.5	1.4	18	10	8	9	108
	22.2	12.9	10.8	23.0	17.0	11.1	10.9	
2 ILE		10	1.3	13	4	2	4	63
		7.5	6.3	13.4	9.9	6.5	6.3	
3 LEU			8	8	6	3	5	53
			5.3	11.3	8.4	5.4	5.3	
4 A,G,M,C,P				32	14	13	14	112
				23.8	17.7	11.5	11.2	
5 D,E,N,Q,K,R,H					24	14	11	83
					13.1	8.5	8.4	
6 SER,THR						10	4	54
						5.5	5.4	
7 F,Y,W,							6	53
							5.3	
N_X	108	63	53	112	83	54	53	526

Pair Correlations in β_p

The small number (526) of observed residue-contacts in parallel β -strands makes it impossible to examine at present in detail the individual pair correlations. Only the three aliphatic residues, whose predominance in β -sheets in general and in β_p -strands in particular, was already mentioned in the section on residue preferences, seem to yield significant individual pair correlations. We therefore adopt an analysis of results similar to that presented in Tables 1, 2 and 3 for β_A , namely, that of grouping together low count residue contacts^(1,3) provided the residues are sufficiently close to each other in polarity, size etc. However, since the counts in β_p are much lower than in β_A , more residues had to be incorporated in a group. We require that the uncorrelated expectation value for pair count should always exceed 5. For a total contact count $N=526$, this implies that each entry (individual residue or a group of residue) should have a single residue contact count larger than 50. The polar residues Lys, Arg, Glu, Gln, Asp, Asp and His have a total $N_X=83$ and must therefore form a single group. The aromatic residues have a total $N_X=53$, and form a natural group, due to their aromaticity and big ring size (see also the stereo-pictures and their discussion below). Ser and Thr, with $N_X=54$ seem also to be natural partners. The remaining residues, comprising Ala, Gly, Met, Cys and Pro, have a total $N_X=112$, and there seems to be no reasonable way of splitting them into 2 sufficiently large groups.

Table 4 represents the resulting contracted set of residue pair correlation N_{XY} for β_p -strands, together with the corresponding values of contact counts N_X , random pairs expectation values ($E_{XY}=N_X N_Y / N$) and finally (in italics) the pair correlations ($g_{XY}=N_{XY} / E_{XY}$). One notices immediately that the grouping did not wipe out interesting correlations. One observes a large number of rather high correlations and a similar number of rather low correlations, i.e. high anticorrelations (which are as relevant deviations from randomness as high correlations). This may serve as an a posteriori argument that our grouping of

residues is a good one. Altogether, one notices high correlations among the three "large, branched and hydrophobic" residues Val, Ile and Leu, and similarly high correlations among the polar and the semi-polar (Ser-Thr) groups.

Recognition in Stereo-Pictures

As we have repeatedly noted above, the residue pair correlations (Tables 3 and 4) are derived from a simplified model of recognition and small data bases of β_A and β_P residue pairs. To what extent can we consider the g_{XY} as true correlations, rather than random fluctuations? To what extent can we correlate the g_{XY} values of the table with the observed structure of β -sheets? In short, to what extent does the actual recognition process resemble the model? One way of seeking answers to such questions is to use the results as presented in the Tables as a guide to a more detailed inspection of the recognition process.

We have singled out two questions about pair correlations in β -sheets which stimulated our interest while trying to interpret the correlation tables. One question concerns the branched aliphatic residues V, L and I. The other relates to the bulky aromatic residues F, Y and W.

We sought the answers through the use of stereo-pictures, computer-drawn by PLUTO, a program for plotting molecular and crystal structures⁽⁷⁾, from protein atomic coordinates taken from AMSOM⁽⁴⁾. In these pictures the plane through the central area of the β -sheet coincides with the plane of the page. The R-groups on the upper side of the β -sheet thus show up clearly. They are drawn with atomic radii of 1.2\AA while all other atoms (excluding hydrogens) are given radii of 0.3\AA . Thus van der Waals interactions between the side groups show up clearly, yet a radius of 1.2\AA is sufficiently small not to screen the main structure of the rest of the sheet.

First, we examined the way aromatic side chains, Phe, Tyr and Trp, interact with each other and with other side-group. They are the largest side groups, and their contribution to the specific formation and stability of β -sheets may be therefore considerable. According to the correlation tables they are

relations among the
and Leu, and similarly
(Thr) groups.

correlations (Tables
and small data bases
the g_{XY} as true
can we correlate
sheets? In short,
this model? One way
presented in the
tion process.

in β -sheets
relation tables.
the other

computer-drawn
(7), from
the plane
of the page.

They are
(proteins) are
the groups
across the

to, inter-
the groups,
ments may

correlated to the large aliphatic side chains favourably in β_A but unfavourably in β_P . Examining Figs. 2-5 one sees that their orientation relative to the sheet and towards each other varies greatly. Sometimes they protrude high above the sheet, perhaps to interact with non-sheet residues, while in some cases they lie flat on the sheet reaching out even to the next nearest strand. More details are discussed in the Figure Captions.

Second, we wished to find out how V, I and L interact in β -sheets, why they are highly preferred, why they prefer the company of each other, and whether it is possible to explain the differences between their pair correlations. We picked out all strand pairs containing pairs of V, L and I. Figures 6-13 are stereo-pictures of some of the more interesting parts of β -sheets containing such pairs.

These pictures reveal beautiful patterns where each sheet manifests clearly its signature of individuality. Seeking some general trends in these patterns and comparing them with the statistically derived pair correlations, we are led to observations of some interest:

1. The groups V, L and I often tend to concentrate in the center of the β -sheets, while the edges are more inhabited by other residues. Thus one might expect that Table 1, if related only to the core of β -sheets would yield higher preferences for V, I and L (and correspondingly lower values for some other residues). This is in accord with the results obtained by Heijne and Blomberg⁽⁶⁾ who found internal residues of β -sheets to be more hydrophobic than peripheral ones.
2. The pair correlations between V, L and I tend to extend cooperatively along lines across the β -sheets, perpendicular to the direction of the strands. These lines are more pronounced at the center of the sheet, as for example the line V, V, I, V, V, V in Fig. 8 L, L, I, L in Fig. 10 or I, L, V, L in Fig. 12. This cooperative linear correlation may be of importance for the nucleation of β -sheets.

In the following stereo views (Figs. 2-13) of β -sheets (or parts of β -sheets) side groups on one side of the sheet are shown nearly space-filling; the backbone and the side groups on the bottom side of the sheet are drawn with a much smaller atomic radius. This unusual representation brings out nicely patterns of side group interaction. The direction of the strands and residue numbers can be read off the schematic, in which top-side residues are represented by their standard one-letter name, and bottom-side residues by a dot. The one letter code is: Ala-A, Gly-G, Ser-S, Thr-T, Cys-C, Asp-D, Asn-N, Glu-E, Gln-Q, Lys-K, Arg-R, His-H, Trp-W, Phe-F, Tyr-Y, Val-V, Ile-I, Leu-L, Met-M, Pro-P.

```

131 197 67
 107 48 72
+ + + + +
 L F G Y
. . . . .
 F W F A V T
. . . . .
 F F A I A V
. . . . .
 L S F Y L Y
+ + + + +
126 191 61
 113 54 77

```

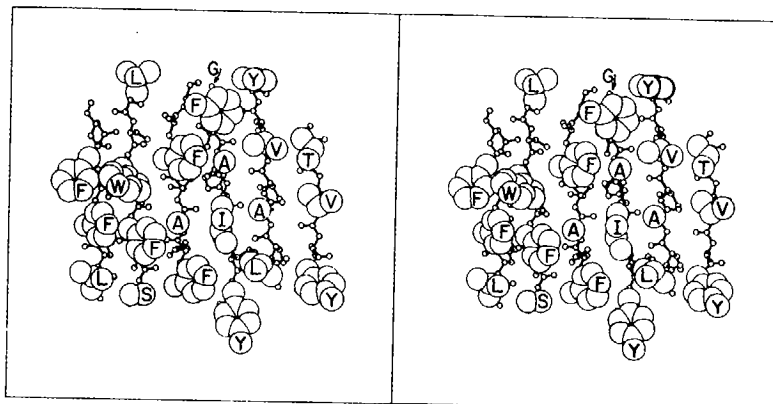


Fig. 2. Concanavalin A (AM9.1.1.2.1). Pure β_A . The location of the small Ala between the large aromatic and aliphatic side groups accentuates the trough due to the slight twist of the sheet. Most of the aromatic ring side groups are aligned cooperatively from strand to strand. Phe forms bridges to the C^β of nearby residues, in particular to Ala in the nearest neighbour position. Phe 197 covers Gly 48 and reaches over to contact Tyr 67 of the next-nearest strand.

```

36 99
 87
+ + +
 Y Y
. . .
 N Q F
. . .
 Y Y Y
+ + +
32 96
 91

```

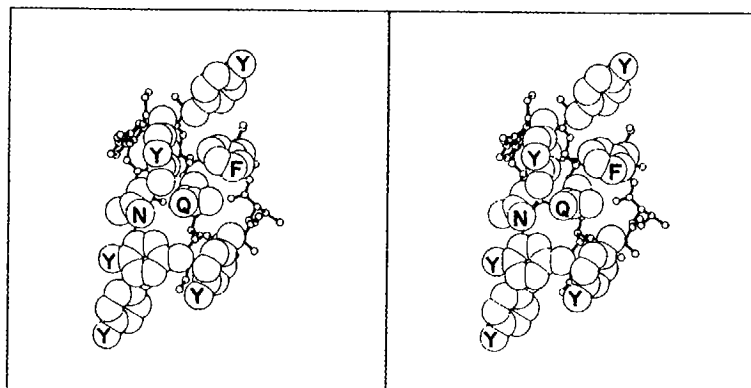


Fig. 3. Bence-Jones Protein 'REI' (AM6.1.2.1.1). Pure β_A . Here the sheet is covered by a patch of mostly Tyr side-groups. The polar hydroxyls of Tyr together with the polar heads of Asn and Gln form a highly polar surface. Thus these side groups cooperatively play a dual role as both hydrophobic and polar residue: they coat the β -sheet with a 'soapy' film, as can be observed also in some other figures.

heets (or parts of β -sheets) space-filling; the back- are drawn with a much rings out nicely patterns and residue numbers are represented by a dot. The one Asn-N, Glu-E, Gln-Q, Leu-L, Met-M, Pro-P.

117 65
94
↓ ↑ ↓
L F F
· · ·
L F V
· · ·
H L F
↓ ↑ ↓
121 69
89

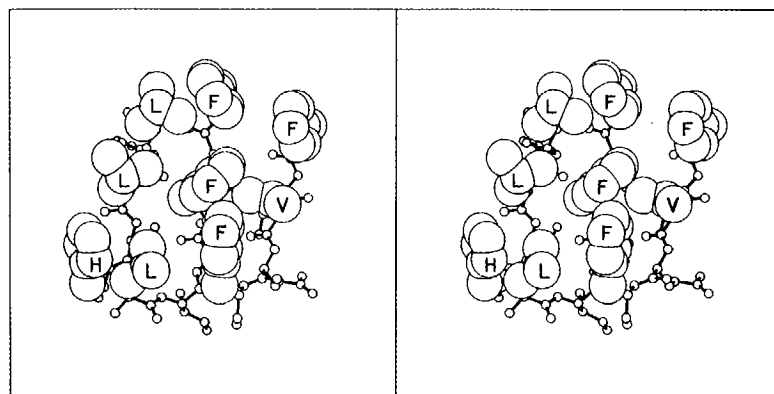


Fig. 4. Carbonic Anhydrase C (AM12.3.1.2.1). Pure β_A . The three central Phe rings protrude high above the backbone of this sheet, and together with the other bulky groups Leu, Val and Phe form a highly hydrophobic saddle surface. Side groups Ile 90 and Gln 91, as part of a 'classic' β -bulge⁽¹⁰⁾ involving Val 120, both point to the bottom side.

on of the small Ala raises the trough due side groups are to the C β of position. Phe 197 nearest strand.

49 11
6
↓ ↑ ↓
F C Y
· · ·
E Y Y
· · ·
D K P
↓ ↑ ↓
53 2 15

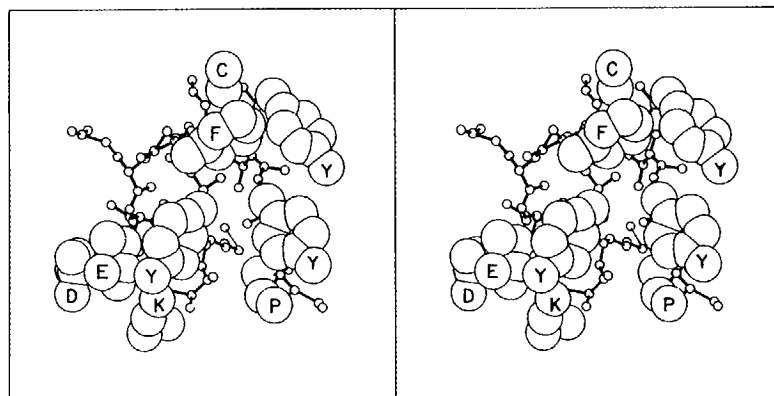
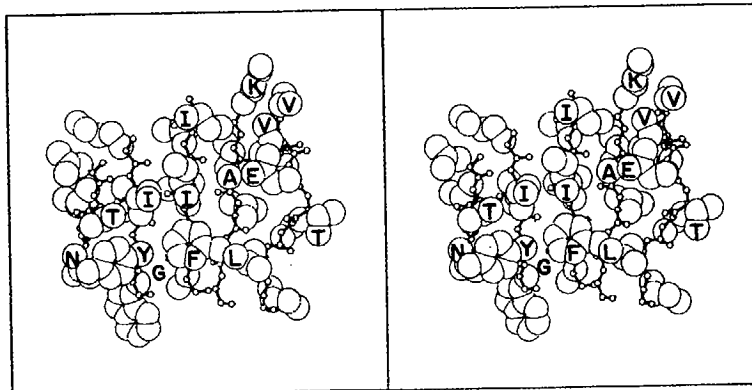


Fig. 5. Rubredoxin (AM2.1.1.2.1). Pure β_A . The strong right-handed twist of this β_A -sheet, containing a cluster of four aromatic side-chains, appears to be stabilized by the contact of Phe 49 and Tyr 11 across an intervening strand. The Tyr side chains surround Phe with purely hydrophobic contacts, while their polar ends stick out. Again, the dual role of Tyr as both hydrophobic and polar residue is expressed cooperatively.

```

31 48 109
  2 81
↓ ↓ ↓ ↓ ↓
  I K V
. . . . V
T I I A E
. . . . .
N Y G F L
↓ ↓ ↓ ↓ ↓
34 6 86
    53 116

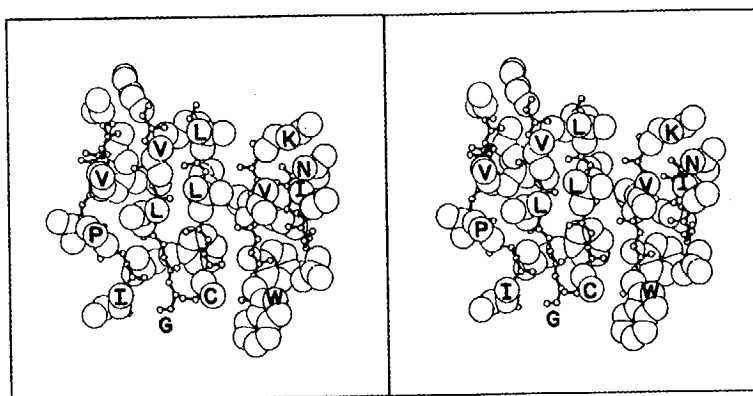
```



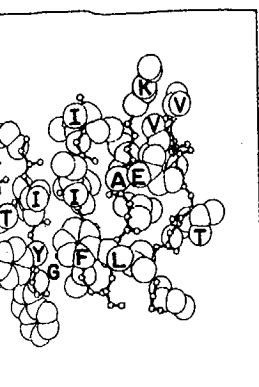
```

109 48 31
  81 2
↓ ↓ ↓ ↓ ↓
. . . . .
V V L K N
. . . . .
P L L V I
. . . . .
I G C W
↓ ↓ ↓ ↓ ↓
116 53 34
    86 6

```



Figs. 6,7. Flavodoxin (AM4.2.1.2.1). Pure β . These two figures show two sides of the same sheet. Side groups on both sides are enlarged, but only those on the top side are labelled. Both sides have a central hydrophobic patch dominated by Ile, Val and Leu, making both nearest neighbour and diagonal contacts, while all polar groups of Tyr, Thr, Cys, Trp, Asn (except Glu 112) point away from the center of the sheet. In Fig. 6 Phe 85 neatly covers a hole left by Gly 53, but in Fig. 7 the backbones of three consecutive nearest neighbours lie exposed. Side chains, Val 111, Glu 112 and Thr 113 (on top in Fig. 6) form part of a 'super'- β -bulge in which the strand-strand register is shifted by two residues. The variety of features observed here underscores the individuality of β -sheet folding.



```

71  3 115 126
28  91 144
+ + + + +
I V I Y K
. . . . .
V V I V V V P
. . . . .
N D G S S N F
+ + + + +
75  7 119 128
32  95 146

```

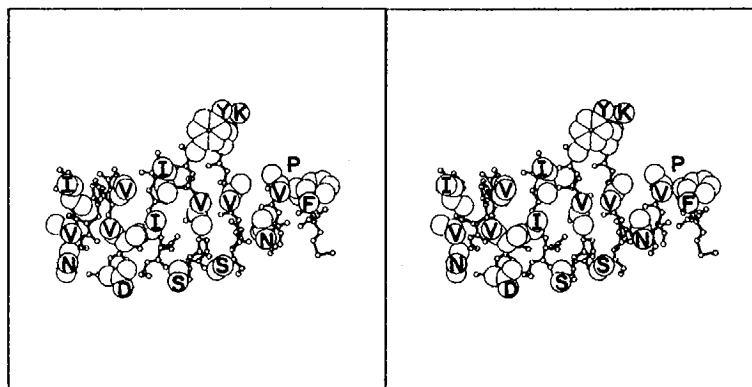
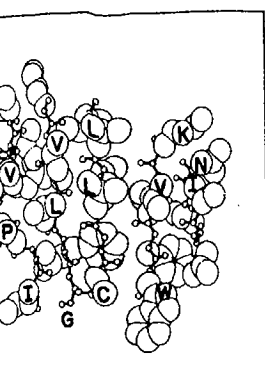


Fig. 8. Glyceraldehyde-3-P Dehydrogenase (AM4.1.2.2.1). Pure β . The central residues form the remarkable cooperative stretch VVIWV clear across the sheet. In the upper row I, V, I, Y, and K form a cooperative string of hydrophobic contacts. Note particularly the hydrophobic nature of the contact between Tyr and Lys. The lower edge of the sheet consists of a string of polar side groups.



```

231 10 41 61
+ + + + +
V G C V
. . . . .
F V V I
. . . . .
F T
+ + + + +
228 6 37 59

```

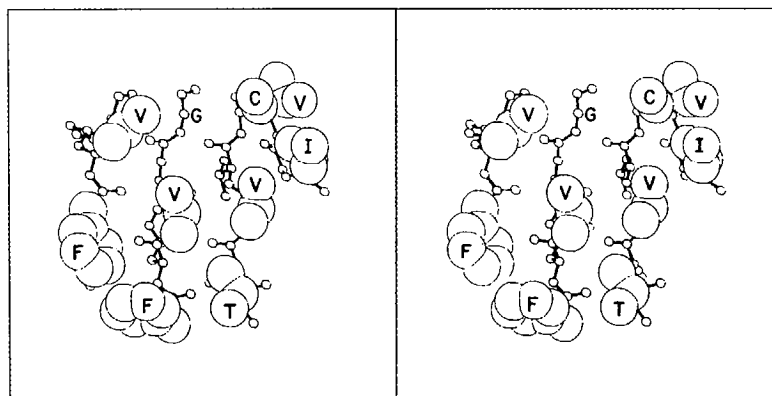


Fig. 9. Triose Phosphate Isomerase (AM11.1.1.1.1). Pure β . Here the C^{β} -branched side groups (4xV, I, T) have nearly identical conformation relative to the backbone. At the top right and bottom left Cys-Ile and Phe-Phe form strong diagonal (i, j+2) contacts. The CH_3 group of Thr is in contact with two Val's while its polar OH group points away from the hydrophobic core patch.

two figures show two sides
ged, but only those on the
phobic patch dominated by
gonal contacts, while all
) point away from the
hole left by Gly 53, but
ighbours lie exposed.
g. 6) form part of a 'super'-
by two residues. The variety
of β -sheet folding.

52 62
 105 191
 † † † †
 F I I A

 L L I L

 Y I L I
 † † † †
 48 66
 109 195

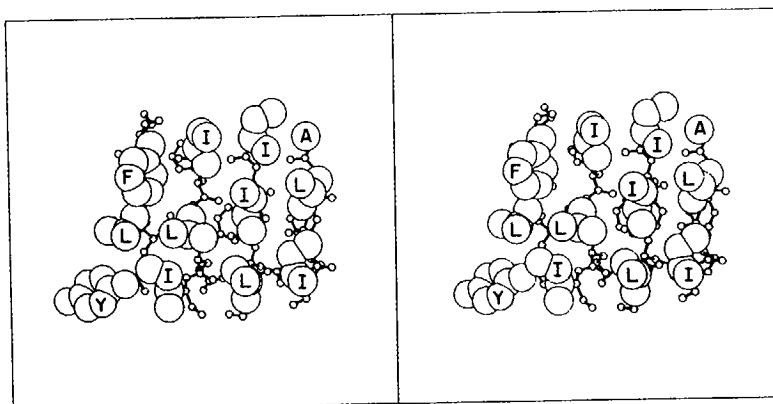


Fig. 10. Carboxipeptidase A (AM3.5.1.1.1). Mostly β_p , but one β_A strand pair. The complete non-polar top surface shows remarkable packing of Leu and Ile residues, both as nearest neighbours (i,j) and diagonally ($i,j+2$), but their side chain conformations are varied. The β_A strand pair on the left is strongly twisted, packing Phe 52 against both Ile 105 and Leu 107 on the neighbour strand in characteristic trigonal fashion.

118 311 254
 316 218
 † † † † †
 T Y L D

 Y T A V

 A I V L

 V Y L
 † † † † †
 116 305 256
 322 212

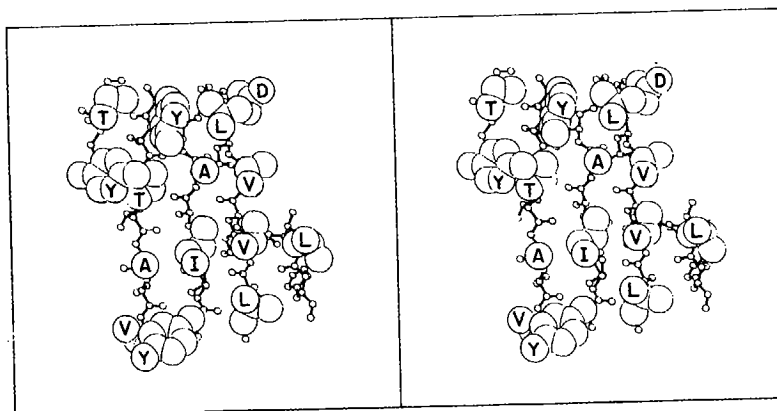
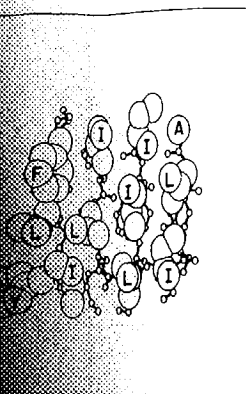


Fig. 11. Prealbumin (AM12.9.1.1.1). Mostly β_A , but one β_p strand pair. Another example of Leu-Val-Ile cooperativity. The top-right to left-bottom trough is deepened by the small size of two Ala's and hemmed in by a wall of Tyr's. Unusual diagonal packing ($i,j-2$) different from the usual ($i,j+2$) is made between Tyr 116 and Tyr 316, in spite of the large distance between their C^α atoms.



91 113⁵⁴50
 ↓ ↑ ↓ ↑

 I L V L

 V A T G
 . .
 ↓ ↑ ↓ ↑
 95 109⁵⁷47

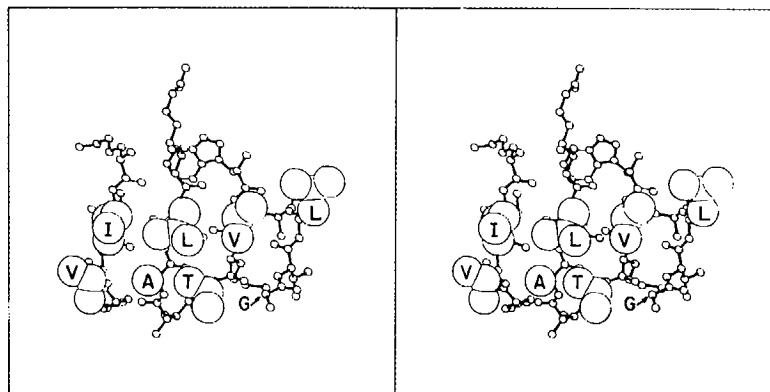
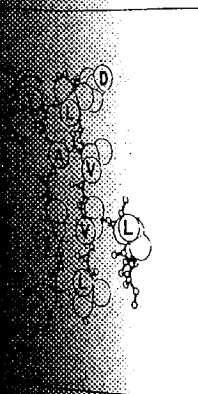


Fig. 12. Tosyl Elastase (AM3.1.3.1). Pure β_A . Interesting alignment of Ile, Val, Leu. The presence of Gly is accompanied by a strong twist. The triangular packing of Leu 112, Val 55 and the non-polar part of Thr 57 appears to allow little local variation in the twist of the sheet.



12 72⁹⁴97
 ↑ ↓ ↑ ↓
 A I A K

 E V I V

 F A
 ↑ ↓ ↑ ↓
 9 76⁹⁰99

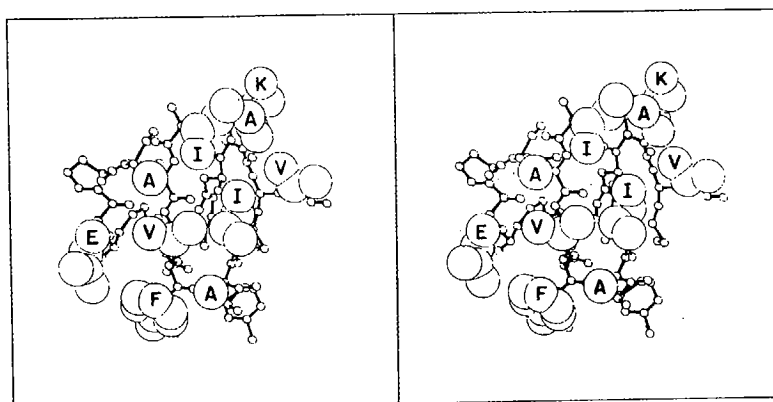


Fig. 13. Bacterial Nuclease (AM5.2.1.1). Pure β_A . The strong twist of this sheet is consistent with the strong diagonal $(i, j+2)$ packing, and seems enhanced by the small size of the three Ala residues.

3. While the mutual juxtaposition, or geometric orientation, of the pairs among V, L, I are rather varied, one recognizes a trend for a special role played by the atom C_β , which is often in close contact with some of the atoms of the other group. C_β is the only atom of side chains whose position is fixed relative to the residue's backbone. Also, bifurcation of the chain at the C_β carbon enhances close packing of the hydrophobic side chains. Is this related to the low correlation of Leu-Leu (0.6) which is the only pair among the three residues with bifurcation solely located at C_γ ? At present such a conjecture is speculative, but further investigation of this question may be of interest.

4. One observes clearly the horizontal lines of neighbouring residues across the stereo-pictures, perpendicular to the β -strands. However, in addition to the horizontal lines one also observes oblique lines, from top-left to bottom-right on the side of the sheet facing the viewer. These represent neighbour interactions among the pairs $(i, j+2)$, where $j+2$ is (disregarding for the moment the common N to C terminal enumeration of residues) the residue above j (above = toward top of page), when j is on the left of i , or alternatively that below j , when j is on the right. A mnemonic for this is the capital letter N: the two vertical bars are the backbones of neighbouring strands, the diagonal line connects the side chains of residue i and $j+2$ above the plane of the paper. These lines are a direct consequence of the right-handed twist of β -sheets, analyzed by Chothia⁽⁸⁾. Such a twist brings i and $j+2$ closer together, while pushing j and $i+2$ away from each other. It seems that while the origin of the right-handed twist has been correctly related by Chothia to the inherent asymmetry of the (ϕ, ψ) maps of L-amino acids, the interaction of the $(i, j+2)$ pairs affect the extent and the detailed geometry of this twist. The pair interactions among (i, j) and $(i, j+2)$ pairs together appear to be the major contributors to the hydrophobic close packing of the surface of many β -sheets. One may therefore expect a pair correlation to exist among $(i, j+2)$ pairs, similarly to (i, j) pairs, which is a worthwhile subject of further investigation.

5. The close packing of sidechains produces a surface on each side of the sheet. Due to the twist of the sheet, this surface generally has a gentle trough running from bottom left to top right⁽⁹⁾, like the diagonal of an inverted capital N: **N**. An example is Fig. 116. In some proteins the trough is accentuated by small residues in the depression (Gly, Ala) and/or large residues (Phe, Tyr, Ile) in the walls (Fig. 11). In other sheets the trough is filled in by large sidechains in the depression, altering radically the shape of the surface the β -sheet presents to other parts of the protein.

The most interesting aspect of our analysis, so far, is the use of statistics (pair correlations) to pinpoint interesting questions, followed by a look at these questions by other means (stereo pictures, packing considerations). By this method we have obtained indications for individual residue-residue recognition among the most prolific β -residues and for a connection between packing of side groups and the twist of β -sheets. The quantitative method of analyzing the concept of molecular recognition used here may be of more general applicability.

References

1. Lifson, S. & Sander, C. (1979). In *Molecular Mechanisms of Biological Recognition*, Ed. M. Balaban, Elsevier, Amsterdam, pp.145-156.
2. Lifson, S. & Sander, C. (1979). *Nature*, in press.
3. Lifson, S. & Sander, C. *J.Mol.Biol.*, submitted.
4. Feldmann, R.J., *AMSOM Atlas of Molecular Structures on Microfiche*, U.S. National Institutes of Health (1976) and updates as of August 1978.
5. von Heijne, G. & Blomberg, C. (1977). *J.Mol.Biol.* 117, 821-824.
6. von Heijne, G. & Blomberg, C. (1978). *Biopolymers* 17, 2033-2037.
7. Motherwell, S., *PLUTO*, a Program for Plotting Molecular and Crystal Structures, University Chemical Laboratory, Cambridge, England.
8. Chothia, C. (1973). *J.Mol.Biol.* 75, 295-302.
9. Cohen, F.E., Sternberg, M.J.E. & Taylor, W. (1979). This volume, p. 131. Figure 4.
10. Richardson, J.S., Getzoff, E.D. & Richardson, D.C. (1978). *Proc.Natl. Acad.Sci. USA* 75, 2574-2578.

DISCUSSION

Richardson, Durham:

The question of packing at β carbons is very interesting, but more complicated because it is quite different for β_p and β_A sheets. In parallel sheets branched β carbons next to each other are quite favored, whereas in antiparallel there is a strong tendency to alternate branched with unbranched β carbons. This difference showed up quite beautifully on your stereo slides. The effect can be explained fairly easily, I think, in terms of packing; the strongly preferred conformation for a branched β carbon side chain allows them to "cup" neatly against each other in the β_p case, but puts them either back-to-back or face-to-face in the β_A case.

Lifson, Rehovot:

This is an interesting observation. Indeed the parallel and antiparallel aliphatic side chains Val, Ile and Leu do pack differently in β_p and β_A . However, their side chain conformations are quite varied in both β_p and β_A , so it is difficult to state a rule. The pair correlations between Val, Ile and Leu, Ala (Tables 3,4) agree with your hypothesis in 4 out of 7 cases.

Scheraga, Ithaca:

Did you say that the need to pack side chain non-polar groups properly necessarily requires a right-handed twist of the β -sheet; i.e. is this the reason for the right-handed twist? Cannot such packing be accommodated by a left-handed twist?

Lifson, Rehovot:

In principle it can. What I said was that the handedness of the twist is due to the chirality of the side chain backbone interactions, as suggested by Choithia⁽⁸⁾, but that the packing of side chains as observed in β -sheets appears to stabilize the twist.

Sander, Rehovot:

About the Ile-Ala preferred pair in β_A : the Ile side chain has one long arm and in one of the stereo pictures you can see how this arm reaches over to make contact with the C^β of Ala.

One should be somewhat careful in analyzing side chain conformations, as X-ray crystallography cannot always give the precise orientation.

One comment about tertiary strand-strand versus tertiary sheet-sheet interactions. Statistically, both can give rise to preferred pairings between residues on neighbour strands; in the first case, the recognition is a direct one; in the second case, the recognition is indirect, i.e. mediated by the environment of the contacting sheet. It will be interesting to see to what extent one or the other recognition process dominates, if any.