Request # 18652570 **NOV 28, 2005**

Ariel To: 64.40.17.85/140.163.217.217

Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center (METRO #146)
1275 York Avenue
New York, NY 10021

DOCLINE: Journal Copy Epayment

Title:	Journal of molecular biology.
Title Abbrev:	J Mol Biol
Citation:	1980 Jun 5;139(4):627-39
Article:	Specific recognition in the tertiary structure of
Author:	Lifson S; Sander C
NLM Unique ID:	2985088R Verify: PubMed
PubMed UI:	7411635
ISSN:	0022-2836 (Print) 1089-8638 (Electronic)
Publisher:	Academic Press, London
Copyright:	Copyright Compliance Law
Authorization:	A Artale
Need By:	DEC 01, 2005
Maximum Cost:	**$25.00**
Patron Name:	Gangi-Dino, Rita
Referral Reason:	Lacking
Library Groups:	BQSIMB,EFTS,RESOURCE,METRO
Phone:	1.212.639-7441
Fax:	1.646.422-2316
Email:	ill@mskcc.org
Comments:	**Please, we prefer as PDF or Ariel. Thank you! EFTS, METRO, BQSI.**
Routing Reason:	Routed to NYUSSY in Serial Routing - cell 1
Received:	Nov 28, 2005 (02:10 PM EST)
Lender:	SUNY Upstate Medical University/ Syracuse/ NY USA (NYUSSY)

Bill to: NYUMSK

Memorial Sloan-Kettering Cancer Center
Medical Library Nathan Cummings Center
1275 York Avenue
New York, NY 10021

J. Mol. Biol. (1980) **139**, 627–639

Specific Recognition in the Tertiary Structure of β-Sheets of Proteins

S. Lifson†

Chemical Physics Department
The Weizmann Institute of Science
Rehovot, Israel

and C. Sander†

Biophysics Department
Max Planck Institute of Medical Research
Heidelberg, Federal Republic of Germany

(Received 25 May 1979, and in revised form 22 January 1980)

The frequency of occurrence of nearest neighbour residue pairs on adjacent antiparallel (β_A) and parallel (β_P) strands is obtained from 30 known protein structures. The specificity of interstrand recognition due to such pairing as a factor in the folding of β-sheets is studied by statistical methods. Residues of sufficiently high count for statistical analysis are treated individually while the rest are combined into small groups of similar size, polarity, and/or genetic exchangeability. The hypothesis of specific recognition between individuals and small groups is contrasted with the alternative hypothesis of non-specific recognition between broad classes (hydrophobic, neutral, polar) of residues. A χ^2 test of pair correlations favours specific recognition against non-specific recognition with a high level of confidence. The largest and most significant correlations are: Ser/Thr (1·9 ± 0·3), Ile/Val (1·7 ± 0·3) and Lys-Arg/Asp-Gln (1·8 ± 0·3) in β_A, and Ile/Leu (1·9 ± 0·4) in β_P. The pair Gly/Gly never occurs in any β-sheet. The specific residue-pair correlations derived here may be useful in statistical prediction methods of protein tertiary structure.

1. Introduction

In globular proteins, *β-strands* are an element of secondary structure, i.e. a regular repeat of backbone conformation. *β-sheets*, on the other hand, are an element of tertiary structure. In the assembly of β-strands into sheets, parts of the protein distant along the polypeptide chain come to lie close to each other. For proteins with a considerable percentage of β structure, once the arrangement of strands within the sheets is fixed, conformational flexibility for the rest of the protein is greatly reduced. Thus, whatever the exact folding pathway, the factors responsible for sheet formation are major determinants of the overall tertiary folding. The most obvious candidates for such factors, each worth studying, are: (1) formation of short loops (reverse turns), which puts strands next to one another. (2) Packing of strands against the surface of neighbouring α-helices or other substructures. (3) Specific pairing of β-strands, as a

† Correspondence may be addressed to either S. L. in Rehovot or C. S. in Heidelberg.

0022-2836/80/160627-13 $02.00/0

result of direct interactions between their residues. Here we focus our attention on the last factor, the residue–residue recognition process (Fig. 1).

Previous attempts at analysing the tertiary structure of β-sheets by statistical means have been of two types. (1) Analysis of sheet topology: Schulz & Schirmer (1974), Richardson (1976,1977), Levitt & Chothia (1976) and Sternberg & Thornton (1977*a*,*b*) have analysed known β-sheets as to: length, direction and number of strands; ordering of strands within the sheet; length, type and handedness of cross-over connections; and statistical significance of the occurrence of folding units containing β-strands. (2) Analysis of the amino acid content of β-sheets: Sternberg & Thornton (1977*c*) observed that the most hydrophobic strands tend to occur at the centre of the sheet and put forward the hypothesis of hydrophobic ordering of strands. von Heijne & Blomberg (1977,1978) analysed pair correlations among hydrophobic, neutral and polar classes of residues. They found that interstrand pairs between residues of the same class occur more often than expected by random chance. They concluded from their observations "that inter- and intrastrand nearest neighbour interactions of a *rather unspecific* character are responsible for the main stabilizing forces in the β-sheet".

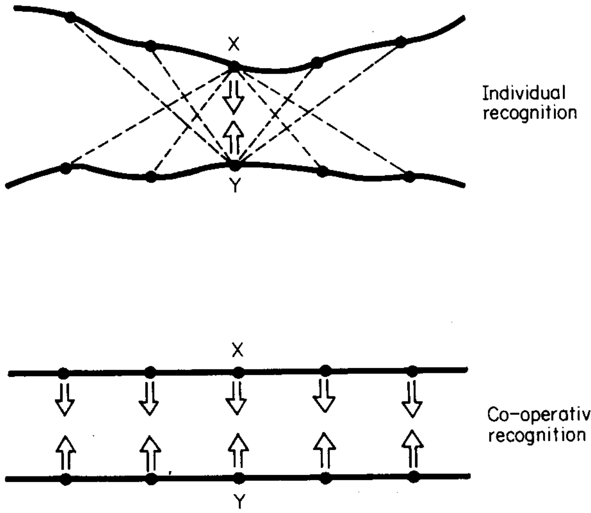

FIG. 1. The concept of individual residue–residue recognition is defined here as preferential pairing of 2 residues: the recognizing partners (→ ←) are more likely to interact with each other than with other candidates (-----). Co-operative strand–strand recognition is the cumulative effect of individual recognition between rows of residues on 2 strands.

In the present study we challenge the above conclusion. We carry the statistical analysis of pair correlations to the level of *specific* individual recognition among the more frequent amino acid residues. Less frequent residues are grouped according to their resemblance in size, structure, polarity and genetic exchangeability (Sander & Schulz, 1979). For antiparallel strands, with a data base of 788 residue pairs, the statistical analysis of pair correlations involves seven individual residues, five groups of two residues each, and one group of three residues. For parallel strands, with a data base of only 263 pairs, the grouping is more extensive.

A χ^2 test of significance is used to analyse the results. We find that the hypotheses of non-specific recognition and of specific recognition are both highly plausible against the "null hypothesis" of random pairing of interstrand residues. We then apply the χ^2 test to the hypothesis that there is specific recognition over and above non-specific recognition, by using the non-specific recognition as the null hypothesis. We find that non-specific recognition can be rejected in favour of specific recognition with a level of confidence of 98·9% for antiparallel strands, 76% for parallel strands and 99·99% for all β strands.

2. Data Base

The data base consists of the atomic co-ordinates selected from data sets generously supplied by Richard Feldmann (1976). The selection comprises only proteins whose atomic co-ordinates include the side-chains. A total of 42 such proteins contained β-sheets. Among these we selected 30 proteins whose β-sheets are significantly different in their sequences and pair counts. They are listed in Table 1.

3. Criteria for β-Strands

Previously available compilations of β-sheets (Feldmann, 1976; Levitt & Greer, 1977) do not always specify which residues make hydrogen bonds and/or have their side groups on the same side of the sheet. We therefore made an independent compilation of β-strands, using criteria which are tailored to the main purpose of this investigation. The criteria are based exclusively on interstrand quantities (tertiary structure), rather than on (ϕ, ψ) backbone angles within each strand (secondary structure). Distinction is made between parallel and antiparallel strands. Residues are considered to be interacting, i.e. making contact, if they fulfil simultaneously three geometrical criteria. Criterion (1) defines neighbours; (2) ensures that the side-chains point in roughly the same direction, i.e. lie on the same side of the β-sheet; (3) determines the hydrogen-bonding type (bonding or non-bonding; parallel or antiparallel) by the relative orientation of their backbone NH and C=O groups (see Fig. 2). A computer program employs the following algorithm.

If $\mathbf{C}_\alpha(i)$ is the position vector of the C$_\alpha$ atom of residue i, and similarly $\mathbf{C}_\beta(i)$ the position of the side group C$_\beta$ atom (for glycine the geometry of alanine was assumed) etc., then the three criteria are fulfilled for two residues i and j if

$$|\mathbf{C}_\alpha(i) - \mathbf{C}_\alpha(j)| < 7\ \text{Å}, \tag{1}$$

$$[\mathbf{C}_\beta(i) - \mathbf{C}_\alpha(i)]\cdot[\mathbf{C}_\beta(j) - \mathbf{C}_\alpha(j)] > 0, \tag{2}$$

$$[\mathbf{H}_\text{N}(i) - \mathbf{N}(i)]\cdot[\mathbf{O}(j) - \mathbf{N}(i)]$$

and
$$[\mathbf{H}_\text{N}(j) - \mathbf{N}(j)]\cdot[\mathbf{O}(i) - \mathbf{N}(j)] \tag{3a}$$

have the same sign, if the two strands are antiparallel, or alternatively

$$[\mathbf{H}_\text{N}(i+1) - \mathbf{N}(i+1)]\cdot[(\mathbf{O}j) - \mathbf{N}(i+1)] \tag{3b}$$

and
$$[\mathbf{H}_\text{N}(j) - \mathbf{N}(j)]\cdot[\mathbf{O}(i-1) - \mathbf{N}(j)]$$

(or the same expressions with i and j exchanged) have the same sign, if the two

TABLE 1

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

AMSOM identifier	Residue pairs		Strand pairs		Protein name
	β_A	β_P	β_A	β_P	
AM 1.2.1.1.1	5	4	1	1	Bovine ferricytochrome *b*5
AM 1.3.1.1.1	3	—	1	—	Tuna ferrocytochrome *c*
AM 1.3.2.1.1	3	—	1	—	Bacterial ferricytochrome *c*2
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 3.1.2.8.1‡	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	—	Porcine 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	Porcine 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	—	Bacteriophage T4 lysozyme
AM 11.1.1.1.1	—	37	—	7	Chicken triose phosphate isomerase (monomer 1)
AM 12.1.1.1.1	3	—	1	—	Carp calcium-binding protein B
AM 12.3.1.1.1	53	17	12	4	Human carbonic anhydrase B
AM 12.3.1.2.1	51	9	11	2	Human carbonic anhydrase C
AM 12.9.1.1.1	40	7	7	1	Human prealbumin
	788	263	170	56	Total

† Co-ordinates supplied by Richard Feldmann (1976).

‡ Updates November 1976 to August 1978 (Feldmann)

strands are parallel. A β-strand is recognized as such and is included in our data base, if the above criteria are fulfilled by at least three consecutive residues. This condition eliminates artifacts where two crossing stretches of the protein chain would be mistakenly identified as β-strands.

Our automatic identification of β-sheets†, a simple type of pattern recognition, is similar in spirit to that of Levitt & Greer (1977); however, our approach requires

† The following items are available on request. (1) List of all sheets and strand pairs with types of hydrogen bonding and geometrical characteristics of each residue pair; (2) list of all occurrences of a particular pair of type XY.

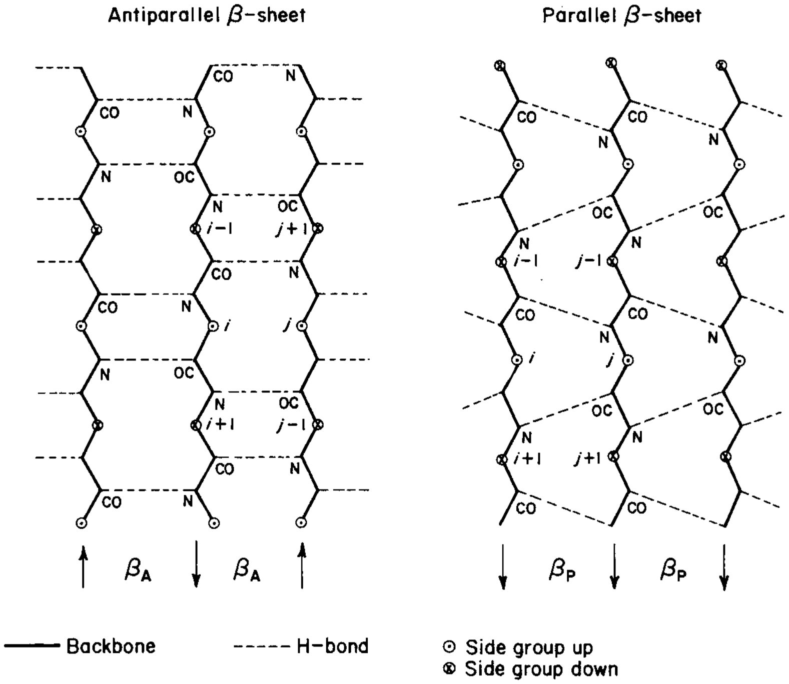

FIG. 2. Schematic drawing of antiparallel β-strands (a) and parallel β-strands (b). The amino acid side groups alternately point to one or the other side of the β-sheet. The correlations investigated in this paper are between nearst neighbours (i, j) on adjacent strands.

knowledge of the co-ordinates of all backbone atoms and of $\mathbf{C}_\beta$, resulting in a more precise identification at the cost of excluding from the data base proteins for which only $\mathbf{C}_\alpha$ co-ordinates are available. For regular stretches of double strands (generally away from the edges of sheets) the two approaches make identical identifications of β residues and agree with those of the crystallographers.

4. Residue–Residue Pair Counts

Once all β-sheets have been identified, we count all residue–residue nearest neighbour contacts of the type (i, j). The counting is done separately for antiparallel (β_A) and parallel (β_P) strands. (What follows in this section is equally applicable to either β_A or β_P.) Note that residues in the interior of a sheet make contacts with residues on two neighbour strands, while residues on the edge of a sheet make only one such contact. The *pair contact count* (or, in short, *pair count*) N_{XY} is thus defined as the number of times a residue of type X is found to make a contact with a residue of type Y. Obviously N_{XY} equals N_{YX}. Counting pair contacts in this way is numerically equivalent to counting pairs, except that for X = Y two pair contacts are counted for one pair: N_{XX} is twice the number of pairs X — X. Single residue counts N_X are defined as the number of times a residue of type X is found in pair contact with another residue. Thus, N_X is twice the number of interior residues (involved in two contacts) plus the number of edge residues (involved in one contact) of type X.

The total number of contacts is twice the number of pairs, and is denoted by N, and the following simple relations hold

$$\sum_{\mathrm{Y}} N_{\mathrm{XY}} = N_{\mathrm{X}}, \quad \sum_{\mathrm{X}} N_{\mathrm{X}} = N. \tag{4}$$

Note that the present definition of single residue frequencies N_{X}/N in β-sheets (Lifson & Sander, 1979*b*) is necessarily different from previous definitions (for a recent review, see Schulz & Schirmer (1979)). It is related to tertiary structure, while the others are related to secondary structure.

5. Expectation Values, Correlations and Confidence Level

If there were no preferred pairing between residues, one would expect the pair frequencies N_{XY}/N to be porportional to the single residue frequencies in β-sheets N_{X}/N and N_{Y}/N. Significant deviations from such a null hypothesis (as it is called in the language of statistical analysis) is used as evidence for recognition. We therefore define E_{XY}, the hypothetical count which would be expected for uncorrelated pairs, as

$$E_{\mathrm{XY}} = N_{\mathrm{X}} N_{\mathrm{Y}}/N \text{ (random pairing)}. \tag{5}$$

The pair correlation, the ratio of actually observed to hypothetical counts, is defined as

$$g_{\mathrm{XY}} = N_{\mathrm{XY}}/E_{\mathrm{XY}}. \tag{6}$$

When the correlation is favourable, $g_{\mathrm{XY}} > 1$, while $g_{\mathrm{XY}} < 1$ indicates an unfavourable correlation. In a nutshell, this definition of pair correlation, g_{XY}, is a way of quantifying the concept of recognition between amino acid residues of type X and those of type Y.

Since we deal with a finite data base, deviations of g_{XY} from 1·0 may be caused by random fluctuations in N_{XY}. We therefore evaluate the level of confidence in the observed correlations by χ^2 analysis, where χ^2 depends on the observed values N_{XY} and the hypothetical values E_{XY} and is defined by

$$\chi^2 \equiv \sum_{\mathrm{X,Y}} (N_{\mathrm{XY}} - E_{\mathrm{XY}})^2/E_{\mathrm{XY}}. \tag{7}$$

Tables giving confidence level as a function of χ^2 and the number of degrees of freedom of the system (see below) are available in most text books on statistics. The χ^2 test can be used to test any hypothesis, simply by substituting for E_{XY} the values expected from the particular hypothesis. To reject the hypothesis of random pairing, we use E_{XY} from equation (5); to reject non-specific recognition, E_{XY} from equation (9).

The single residue frequencies N_{X}/N and N_{Y}/N and the correlations g_{XY} represent properties of the system. Their numerical values should converge as the size of the data base increases. The average (root mean square) deviation of g_{XY} from 1·0:

$$\gamma \equiv \Big(\sum_{\mathrm{X,Y}} (N_{\mathrm{X}}/N)\,(N_{\mathrm{Y}}/N)\,(g_{\mathrm{XY}} - 1)^2\Big)^{\frac{1}{2}} = (\chi^2/N)^{\frac{1}{2}} \tag{8}$$

can therefore be used as a quantitative estimate of the size of the recognition effect.

Note that as γ approaches its limit for large N, χ^2 tends to $\gamma^2 N$ and thus becomes proportional to the size of the data base.

6. Non-specific Recognition: Hydrophobic, Polar and Neutral Residues

It is by now common knowledge that "hydrophobic" and polar interactions dominate the process of protein folding, that polar residues prefer the protein surface, while hydrophobic residues are more abundant in its interior, and this must be true also for β-sheets. Indeed, von Heijne & Blomberg (1977,1978) performed a statistical analysis of polar/non-polar (*non-specific*) interstrand pair correlations in β-sheets, and showed that such correlations are definitely non-random. However, the distribution of residue-pair correlations may be meaningful on the level of individual, or specific, recognition among residues; thus, to establish the level of specificity it is not sufficient to show that the correlations have a high confidence level against the null hypothesis of *random* pairing. In the following we repeat the analysis of von Heijne & Blomberg on our data base, and then use the results as a null hypothesis against which we examine our conjecture of *specific* recognition.

In the classification of von Heijne & Blomberg, the hydrophobic residues are Val (V), Leu (L), Ile (I), Phe (F), Tyr (Y), Trp (W), Met (M) and Cys (C); the neutral residues are Thr (T), Ser (S), Ala (A), Gly (G), Pro (P); and the polar residues are Lys (K), Arg (R), Asp (D), Asn (N), His (H), Glu (E) and Gln (Q). The same classification is adopted here, for ease of comparison.

Table 2 presents $N_{\rm U}$, $N_{\rm UV}$ and $g_{\rm UV} \pm \sigma_{\rm UV}$ (pair correlation ± variance) for $\beta_{\rm A}$, $\beta_{\rm P}$ and all-β; the indices U and V represent the hydrophobic, neutral and polar classes of residues. The variance $\sigma_{\rm UV}$ is estimated as follows: $N_{\rm UV}$ are subject to random fluctuations of the order of $\sqrt{E_{\rm UV}}(\sqrt{(2E_{\rm UU})}$ for U = V). Therefore fluctuations of $g_{\rm UV}$ are $\sigma_{\rm UV} = 1/\sqrt{E_{\rm UV}}(\sigma_{\rm UU} = \sqrt{(2/E_{\rm UU})})$. The χ^2 test in Table 2 using equations (5) and (7) allows rejection of the hypothesis of random pairing with a high level of confidence: >99·9%, 99·3%, >99·9%, respectively, for $\beta_{\rm A}$, $\beta_{\rm P}$, all-β ($\chi^2 = 33{\cdot}1$, 12·3, 47·6 for 3 degrees of freedom). One notes differences in pair correlations between parallel and antiparallel pairs, particularly in hydrophobic pairs. A χ^2 test on the observed values $N_{\rm UV}(\beta_{\rm P})$ using expected values $E'_{\rm UV}(\beta_{\rm P}) = E_{\rm UV}(\beta_{\rm P}) \cdot g_{\rm UV}(\beta_{\rm A})$ gives a confidence level of 50%. Thus, whether these differences are significant or due to random fluctuation is an open question, to be settled only by accumulating a larger data base. Our results agree with those of von Heijne & Blomberg (1978) within the limits $\pm\sigma$.

7. Specific Recognition: Individual Residues and Residue-grouping

The data base for antiparallel $\beta_{\rm A}$ strands consists of 1576 contacts (788 residue pairs). It is too small to deduce reliable values for all 210 individual pair correlations $g_{\rm XY}$. Conventional χ^2 tests of significance for small data bases require that all expected values $E_{\rm XY}$ should be at least ~ 5. According to equation (5) this condition is fulfilled for $N_{\rm X} > \sim\sqrt{(5 \times 1576)} \sim 90$. Only Val, Leu, Ile, Tyr, Thr, Ser, Ala satisfy this criterion in $\beta_{\rm A}$. Previously (Lifson & Sander, 1979*a*,1980) we avoided this difficulty by adopting different, less conventional tests of significance. Here we overcome the

TABLE 2

Non-specific recognition between three major classes of residues. Pair contact counts N_{UV} *(top) and pair correlations* $g_{UV} \pm \sigma_{UV}$ *(bottom)*

		Antiparallel strands				Parallel strands				All-beta strands			
		1	2	3		1	2	3		1	2	3	
1	VLIYFWMC	404 $\underline{1{\cdot}2\pm0{\cdot}1}$	195 0·9±0·1	124 $\underline{0{\cdot}7\pm0{\cdot}1}$		192 1·1±0·1	75 0·9±0·1	32 0·7±0·1		596 $\underline{1{\cdot}2\pm0{\cdot}1}$	270 0·9±0·1	156 $\underline{0{\cdot}7\pm0{\cdot}1}$	
2	TSAPG	195 0·9±0·1	178 $\underline{1{\cdot}2\pm0{\cdot}1}$	118 1·0±0·1		75 0·9±0·1	42 1·1±0·2	27 1·2±0·2		270 0·9±0·1	220 1·1±0·1	145 1·1±0·1	
3	KRDNHQE	124 $\underline{0{\cdot}7\pm0{\cdot}1}$	118 1·0±0·1	120 $\underline{1{\cdot}4\pm0{\cdot}2}$		32 0·7±0·1	27 1·2±0·2	24 $\underline{1{\cdot}8\pm0{\cdot}4}$		156 $\underline{0{\cdot}7\pm0{\cdot}1}$	145 1·1±0·1	144 $\underline{1{\cdot}5\pm0{\cdot}1}$	
N_U		723	491	362	1576	299	144	83	526	1022	635	445	2102
%		45·9	31·2	23·0	100	56·8	27·4	15·8	100	48·6	30·2	21·2	100
Average deviation of g_{UV} from random (γ, see text for definition)					0·205				0·216				0·213
Confidence level for rejecting hypothesis of random pairing (%)					>99·9				99·3				>99·9

Values of g_{UV} which differ from the random value 1·0 by more than their estimated variance are underlined.
The variance σ_{UV} is estimated as $1/\sqrt{E_{XY}}$ (see text Section 6).

difficulty by minimal grouping at the expense of losing individuality for some residues. We group residues which are less frequent into groups which are large enough to fulfil the condition $E_{XY} \gtrsim 5$ and are composed of residues which resemble each other best in size, structure, polarity and genetic exchangeability (Sander & Schulz, 1979). The groups in β_A are (N_X is given in parenthesis): Phe(68)-Trp(47), Met(24)-Cys(45), Ala(119)-Pro(29), Lys(83)-Arg(45), Asp(39)-Asn(41)-His(44) and Glu(47)-Gln(63). Gly(81) was retained as an individual residue because of its special role. Note, for example, that $N_{Gly,Gly} = 0$, $E_{Gly,Gly} = 6{\cdot}7$ for all-β (β_A and β_P combined). Thus the absence of Gly-Gly pairs seems statistically significant, in spite of the relatively low N_X counts in β_A and β_P taken separately. Pro was paired with Ala because of their close genetic exchangeability. Met and Cys were kept as a pair although their joint count N_X is low because of the large count $N_{XY} = 8$. As the data base for parallel (β_P) strands consists of only 526 contacts (263 residue pairs), the grouping has to be more extensive. The groups are: Val(108), Ile(63), Leu(53)-Met(11)-Cys(11), Phe(24)-Tyr(21)-Trp(8), Ala(44)-Gly(38)-Pro(8), Thr(21)-Ser(33), and Lys(22)-Arg(10)-Asp(14)-Asn(12)-His(5)-Gln(5)-Glu(15).

Table 3 presents N_X, N_{XY} and $g_{XY} \pm \sigma_{XY}$ for (A) antiparallel and (B) parallel strands; the standard deviation σ_{XY} is estimated here also by $1/\sqrt{E_{XY}}$ (see discussion of Table 2 in Section 6, above). The average deviation from random, γ, indicates the overall magnitude of the recognition effect; even though part of γ comes from random fluctuations due to the finite size of the data base, one can draw the tentative conclusion that the effect is stronger in β_A ($\gamma = 0{\cdot}423$) than in β_P ($\gamma = 0{\cdot}374$). (If the effect were equal one would have $\gamma(\beta_P) > \gamma(\beta_A)$ since a smaller data base is relatively more sensitive to statistical fluctuations, but in fact the opposite is the case.)

The χ^2 level of confidence for rejecting the null hypothesis of random pairing between β-strands is 99·998% for β_A ($\chi^2 = 109{\cdot}5$ for the half-table with 78 degrees of freedom) and 98·2% for β_P ($\chi^2 = 25{\cdot}3$ for the half-table with 21 degrees of freedom). For this test of specific *versus* random recognition the conclusion is the same as in the test of non-specific recognition: there *is* statistically significant recognition. However, whether there is specific recognition over and above the non-specific one is an open question, requiring further analysis (see next section).

8. Specific *versus* Non-specific Recognition

In order to examine the existence of specific recognition over and above non-specific recognition, we test the deviation of the counts N_{XY} (Table 3A and B) from the values

$$E_{XY} = g_{UV} N_X N_Y / N \quad \text{(non-specific pairing)}, \tag{9}$$

i.e. the values expected if recognition only acts on the level of the three classes of hydrophobic, neutral and polar residues. The χ^2 test gives a level of confidence for rejecting the hypothesis of non-specific recognition in favour of specific recognition: 98·9% for β_A and 76% for β_P. The lower level of significance for β_P is at least in part due to the smaller data base, i.e. smaller N, since χ^2 is proportional to N for given γ (eqn (8)). In addition, with the more extensive grouping for β_P more of the specificity is lost. In particular the polar residues in β_P were grouped the same way in the "specific"

TABLE 3

Specific recognition between individual or minimally grouped residues. Pair contact counts N_{XY} *(top) and pair correlations* $g_{XY} \pm \sigma_{XY}$ *(bottom)*

A. *Antiparallel strands*

	1 V	2 L	3 I	4 Y	5 FW	6 MC	7 T	8 S	9 AP	10 G	11 KR	12 DNH	13 QE	
1 V	26 1·2±0·3	25 1·5±0·2	23 1·7±0·3	11 0·9±0·3	18 1·3±0·3	7 0·9±0·3	6 0·4±0·2	19 1·3±0·3	14 0·8±0·2	12 1·2±0·3	13 0·9±0·3	8 0·5±0·3	6 0·5±0·3	
2 L	25 1·5±0·2	8 0·6±0·4	12 1·2±0·3	12 1·4±0·3	14 1·4±0·3	11 1·8±0·4	12 1·0±0·3	5 0·5±0·3	13 1·0±0·3	10 1·4±0·4	5 0·4±0·3	10 0·9±0·3	4 0·4±0·3	
3 I	23 1·7±0·3	12 1·2±0·3	8 1·0±0·5	5 0·7±0·4	9 1·1±0·3	0 0·0±0·5	9 0·9±0·3	4 0·5±0·3	20 1·9±0·3	2 0·3±0·4	8 0·9±0·3	9 1·0±0·3	3 0·4±0·4	
4 Y	11 0·9±0·3	12 1·4±0·3	5 0·7±0·4	12 2·0±0·6	8 1·1±0·4	6 1·4±0·5	9 1·0±0·3	1 0·1±0·4	7 0·8±0·3	6 1·2±0·4	7 0·9±0·4	6 0·8±0·4	8 1·2±0·4	
5 FW	18 1·3±0·3	14 1·4±0·3	9 1·1±0·3	8 1·1±0·4	8 1·0±0·5	6 1·2±0·4	3 0·3±0·3	4 0·4±0·3	9 0·8±0·3	7 1·2±0·4	12 1·3±0·3	10 1·1±0·3	7 0·9±0·4	
6 MC	7 0·9±0·3	11 1·8±0·4	0 0·0±0·5	6 1·4±0·5	6 1·2±0·4	8 2·6±0·8	5 0·8±0·4	6 1·1±0·4	8 1·2±0·4	4 1·1±0·5	2 0·4±0·4	2 0·4±0·4	4 0·8±0·5	
7 T	6 0·4±0·2	12 1·0±0·3	9 0·9±0·3	9 1·0±0·3	3 0·3±0·3	5 0·8±0·4	18 1·5±0·4	21 1·9±0·3	18 1·4±0·3	4 0·6±0·4	13 1·2±0·3	7 0·6±0·3	14 1·4±0·3	
8 S	19 1·3±0·3	5 0·5±0·3	4 0·5±0·3	1 0·1±0·4	4 0·4±0·3	6 1·1±0·4	21 1·9±0·3	14 1·5±0·5	13 1·1±0·3	5 0·8±0·4	11 1·1±0·3	13 1·3±0·3	7 0·8±0·3	
9 AP	14 0·8±0·2	13 1·0±0·3	20 1·9±0·3	7 0·8±0·3	9 0·8±0·3	8 1·2±0·4	18 1·4±0·3	13 1·1±0·3	10 0·7±0·4	7 0·9±0·4	11 0·9±0·3	7 0·6±0·3	11 1·1±0·3	
10 G	12 1·2±0·3	10 1·4±0·4	2 0·3±0·4	6 1·2±0·4	7 1·2±0·4	4 1·1±0·5	4 0·6±0·4	5 0·8±0·4	7 0·9±0·4	0 0·0±0·7	7 1·1±0·4	9 1·4±0·4	8 1·4±0·4	
11 KR	13 0·9±0·3	5 0·4±0·3	8 0·9±0·3	7 0·9±0·4	12 1·3±0·3	2 0·4±0·4	13 1·2±0·3	11 1·1±0·3	11 0·9±0·3	7 1·1±0·4	12 1·2±0·4	11 1·1±0·3	16 1·8±0·3	
12 DNH	8 0·5±0·3	10 0·9±0·3	9 1·0±0·3	6 0·8±0·4	10 1·1±0·3	2 0·4±0·4	7 0·6±0·3	13 1·3±0·3	7 0·6±0·3	9 1·4±0·4	11 1·1±0·3	20 2·0±0·5	12 1·4±0·3	
13 QE	6 0·5±0·3	4 0·4±0·3	3 0·4±0·4	8 1·2±0·4	7 0·9±0·4	4 0·8±0·5	14 1·4±0·3	7 0·8±0·3	11 1·1±0·3	8 1·4±0·4	16 1·8±0·3	12 1·4±0·3	10 1·3±0·5	
N_X	188	141	112	98	115	69	139	123	148	81	128	124	110	Total 1576
%	11·9	8·9	7·1	6·2	7·3	4·4	8·8	7·8	9·4	5·1	8·1	7·9	7·0	100

Average deviation of g_{XY} from random (γ, see text for definition) 0·423

Confidence level for rejecting hypothesis of non-specific pairing (%) 98·9.

Values of g_{XY} which differ from the random value 1·0 by more than their estimated variance are underlined.

[illegible] (see text Section 6)

Values of g_{XY} which differ from the random value 1·0 by more than their estimated variance are underlined.
The variance σ_{XY} is estimated as $1/\sqrt{E_{XY}}$ (see text Section 6).

TABLE 3 *(continued)*

B. *Parallel strands*

		1 V	2 I	3 LMC	4 YFW	5 AGP	6 TS	7 KRDNHQE	
1	V	34 $\underline{1{\cdot}5\pm0{\cdot}3}$	18 $\underline{1{\cdot}4\pm0{\cdot}3}$	13 0·8±0·3	9 0·8±0·3	16 0·9±0·2	8 0·7±0·3	10 $\underline{0{\cdot}6\pm0{\cdot}2}$	
2	I	18 $\underline{1{\cdot}4\pm0{\cdot}3}$	10 1·3±0·5	15 $\underline{1{\cdot}7\pm0{\cdot}3}$	4 0·6±0·4	10 0·9±0·3	2 $\underline{0{\cdot}3\pm0{\cdot}4}$	4 $\underline{0{\cdot}4\pm0{\cdot}3}$	
3	LMC	13 0·8±0·3	15 $\underline{1{\cdot}7\pm0{\cdot}3}$	10 0·9±0·4	7 0·9±0·4	15 1·2±0·3	8 1·0±0·4	7 $\underline{0{\cdot}6\pm0{\cdot}3}$	
4	YFW	9 0·8±0·3	4 0·6±0·4	7 0·9±0·4	6 1·1±0·6	12 1·3±0·3	4 0·7±0·4	11 1·3±0·3	
5	AGP	16 0·9±0·2	10 0·9±0·3	15 1·2±0·3	12 1·3±0·3	16 1·0±0·4	8 0·9±0·3	13 0·9±0·3	
6	TS	8 0·7±0·3	2 $\underline{0{\cdot}3\pm0{\cdot}4}$	8 1·0±0·4	4 0·7±0·4	8 0·9±0·3	10 $\underline{1{\cdot}8\pm0{\cdot}6}$	14 $\underline{1{\cdot}6\pm0{\cdot}3}$	
7	KRDNHQE	10 $\underline{0{\cdot}6\pm0{\cdot}2}$	4 $\underline{0{\cdot}4\pm0{\cdot}3}$	7 $\underline{0{\cdot}6\pm0{\cdot}3}$	11 1·3±0·3	13 0·9±0·3	14 $\underline{1{\cdot}6\pm0{\cdot}3}$	24 $\underline{1{\cdot}8\pm0{\cdot}4}$	
N_X		108	63	75	53	90	54	83	Total 526
%		20·5	12·0	14·3	10·1	17·1	10·3	15·8	100

Average deviation of g_{XY} from random (γ, see text for definition) 0·374

Confidence level for rejecting hypothesis of non-specific pairing (%) 76.

Values of g_{XY} which differ from the random value 1·0 by more than their estimated variance are underlined.
The variance of σ_{XY} is estimated as $1/\sqrt{E_{XY}}$ (see text Section 6).

and "non-specific" hypothesis. For "all-β" the χ^2 test yields a confidence level of 99·99% for rejecting the hypothesis of non-specificity.

In other words, the recognition clearly goes beyond the level of the three classes: polar, neutral and non-polar. It is reasonable to expect that pair correlations among *all twenty individual* residues in both parallel and antiparallel β-sheets will turn out to reflect specific recognition as more data on β-sheets accumulate.

9. Specific Recognition: Correlations

Once specific recognition is seen to be statistically significant, it becomes interesting to discuss the most significant correlations between particular pairs. In Table 3A and B, g_{XY} values which differ from 1·0 (either above or below) beyond their estimated various σ_{XY} are underlined. For β_A there are 15 favourable ($>$1·0) and 18 unfavourable ($<$1·0) such correlations; for β_P there are six and four, respectively. In β_A the charged residues show particularly high correlation, especially the group Asp-Asn-His with itself (2·0±0·5), and Lys-Arg with Glu-Gln (1·8±0·3); these are probably due to specific charge–charge and dipole–dipole interactions. Also high in β_A: 1·9±0·3 for Ser/Thr and for Ile/Ala-Pro, 2·0±0·6 for Tyr with itself and 1·7±0·3 for Ile/Val. The most interesting unfavourable correlations between antiparallel pairs are 0·4±0·2 for Thr/Val, 0·3±0·3 for Thr/Phe-Trp, 0·4±0·3 for Lys-Arg/Leu and Glu-Gln/Leu,

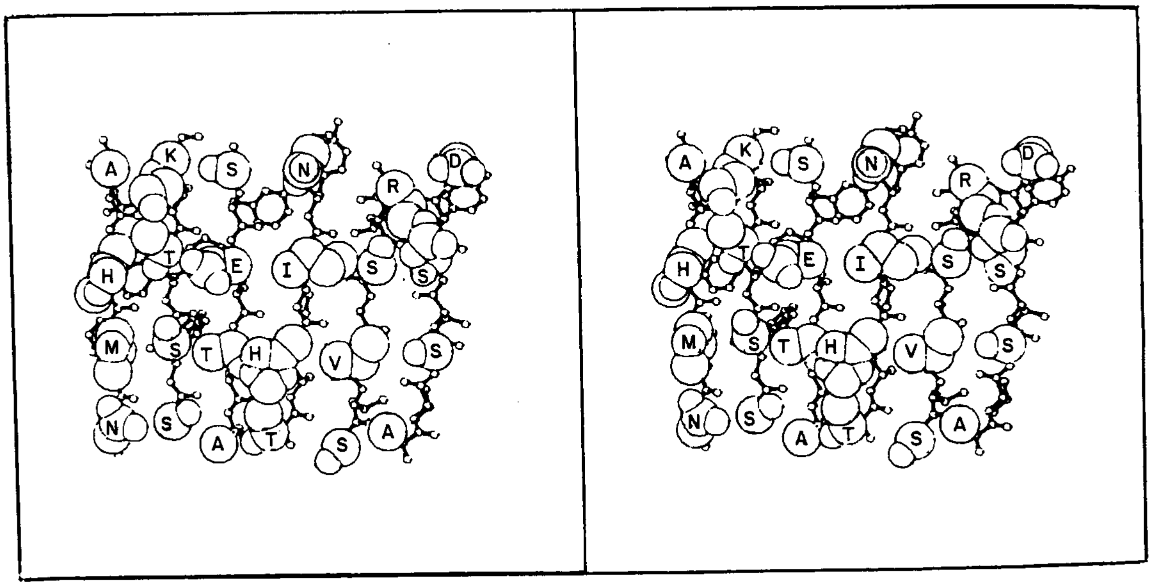

FIG. 3. This view of a section of the large antiparallel β-sheet of concanavalin A illustrates the variety of side-group–side-group interactions between neighbouring strands typical of globular proteins. Side-group atoms on one face of the sheet are shown at 70% of their atomic contact radius, while atoms of the backbone and opposite-face side-groups are reduced to 0·3 Å radius. The side-groups are labelled in the standard one-letter code (see text). The β-strands run vertically and the side-groups on neighbouring strands form horizontal rows of interacting nearest neighbours. The strands are (from top to bottom and left to right): Ala125-Asn131, Lys114-Ser108, Ser190-Ala196, Asn55-Thr49, Arg60-Ser66, Asp78-Ala72. The co-ordinates are taken from data set AM 9.1.1.2.1 (see Table 1). Drawn using PLOTRES (by C. Sander) and PLUTO (by S. Motherwell, Cambridge). A stereoview of the other side of this sheet as well as of other β-sheets is included by Lifson & Sander (1980).

0·3±0·4 for Gly/Ile and 0·4±0·4, for Glu-Gln/Ile. Putative "strand pair breakers" with $g_{XY} = 0·0$ (Met-Cys/Ile and Gly/Gly) are difficult to interpret because of relatively low expected counts, but the extreme backbone flexibility of Gly is likely to exclude Gly-Gly pairs across strands. In β_P, Ile correlates particularly well with Leu-Met-Cys (1·7±0·3) and avoids pairing with Thr-Ser (0·3±0·4) and the other polar residues (0·4±0·3). Val, however, shows no preference for Leu-Met-Cys.

The physical reasons for the particularly favourable or unfavourable pairs are impossible to determine by statistical analysis. In general, the correlations may be due to direct interaction between the side groups of a nearest neighbour pair such as packing and dipole–dipole interactions; alternatively, they may be due to indirect interaction mediated by solvent molecules, such as hydrogen bonding between side groups and water, or by the packing environment in the interior of the protein. In addition, the pair interactions may be modified by co-operative interaction of residues in the plane of the sheet, such as close packing of rows. The statistical results of this paper can be used as a pointer to particular interesting correlations, which are worth studying by other means.

The specific residue–residue correlations derived here, when combined with other known conformational preferences, may be useful in the prediction of the three-dimensional structure of proteins.

We would like to thank Michael Levitt, John Moult, Georg Schulz, Peter Stern, Joel Sussman and Ada Yonath for interesting discussions and/or comments on the manuscript.

REFERENCES

Feldmann, R. J. (1976). *AMSOM Atlas of Molecular Structures on Microfiche*, U.S. National Institutes of Health, Maryland.

von Heijne, G. & Blomberg, C. (1977). *J. Mol. Biol.* **117**, 821–824.

von Heijne, G. & Blomberg, C. (1978). *Biopolymers*, **17**, 2033–2037.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Levitt, M. & Greer, J. (1977). *J. Mol. Biol.* **114**, 181–293.

Lifson, S. & Sander, C. (1979*a*). In *Molecular Mechanisms of Biological Recognition* (Balaban, M., ed.), pp. 145–156, Elsevier, Amsterdam.

Lifson, S. & Sander, C. (1979*b*). *Nature (London)*, **282**, 109–111.

Lifson, S. & Sander, C. (1980). In *Protein Folding* (Jaenicke, R., ed.), pp. 289–316, Elsevier, Amsterdam.

Richardson, J. S. (1976). *Proc. Nat. Acad. Sci., U.S.A.* **73**, 2619–2623.

Richardson, J. S. (1977). *Nature (London)*, **268**, 495–500.

Sander, C. & Schulz, G. E. (1979). *J. Mol. Evol.* **13**, 245–252.

Schulz, G. E. & Schirmer, R. H. (1974). *Nature (London)*, **250**, 142–144.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, Springer-Verlag, New York.

Sternberg, M. J. E. & Thornton, J. M. (1977*a*). *J. Mol. Biol.* **110**, 269–283.

Sternberg, M. J. E. & Thornton, J. M. (1977*b*). *J. Mol. Biol.* **110**, 285–296.

Sternberg, M. J. E. & Thornton, J. M. (1977*c*). *J. Mol. Biol.* **115**, 1–17.