ANTHONY ARTALE
MED LIB NATHAN CUMMINGS CTR (S-46)
MEMORIAL SLOAN KETTERING CANCER CTR
1275 YORK AVENUE
NEW YORK, NY 10021
UNITED STATES

Client Number:	DDS36862
Title:	**MOLECULAR MECHANISMS OF BIOLOGICAL RECOGNITION : PROCEEDINGS OF THE SIXTH AHARON KATZIR-KATCHALSKY CONFERENCE, IN CONJUNCTION WITH THE MINERVA SYMPOSIA IN BIOLOGY, GOTTINGEN AND BRAUNLAGE/HARZ, SEPTEMBER 24-30, 1978**
Author:	CRAMER, FRIEDRICH,
DB Ref. No.:	IRN10645767
Date:	1979
Pages:	145–156
Article Title:	ON THE MUTUAL RECOGNITION OF STRANDS
Article Author:	LIFSON, S.
Report Number:	IRN10645767
Publisher:	ELSEVIER/NORTH-HOLLAND BIOMEDICAL PRESS,

Molecular Mechanisms of Biological Recognition
Miriam Balaban, Editor

ON THE MUTUAL RECOGNITION OF STRANDS IN β-SHEETS

Shneior Lifson and Christian Sander
Chemical Physics Department, The Weizmann Institute of Science, Rehovot, Israel
Institut für theoretische Physik, Universität Heidelberg

When β-sheets are formed as part of the folding of proteins, neighboring strands in the sheet come to lie either antiparallel or parallel to each other. In both cases, the precise alignment of any two neighboring strands is determined uniquely. Strand-strand interaction, the position of reverse turns and packing of the sheet in its environment are likely determinants of this unique pairing of strands. The question we ask here is: To what extent is the strand-strand pairing and alignment a specific recognition process between the two strands? As the polypeptide backbone is uniform and all specific information is carried by the sequence of side-chains, the question becomes: Do the amino acid residue pairs between the two strands contribute cooperatively to the folding process in that the members of each pair recognize each other by their particular side-chains?

In the following we show how statistical methods may be applied in search for answers to the above question, what the difficulties are, and possible ways of overcoming them. A more detailed discussion of the subject, including some answers to questions left open here, will soon be submitted for publication.

Any recognition process involved in the unique folding of proteins is, at least in part, specific and particular to each protein. Specificity is determined by the protein's evolutionary history, and is designed to fulfil a vital function in the organism to which the protein belongs. Only such recognition capabilities of residue pairs which are common to all or at least many β-sheets may be revealed by a statistical analysis. Our question should therefore be stated more restrictively: Are there residue-residue pair correlations common either to all β-sheets or to distinct classes of β-sheets?

In the final β-sheet structure, only side-chains on the same side of the sheet can interact specifically (excluding possible long-range electrostatic interactions). Such pairs on different strands in contact with each other (see Fig. 1) can be of the type (i,j) - nearest neighbors - or (i,j±2) - next nearest neighbors. Of these, stronger interactions are likely between nearest neighbors. In what follows, 'pair' therefore always refers to nearest neighbor pairs of the type i,j.

A detailed classification of residue pairs in β strands must distinguish,

first, between antiparallel (β_A) and parallel (β_P) strands, since they are different in the mode of interaction of both their backbones and side groups. One might expect, therefore, substantial differences in the residue-residue pair correlations in β_A and β_P, which would be smeared out if data on β_A and β_P are lumped together. A further, more refined classification should involve the two distinct positions of consecutive residues in both β_A and β_P strand pairs. In β_A strands, the residue pairs alternate as hydrogen bonded pairs and non-bonded pairs. In β_P strands one of the residues in a pair is hydrogen bonded and the other one is non-bonded (see Fig. 1).

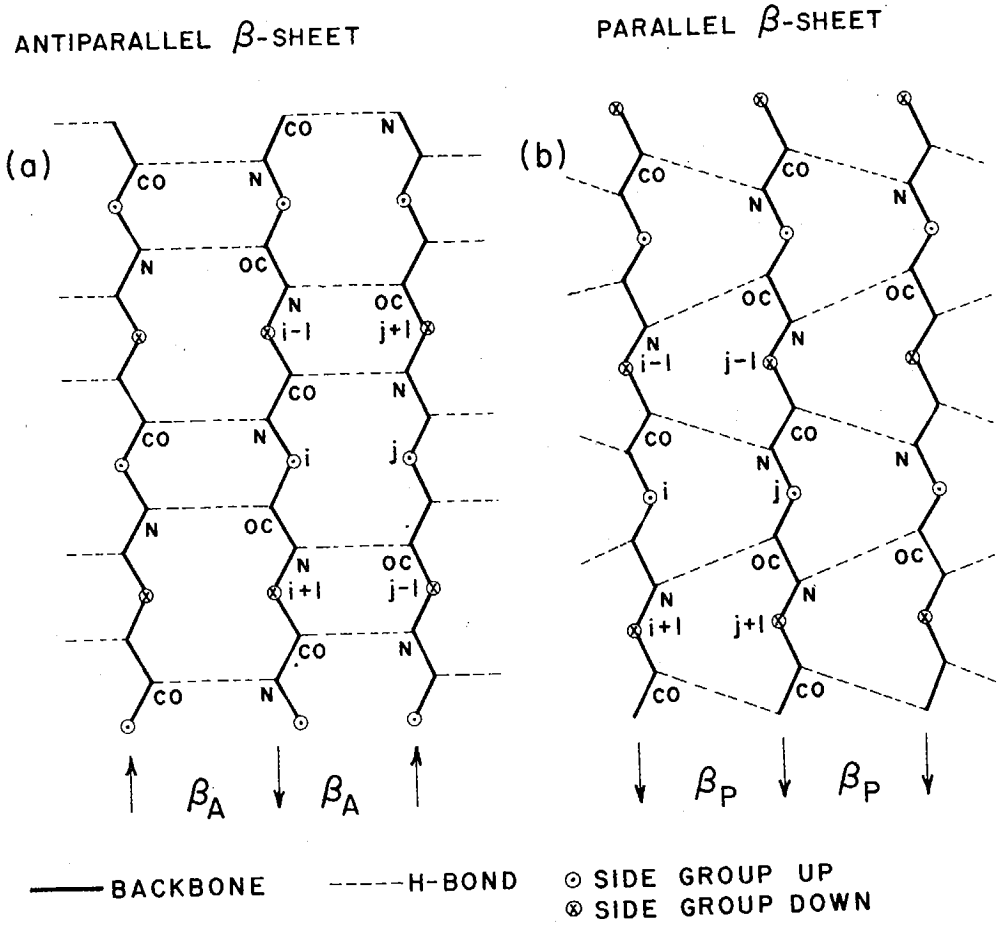

Fig. 1. A schematic representation of antiparallel β-sheets (a) and parallel β-sheets (b).

In the present study we adopt the distinction between antiparallel and parallel strand-pairs, but neglect the more refined classification, because of the smallness of the data base. As we shall see, even at this level of classification the average pair counts are low, necessitating special precautions in the statistical analysis.

Our data base contains 30 different proteins, selected from the AMSOM files, kindly supplied to us by Richard Feldman[1] of the U.S. National Institutes of Health. We wrote a computer program which scans the crystal coordinates and identifies β_A and β_P strand pairs according to geometrical criteria based on the relative positions of the C_α, C_β, N and O atoms of neighboring amino acid residues; details of the criteria are available and will be published elsewhere. The list of the proteins, their residue pair counts and strand pair counts is given in Table I.

We now briefly introduce the statistical analysis of the pair correlations.

Let N_{XY} denote the number of residues of type X which make a pair contact with type Y, where X, Y is the generic notation for the 20 different amino acid residues. $N_X = \sum_Y N_{XY}$ gives the number of contacts made by residues of type X. Note that according to this definition a residue on an inside strand, which has two neighbor strands, is counted twice since it makes two contacts, while a residue on the edge of a β-sheet is counted only once. Thus our definition of β-sheet residue (contact) counts differs from that used by authors interested only in β-strands as secondary structures, rather than in β-sheets as tertiary structures. Let N denote the total number of residue contacts, $N = \sum_X N_X$, equal to twice the number of observed pairs. Then if there were no correlations between residues, the expected number of pair contacts X-Y would simply be proportional to the single residue contact counts N_X and N_Y, namely $E_{XY}=N_X N_Y/N$. We may therefore define *residue-residue* pair correlations g_{XY} as a measure of the deviation of the actual pair distribution N_{XY} from a hypothetical uncorrelated distribution, by

$$g_{XY} = N_{XY}/E_{XY} = N_{XY}\, N/N_X N_Y$$

When $g_{XY}>1$, X and Y are favorably correlated (or simply 'correlated'), while if $g_{XY}<1$, X and Y are unfavorably correlated (or 'anticorrelated').

We further define a cooperative measure of *strand-strand* pair correlation g_s as the product of the g_{XY} of all the pairs of residues constituting the pair of strands

$$g_s = g_{X_1Y_1} g_{X_2Y_2} \cdots g_{X_sY_s}$$

TABLE I

Antiparallel and parallel residue pair counts and strand pair counts in the 30 proteins used as data base

Amsom identifier	Residue pairs β_A	Residue pairs β_P	Strand pairs β_A	Strand pairs β_P	Protein name
AM 1.2.1.1.1	5	4	1	1	Bovine ferricytochrome B5
AM 1.3.1.1.1	3	-	1	-	Tuna ferrocytochrome C
AM 1.3.2.1.1	3	-	1	-	Bacterial ferricytochrome C2
AM 2.1.1.2.1	8	-	2	-	Bacterial rubredoxin
AM 2.2.1.1.1	10	-	3	-	Bacterial high potential iron protein
AM 3.1.1.1.1	14	35	4	8	Subtilisin BPN'
AM 3.1.2.08.1	44	-	10	-	Bovine trypsin
AM 3.1.2.3.1	63	3	15	1	Bovine chymotrypsinogen A
AM 3.1.3.1.1	58	-	12	-	Procine tosyl elastase
AM 3.2.1.8.1	29	-	7	-	Papain
AM 3.4.1.1.1	33	13	7	3	Bacterial thermolysin
AM 3.5.1.1.1	14	15	3	3	Bovine carboxypeptidase A complex
AM 3.6.1.3.1	10	-	2	-	Bovine trypsin inhibitor
AM 4.1.1.3.1	7	19	2	4	Dogfish apo-lactate dehydrogenase
AM 4.1.2.2.1	20	35	4	7	Lobster glyceraldehyde-3-P dehydrogenase "red"
AM 4.1.3.3.1	28	13	8	3	Horse alcohol dehydrogenase complex
AM 4.2.1.2.1	-	26	-	5	Bacterial semiquinone flavodoxin
AM 5.1.1.1.1	26	-	6	-	Bovine ribonuclease S complex
AM 5.2.1.1.1	29	3	7	1	Bacterial nuclease complex
AM 6.1.2.1.1	37	4	7	1	Human bence-jones protein "Rei"
AM 6.1.2.2.1	90	5	17	1	Human immunoglobulin G Fab' "new"
AM 8.1.1.2.1	-	18	-	4	Procine adenylate kinase
AM 9.1.1.2.1	99	-	17	-	Jack bean concanavalin A
AM 10.1.1.2.1	8	-	2	-	Chicken lysozyme
AM 10.1.2.1.1	3	-	1	-	Bacteriophase T4 lysozyme
AM 11.1.1.1.1	-	37	-	7	Chicken triose phosphate isomerase (monomer 1)
AM 12.01.1.1.1	3	-	1	-	Carp calcium-binding protein B
AM 12.03.1.1.1	53	17	12	4	Human carbonic anhydrase B
AM 12.03.1.2.1	51	9	11	2	Human carbonic anhydrase C
AM 12.09.1.1.1	40	7	7	1	Human prealbumin
	788	263	170	56	T o t a l

TABLE 2. Pair counts $N_{\chi\gamma}$ and their uncorrelated values $E_{\chi\gamma}$ (top right); correlations $g_{\chi\gamma}$ and their differential errors $\Delta g_{\chi\gamma}$ (bottom left) in β_A strands

	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W	
	0 4.2	0 1.5	3 2.0	4 2.4	7 6.1	3 2.1	4 3.2	5 6.3	4 7.1	4 4.3	3 2.3	3 2.3	12 9.7	2 5.8	2 2.2	2 2.3	10 7.2	4 3.5	6 5.0	3 2.4	G Gly
		2 0.5	0 0.7	1 0.9	2 2.2	0 0.8	3 1.2	2 2.3	5 2.6	1 1.5	2 0.8	0 0.8	4 3.5	2 2.1	1 0.4	0 0.8	3 2.6	1 1.3	0 1.8	0 0.9	P Pro
			2 1.0	0 1.2	3 2.9	3 1.0	3 1.6	3 3.0	5 3.4	5 2.1	2 1.1	2 1.1	1 4.7	3 2.8	0 0.6	0 1.1	1 3.5	1 1.7	1 2.4	1 1.2	D Asp
G Gly	0.0 0.5			2 1.4	3 3.5	2 1.2	4 1.9	1 3.7	8 4.1	4 2.5	4 1.3	3 1.3	2 5.6	1 3.3	0 0.7	0 1.3	2 4.2	2 2.0	4 2.9	0 1.4	E Glu
P Pro	-		-		4 9.0	1 3.1	4 4.8	11 9.3	13 10.5	8 6.3	0 3.4	3 3.3	10 24.2	18 8.5	3 1.8	4 3.4	10 10.6	5 5.1	7 7.4	3 3.5	A Ala
D Asp	1.5 0.5	-	-			0 1.1	2 1.6	5 3.2	0 3.6	1 2.2	2 1.2	1 1.2	3 4.9	2 2.9	0 0.6	1 2.2	5 3.7	3 1.8	3 2.5	4 1.2	N Asn
E Glu	1.7 0.4	-	-	-			0 2.5	6 4.9	6 5.6	5 3.3	3 1.8	2 1.8	4 7.5	2 4.5	1 1.0	3 1.8	2 5.6	5 2.7	4 3.9	0 1.9	Q Gln
A Ala	1.1 0.2	0.9 0.5	1.0 0.3	0.8 0.3	0.4 0.2			14 9.6	21 20.8	6 6.5	5 3.5	5 3.4	19 24.7	4 8.7	3 2.9	3 3.5	5 11.0	4 5.3	1 7.6	0 3.7	S Ser
N Asn	1.4 0.5	-	3.0 1.0	1.6 0.8	0.3 0.3	-			18 12.3	8 7.3	5 4.0	2 3.9	6 16.6	9 9.9	3 2.1	2 4.0	12 12.4	2 6.0	9 8.6	1 4.1	T Thr
Q Gln	1.2 0.3	2.6 0.9	1.9 0.6	2.1 0.5	0.8 0.2	1.2 0.6	-			6 4.4	3 2.4	1 2.3	8 9.9	7 5.9	1 1.3	1 2.4	4 7.4	2 3.6	5 5.2	3 2.5	K Lys
S Ser	0.8 0.2	0.9 0.4	1.0 0.3	0.3 0.3	1.2 0.1	1.6 0.3	1.2 0.2	1.5 0.2			0 1.3	0 1.3	5 5.4	1 3.2	0 0.7	0 1.3	1 4.0	2 1.9	2 2.8	5 2.3	R Arg
T Thr	0.6 0.1	2.0 0.4	1.5 0.3	1.9 0.2	1.2 0.1	0.0 0.3	1.1 0.2	1.9 0.1	1.5 0.2				6 1.2	4 5.2	4 3.1	1 0.7	0 1.3	4 3.9	0 1.9	2 2.7	H His
K Lys	0.9 0.2	-	2.4 0.5	1.6 0.4	1.3 0.2	0.5 0.5	1.5 0.3	0.9 0.3	1.1 0.2	1.4 0.5			26 22.4	23 13.4	1 2.9	6 5.4	25 16.8	13 5.1	11 11.7	5 5.6	V Val
R Arg	1.3 0.4	2.4 1.2	1.8 0.9	3.0 0.7	0.0 0.3	1.7 0.9	1.7 0.6	1.4 0.3	1.3 0.3	1.3 0.4	-			8 8.0	0 1.7	0 3.2	12 10.0	4 4.8	5 7.0	5 3.3	I Ile
H His	1.3 0.4	-	1.8 0.9	2.3 0.8	0.9 0.3	-	1.1 0.6	1.5 0.7	0.5 0.3	0.4 0.4	-	4.9 1.6			0 0.4	1 0.7	5 2.1	0 1.0	1 2.5	1 0.7	M Met
V Val	1.2 0.1	1.2 0.3	0.2 0.2	0.4 0.2	0.7 0.4	0.6 0.2	0.5 0.2	1.3 0.2	0.4 0.2	0.8 0.2	0.9 0.2	0.8 0.2	1.2 0.1			6 1.3	6 4.0	2 1.9	5 2.8	3 1.3	C Cys
I Ile	0.3 0.2	1.0 0.5	1.1 0.4	0.3 0.3	2.1 0.2	0.7 0.3	0.4 0.2	0.5 0.1	0.9 0.1	1.2 0.2	0.3 0.3	1.3 0.3	1.7 0.2	1.0 0.3			8 12.6	6 6.1	12 8.8	8 4.2	L Leu
M Met	1.6 0.8	-	-	-	1.7 0.6	-	-	1.6 0.5	1.4 0.5	-	-	-	0.3 0.3	-	-			4 2.9	6 4.2	2 2.0	F Phe
C Cys	0.9 0.4	-	-	-	1.2 0.3	-	1.7 0.6	0.9 0.3	0.5 0.3	0.4 0.4	-	-	1.1 0.2	0.0 0.3	-	4.7 1.6			12 6.1	2 2.9	Y Tyr
L Leu	1.4 0.1	1.2 0.4	0.3 0.3	0.5 0.2	0.9 0.1	1.4 0.3	0.4 0.2	0.5 0.1	1.0 0.1	0.5 0.1	0.2 0.2	1.0 0.3	1.5 0.1	1.2 0.1	2.3 0.5	1.5 0.2	0.6 0.2			0 1.4	W Trp
F Phe	1.1 0.3	-	-	1.0 0.5	1.0 0.2	1.7 0.6	1.8 0.4	0.8 0.2	0.3 0.2	0.6 0.2	1.0 0.5	-	1.6 0.1	0.8 0.2	-	1.0 0.5	1.0 0.2	1.4 0.7			
Y Tyr	1.2 0.2	-	0.4 0.4	1.4 0.3	0.9 0.1	1.2 0.4	1.0 0.3	0.1 0.1	1.0 0.1	1.0 0.2	0.7 0.4	0.7 0.4	0.9 0.1	0.7 0.1	-	1.8 0.4	1.4 0.1	1.4 0.2	2.0 0.3		
W Trp	1.2 0.4	-	-	-	0.8 0.3	3.3 0.8	-	0.0 0.3	0.2 0.2	1.2 0.4	3.7 0.7	-	0.9 0.2	1.5 0.3	-	2.2 0.7	1.9 0.2	1.0 0.5	0.7 0.3	-	
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W	Sum
Total (N_χ)	81	29	39	47	119	41	63	123	139	83	45	44	188	112	24	45	141	68	98	47	1576
Freq, %	5.1	1.8	2.5	3.0	7.6	2.6	4.0	7.8	8.8	5.3	2.9	2.8	11.9	7.1	1.5	2.9	8.9	4.3	6.2	3.0	100.0

TABLE 3. Pair counts N_{XY} and their uncorrelated values E_{XY} (top right); correlations g_{XY} and their differential errors Δg_{XY} (bottom left) in β_p strands

	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W	
	0 2.7	1 0.6	0 1.0	2 1.1	4 3.2	2 0.9	1 0.4	0 2.4	0 1.5	2 1.6	0 0.7	0 0.4	7 7.8	3 4.6	2 0.8	3 0.8	3 3.8	4 1.7	3 1.5	1 0.6	G Gly
		0 0.1	0 0.2	2 0.2	0 0.7	0 0.2	0 0.1	0 0.5	0 0.3	0 0.3	0 0.2	0 0.1	2 1.6	0 1.0	1 0.2	0 0.2	1 0.8	1 0.4	0 0.3	0 0.1	P Pro
			0 0.4	2 0.4	1 1.2	3 0.3	0 0.1	3 0.9	1 0.6	1 0.6	0 0.3	0 0.1	2 2.9	0 1.7	0 0.3	0 0.3	1 1.4	0 0.6	0 0.6	0 0.2	D Asp
G Gly	-			0 0.4	0 1.3	0 0.3	0 0.1	0 0.9	0 0.6	2 0.6	1 0.3	0 0.1	1 3.1	2 1.8	0 0.3	0 0.3	0 1.5	2 0.7	1 0.6	0 0.2	E Glu
P Pro	-	-			6 3.7	0 1.0	0 0.4	3 2.8	5 1.8	2 1.8	0 0.8	1 0.4	7 9.0	7 5.3	1 0.9	1 0.9	3 4.4	0 2.0	2 1.8	1 0.7	A Ala
D Asp	-	-	-			0 0.3	0 0.1	1 0.8	0 0.5	1 0.5	0 0.2	0 0.1	2 2.5	0 1.4	0 0.3	0 0.3	0 1.2	2 0.5	1 0.5	0 0.2	N Asn
E Glu	1.8 0.9	8.8 4.4	5.0 2.5	-			0 0.0	1 0.3	0 0.2	0 0.2	0 0.1	0 0.0	1 1.0	1 0.6	0 0.1	0 0.1	0 0.5	0 0.2	1 0.2	0 0.1	Q Gln
A Ala	1.3 0.3	-	-	-	1.6 0.5			6 2.1	2 1.3	5 1.4	0 0.6	0 0.3	5 6.8	1 4.0	2 0.7	0 0.7	2 3.3	0 1.5	2 1.3	0 0.5	S Ser
N Asn	2.3 1.2	-	9.4 3.1	-	-	-			0 0.8	0 0.9	2 0.4	1 0.2	3 4.3	1 2.5	2 0.4	1 0.4	1 2.1	2 1.0	0 0.8	0 0.3	T Thr
Q Gln	-	-	-	-	-	-	-			0 0.9	1 0.4	0 0.2	2 4.5	1 2.6	0 0.5	1 0.5	2 2.2	0 1.0	2 0.9	0 0.3	K Lys
S Ser	0.0 0.4	-	3.4 1.1	-	1.1 0.4	-	-	2.9 1.0			0 0.2	1 0.1	1 2.1	0 1.2	0 0.2	0 0.2	2 1.0	2 0.5	0 0.4	0 0.2	R Arg
T Thr	-	-	-	-	2.8 0.6	-	-	1.5 0.8	-			0 0.0	1 1.0	0 0.6	0 0.1	0 0.1	1 0.5	0 0.2	0 0.2	0 0.1	H His
K Lys	1.3 0.6	-	-	3.2 1.6	1.1 0.5	-	-	3.6 0.7	-	-			34 22.2	18 12.9	0 2.3	2 2.3	11 10.9	4 4.9	2 4.3	3 1.6	V Val
R Arg	-	-	-	-	-	-	-	-	5.0 2.5	-	-			10 7.5	2 1.3	1 1.3	12 6.3	2 2.9	2 2.5	0 1.0	I Ile
H His	-	-	-	-	-	-	-	-	-	-	-	-			0 0.2	0 0.2	1 1.1	0 0.5	0 0.4	0 0.2	M Met
V Val	0.9 0.1	1.2 0.6	0.7 0.3	0.3 0.3	0.8 0.1	0.8 0.4	-	0.7 0.1	0.7 0.2	0.4 0.2	0.5 0.5	-	1.5 0.1			0 0.2	0 1.1	1 0.5	0 0.4	1 0.2	C Cys
I Ile	0.7 0.2	-	-	1.1 0.6	1.3 0.2	-	-	0.3 0.3	0.4 0.4	0.4 0.4	-	-	1.4 0.1	1.3 0.3			8 5.3	2 2.4	3 2.1	0 0.8	L Leu
M Met	2.5 1.3	-	-	-	-	-	-	2.9 1.4	4.6 2.3	-	-	-	0.0 0.4	1.5 0.8	-			0 1.1	0 1.0	2 0.4	F Phe
C Cys	3.8 1.3	-	-	-	-	-	-	-	-	-	-	-	0.9 0.4	-	-	-			2 0.8	0 0.3	Y Tyr
L Leu	0.8 0.3	-	-	-	0.7 0.2	-	-	0.6 0.3	0.5 0.5	0.9 0.5	2.0 1.0	-	1.0 0.1	1.9 0.2	-	-	1.5 0.4			0 0.1	W Trp
F Phe	2.3 0.6	-	-	2.9 1.5	0.0 0.5	3.7 1.8	-	-	2.1 1.0	-	4.4 2.2	-	0.8 0.2	0.7 0.3	-	-	0.8 0.4	-			
Y Tyr	2.0 0.7	-	-	-	1.1 0.6	-	-	1.5 0.8	-	2.3 1.1	-	-	0.5 0.2	0.8 0.4	-	-	1.4 0.5	-	-		
W Trp	-	-	-	-	-	-	-	-	-	-	-	-	1.8	-	-	-	-	5.5	-	-	
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W	Sum
Total(N_X)	38	8	14	15	44	12	5	33	21	22	10	5	108	63	11	11	53	24	21	8	526
Freq. %	7.2	1.5	2.7	2.9	8.4	2.3	1.0	6.3	4.0	4.2	1.9	1.0	20.5	12.0	2.1	2.1	10.1	4.6	4.0	1.5	100.0

where $X_i Y_i$ (i=1...s) are the residue pairs of the two strands. When $g_s>1$ (or <1) the two strands are favorably (or unfavorably) correlated as compared to strands with uncorrelated residue pairs. These definitions hold equally well for antiparallel and parallel strands.

The results of scanning the β-sheets of all proteins of Table 1 are given in Tables 2 and 3 for antiparallel and parallel β-strands respectively. The actual counts N_{XY} and their corresponding uncorrelated values E_{XY} are given at the top right side of the main diagonal; the pair correlations g_{XY} and their differential errors Δg_{XY} are given at the bottom left side. The $\Delta g_{XY}=1/E_{XY}$ give the change which would occur in g_{XY} if the count N_{XY} would change by 1, assuming E_{XY} to be unaffected. For X=Y, $\Delta g_{XY}=2/E_{XY}$. The differential error measures the relative significance of g_{XY} values: The smaller Δg_{XY} the more significant is g_{XY}. Δg_{XY} as defined here should not be confused with a standard deviation. We omitted from Tables 2 and 3 those g_{XY} and Δg_{XY} for which $N_{XY}<1$ and $E_{XY}<2$ (or twice that for X=Y), since their significance is the lowest. The data necessary to calculate them is still available in the tables, and it is more meaningful to inspect the tables without them.

Some of the g_{XY}'s are quite far from the uncorrelated value 1. In antiparallel strands, for example, $g_{Val,Leu}=1.5$ and $g_{Val,Ile}=1.7$. These two are among the most significant correlations, since they are obtained from among the highest pair counts. In general, hydrophobic-hydrophobic pairs are favorably correlated, as observed already by Heijne and Blomberg[2,3], however, we find differences within the hydrophobic group which are possibly significant. For example, the uncorrelated values $E_{Leu,Leu}$ and $E_{Val,Ile}$ in β_A are similar, 12.6 and 13.4, while their actual counts are 8 and 23 respectively. Such differences within a group indicate specific recognition between individual amino acids, rather than recognition between classes of amino acids (e.g. 'hydrophobic' recognition). Another outstanding pair in β_A is SER, THR with g=1.9. Less significant statistically, but quite interesting, are some of the missing counts. Gly, for example, is quite abundant in many pairs, but $N_{Gly,Gly}=0$ in both antiparallel and parallel strands.

We come now to the original question which motivated this study: How do pair correlations contribute to strand-strand recognition. To answer this question we consider the hypothetical pairing of strands which is obtained when two strands are shifted with respect to each other by ±1 or ±2 residues where ± indicates shifts in either direction. A schematic description of the actual pairing and the shifted pairings is given in Fig. 2 for one of the strand pairs of Concanavalin A. We assume that in the process of folding, the strands try

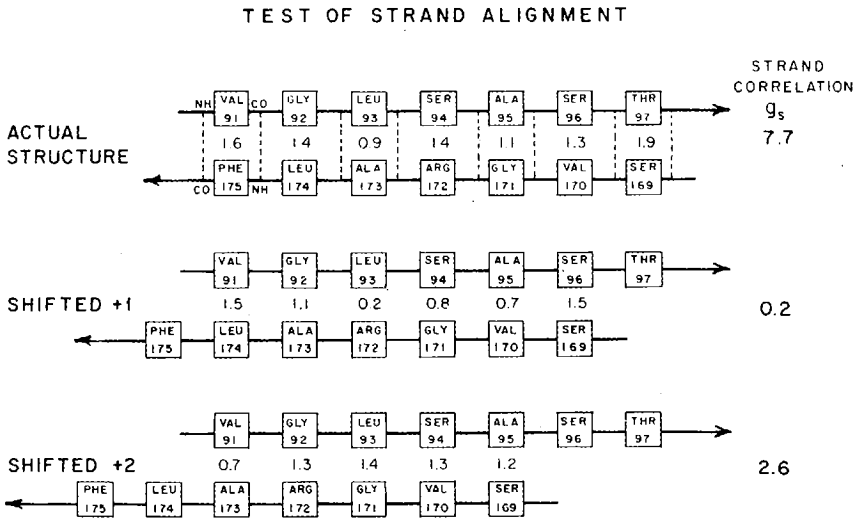

Fig. 2. A schematic representation of strand alignments of an actual structure from Concanavalin A and shifted structures.

various hypothetical pairings such as the shifted ones, but find them insufficiently stable as compared to the actual one. In terms of the strand correlations g_s, we conjecture that the g_s of the actual pairing should be, in general, larger than the g_s of the hypothetical, shifted pairing and larger than the uncorrelated reference value $g_s=1.0$.

In order to examine this conjecture, we calculated the g_s of all antiparallel (as well as parallel) strand pairs which occurred in our data base and then calculated the g_s of the shifted strands. The results are presented by histograms in Fig. 3.

A detailed examination of the six histograms reveals a number of interesting facts. We shall discuss here only the histograms of antiparallel strands, (a)

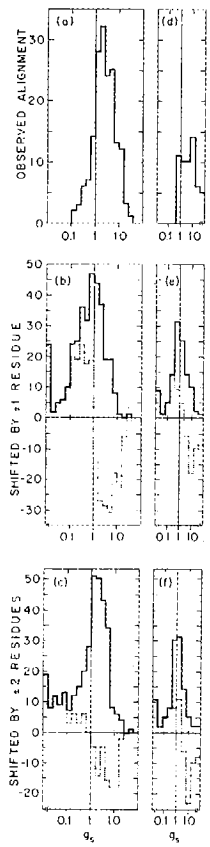

Fig. 3. Histograms of g_s distributions for actual and shifted alignments. Dotted lines represent the difference between the shifted and the actual histograms.

(b) and (c), because they are based on a larger data base. Similar comments may be made, however, with respect to the other three histograms referring to parallel strands. First we note the systematic differences between the three histograms. That of the actual alignment is centered at larger g_s values than the two others, indicating more favorable correlations. Furthermore the histogram of the ±1 shifted strands is centered at lower g_s values than that of the ±2 shifted strands. Our first conclusion is that shifting the strands away from the actual alignment by ±1 residues offers the most unfavorable correlation. This is interesting, since a ±1 shift would require flipping all residues on one strand by 180°, in order to maintain nearest neighbor contacts on both sides of the β-sheet. The ±2 shift, on the other hand, is a pure shift in that it leaves the side groups on the same side of the β-sheet. This distinction between ±1 and ±2 shifts, which is particularly important if one side of the sheet is in a predominantly apolar, the other in a mainly polar environment, seems to be significantly scored, in spite of the smallness of the data base.

One is tempted to draw similar conclusions from comparison of the actual g_s distribution and the ±2 or ±1 hypothetical distribution, since the actual distribution is centered at distinctly larger values of g_s. However, with an average count $N_{XY} \sim 4$ for β_A and $N_{XY} \sim 1$ for β_P the data base is too small for such a comparison. Small data bases tend in general to give favorable correlations for any properties scored within the data base. Thus even if there were no real correlation between the residue pairs, the actual g_s distribution would appear as if there is a favorable correlation. To understand this tendency, note that the random distribution does not give exactly the expected value $E_{XY}=N_X N_Y/N$ but fluctuates around this value. The fluctuations are relative larger, the smaller the data base. By definition, fluctuations toward favorable correlations have

larger counts ($N_{XY}>E_{XY}$) and are therefore weighted more heavily in the strand correlations g_s than fluctuations toward unfavorable correlations ($N_{XY}<E_{XY}$) which appear correspondingly less frequently in the products used to calculate the g_s. This biases any distribution of strand correlations scored with pair correlations g_{XY} derived from the same data base. In particular, the actual distribution of strand correlations g_s contains *no* $g_s=0$ (by definition, since only actually occurring pairs are scored), yet $N_{XY}=0$ may arise as a fluctuation. In contrast, the hypothetical distribution of ±1 and ±2 shifts *do* contain $g_s=0$. Therefore it is not meaningful to compare actual and hypothetical distributions of small data bases without taking into account the bias introduced in the scoring on the same data base from which the correlations were deduced. Thus, some further work is necessary in order to interpret correctly the strand-strand correlations g_s and how they change as the strands are shifted relative to each other.

The following conclusions may be made at this stage:

1) There is tentative evidence for non-random pair correlations in antiparallel and parallel β-strands.

2) The recognition implied by these correlations appears to involve a higher degree of specificity than simply hydrophobic recognition.

3) Nearest neighbor pair correlations are not the only determining factor in fixing β-structures; reverse turns near β-sheets and packing of sheets with other parts of the globular protein are the other most likely contributors.

4) Just as the distinction between antiparallel and parallel β-structures in this type of statistical analysis is useful, so will further distinction between different H-bonding types be interesting.

In summary, we have demonstrated a statistical method of testing an element of tertiary structure, β-sheets, for molecular recognition on the level of individual amino acids. The method allows pinpointing certain amino acid side-chains pairs which preferentially interact in forming double β-strands in sheets, such as pairs involving the side-chains branched at C_β: THR, VAL and ILE. Perhaps the most interesting aspect of the method is the comparison of actual strand alignments with hypothetical ones which might be encountered in a trial and error fashion in protein folding. However, further work on this is necessary, in particular on removal of the bias introduced by the limited size of the data base.

REFERENCES

1. Feldman, Richard J.(1976) AMSOM Atlas of Molecular Structures on Microfiche, and update files as of August 1978.

2. von Heijne, G. and Blomberg, C. (1977) J. Mol. Biol. 117, 821-824.

3. von Heijne, G. and Blomberg, C. (1978) Biopolymers 17, 2033-2037.