DD704731

COPY

CI-05880034-3

CISTI ICIST

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

Telephone: 212/639-7441
Fax: 646/422-2316

ORDER NUMBER:	CI-05880034-3
Account Number:	DD704731
Delivery Mode:	F31
Delivery Address:	arielsf.infotrieve.com/140.163.217.217
Reply Via:	SMT
Reply Address:	conyersd@mskcc.org
Submitted:	2005/11/25 11:12:19
Received:	2005/11/25 11:12:19
Printed:	2005/11/25 13:42:00

Direct	**Periodical**	**OPENURLOPAC**	**UNITED STATES**

IRBM Ser	QP551 P965	v. 1-11; 1986-1998. c. 1.
Internet	QP551 P965	v. 10-16; 1997-2003. c. 1.
MAIN Ser	QP551 P965	v. 1-16; 1986-2003. c. 1.

Protein engineering

10409592

Title:	PROTEIN ENGINEERING
DB Ref. No.:	IRN10409592
ISSN:	ISSN02692139
Vol./Issue:	2
Date:	1989
Pages:	329-334
Article Title:	POLARITY AS A CRITERION IN PROTEIN DESIGN
Article Author:	BAUMANN G, FROMMEL C
Report Number:	IRN10409592
Publisher:	IRL PRESS,
Client Number:	DDS36739/GANGI-DINO, RITA

INSTRUCTIONS: PATRON REQUESTS COLOUR IF AVAILABLE. THANK YOU.

Estimated cost for this 6 page document: $10.2 document supply fee + $25.5 copyright = $35.7

Transmission problem
Please redo
Sharon

Protein Engineering vol.2 no.5 pp.329–334, 1989

Polarity as a criterion in protein design

G.Baumann, C.Frömmel[1,2] and C.Sander[1]

Central Institute of Molecular Biology, Berlin-Buch, GDR, [1]BIOcomputing Programme, EMBL, Heidelberg, FRG

[2]Permanent address: Institute of Biochemistry, Humboldt University, Berlin, GDR

Hypothetical proteins can be tested computationally by determining whether or not the designed sequence-structure pair has the characteristics of a typical globular protein. We have developed such a test by deriving quantities with approximately constant value for all globular proteins, based on empirical analysis of the exposed and buried surfaces of 128 structurally known proteins. The characteristic quantities that best appear to segregate badly designed or deliberately misfolded proteins from their properly folded natural relatives are the polar fraction of side chains on the protein surface and, independently, in the protein interior. Three of the seven hypothetical structures tested here can be rejected as having too many polar side-chain groups in the interior or too few on the protein surface. In addition, a recently designed nutritional protein is identified as being very much unlike globular proteins. These database-derived characteristic quantities are useful in screening designed proteins prior to experiment and may be useful in screening experimentally determined (X-ray, NMR) protein structures for possible errors.

Key words: Electrostatics/globular proteins/hydrophobicity/protein data bank/protein engineering

Introduction

With the tremendous progress in gene synthesis and protein expression, design of completely new proteins has become an experimental reality. Several *de novo* designs are now in the production or testing stage. However, as the protein folding problem is as yet essentially unsolved, theoretical designs have a less than optimal chance of leading to successful experimentation. In order to avoid unnecessary experiments, more stringent design evaluation and improved design methods are needed.

An interesting example of design evaluation is provided by deliberately misfolded proteins in which the sequence of a protein known to have an all-helical 3-D structure is placed into a known structure of a completely different type, an antiparallel β-barrel, and *vice versa*. For the evaluation of the quality of these clearly incorrect hypothetical structures, intramolecular energy, calculated in vacuum using standard potentials, was not a sensitive criterion (Novotny *et al.*, 1984). However, the misfolded structures (not the improved versions used in this paper) were reported to have surface areas 15–20% larger than the native structures and to have a greater proportion of non-polar side-chain atoms exposed to solvent, using Chothia's (1976) definition of non-polar and polar surfaces. The work of these authors indicates that characterization of the distribution of polar atoms in all known globular protein structures may lead to generally useful design criteria.

The key aspect is the development of criteria with sufficient discriminatory power. For example, the free energy difference between the folded and unfolded state would be an optimal criterion, but present theories are not capable of calculating free energy differences to sufficient accuracy. However, faced with the lack of an accurate theory of protein folding, empirical observations of regularities gleaned from the database of solved 3-D protein structures can be very useful, assuming that the current database is sufficiently representative of the range of possible structures. It is on such empirical analysis of the database that we base our criteria.

Derivation of characteristic quantities by empirical analysis

The aim of our analysis is to derive quantities that are characteristic of globular proteins in that their distribution in the database can be used to discriminate against non-globular proteins, i.e. against sequences that will not fold up into any globular shape or not fold into a *given* globular shape. Previous analyses (e.g. Klein *et al.*, 1984) have provided us with a long list of properties useful for classifying or describing sequences (amino acid composition, number of charged residues, net charge, charge segregation along the chain, number of hydrophobic residues, hydrophobic periodicity, number of hydrophobic segments, strength of helix/sheet preferences, etc.) or 3-D structures (compactness of shape, total solvent accessible surface area, hydrophobic contacts, buried charges, etc.). Our task is to select and invent the quantities that are most discriminating.

In order to arrive at generally useful criteria, the quantities used must be properly normalized. To illustrate by a trivial example, natural proteins have a low number of Trp residues; the appropriate quantity is the relative number of Trp residues, i.e. the absolute number divided by the total number of residues. In general, normalization should be such that the *distribution* of values in the database is *narrow* compared to the difference between a protein in the database and e.g. a misfolded protein.

Below, we first define and analyse quantities that depend only on the amino acid sequence (called sequence-dependent characteristics) and then quantities that depend on the 3-D structure (called structure-dependent characteristics). We discuss the distribution of these quantities in the database of proteins of known 3-D structure (hereafter called database), derive reasonable lower and upper limits for each quantity and apply the resulting criteria to hypothetical sequences and structures.

Sequence-dependent characteristics

Molecular weight per residue

The simplest sequence-dependent properties reflect the amino acid composition. MWRES, the molecular weight per residue, quantifies the average size per residue

$$\mathrm{MWRES} = \frac{\mathrm{MW}}{\mathrm{NRES}}$$

where MW is the total molecular weight of the protein in daltons and NRES is the number of residues in the protein. As there is a linear relationship between molecular weight and maximal solvent accessible surface of an extended chain, the latter quantity could be used equivalently. The range of MWRES values in the database (Table II, in units of daltons/residue) is sufficiently narrow to identify, e.g. polyalanine (MWRES=71), polytyrosine (MWRES=163) and collagen (MWRES=90.2) as non-globular, but not keratin (MWRES=107.1).

In the database of known structures, the porcine hormone glucagon, a 29-residue helical peptide (1GCN, MWRES=120) and the bovine eye lens protein γ-crystallin (1GCR, MWRES=120) have the highest values of MWRES. The unusual amino acid composition of γ-crystallin, low in Ala, Thr and Lys residues (1.1, 2.2, 1.1%) and high in Arg, Phe, Tyr, Trp (11.4, 5.1, 8.6, 2.2%), may reflect unusual functional requirements as an eye lens protein. Excluding 1GCN, 1PPT and 1GCR, we define as normal the range 97 < MWRES < 118 daltons/residue.

Maximum polar fraction

Another simple quantity that depends only on the amino acid composition is POLFRAC__MAX, the polar fraction (defined in physical terms below) in an extended chain conformation, calculated as a simple function of single residue values POLSURF(k) and SURF(k), taken from Table I. Here, NRES is the total number of residues in the protein and POLFRAC__MAX is in units of e, the electronic charge.

$$\mathrm{POLFRAC\_\_MAX} = \sum_{k=1}^{\mathrm{NRES}} \mathrm{POLSURF}(k) / \sum_{k=1}^{\mathrm{NRES}} \mathrm{SURF}(k)$$

Polyaspartic acid (POLFRAC__MAX = 0.256) is clearly much more polar, keratin (POLFRAC__MAX = 0.192) somewhat more polar and polyvaline (POLFRAC__MAX = 0.112) much more hydrophobic than sequences of typical globular proteins (Table II).

Wheat germ agglutinin (isolectin 2, 3WGA) has an unusually polar sequence (POLFRAC__MAX = 0.194). Both the amino acid composition (25% Gly, 18% Cys—somewhat like metallothioneins) and the tertiary structure (dimer, 16 S-S bridges per monomer, little hydrogen-bonded secondary structure) are atypical of globular proteins. Melittin, 1MLT, a small 26-residue protein capable of membrane insertion is significantly less polar (0.155 e) than water-soluble globular proteins. Excluding 1MLT and 3WGA, we define as normal the range 0.160 < POLFRAC__MAX < 0.185 e.

A recently designed nutritional protein (Biernat *et al.*, 1987) clearly falls outside the class of globular proteins on sequence-based criteria alone, with a very high average residue size (MWRES = 174.0) and very low polarity (POLFRAC__MAX = 0.138). Note the absence of small residues (AGSC) in its sequence: MVWYL VIKVI RLIRL THKHT LITLR (repeated). The low polarity of the sequence may indicate that the protein would not be water soluble in physiological conditions and would not fold into a globular shape.

Structure dependent characteristics

Accessible surface area per molecular weight

Evaluation of designs becomes more sensitive when both the designed sequence and the corresponding hypothetical structure are known, as is the case with the five proteins designed *de novo* during an EMBO course in 1986 (Sander, 1987). For example, the relative compactness of globular proteins can be quantified by calculating the solvent accessible surface area, as conceptually defined by Lee and Richards (1971): for a given chain length an extended conformation has maximal accessible surface, while a perfectly spherical protein would have minimal accessible surface. The actual surface values SURF(MW) for globular proteins are somewhat larger than those for a perfect sphere and are proportional to molecular weight (MW), with an offset of SURF(0)=1178 Å^2 at zero molecular weight (Figure 1). The appropriate normalized characteristic quantity removes this dependence

$$\mathrm{SURFMW} = (\ \mathrm{SURF(MW)} - \mathrm{SURF(0)}\ )\ /\ \mathrm{MW}$$

so that SURFMW is approximately constant over the entire range of molecular weights with an average value of 0.403 Å^2/dalton. Non-globular shapes like a long α-helix clearly have an excessively large surface: e.g. SURFMW = 0.661 for residues 75–126 in data set 1HMG, haemagglutinin, taken in isolation. Application of the corresponding criterion (does the value for a protein fall inside or outside the range of values in the database?) leads to some interesting results.

Unusually large values (Table III) of external surface per molecular weight in the database are from an immunoglobulin (1PFC, fragment, only 3.1 Å resolution), α-bungarotoxin (2ABX, little secondary structure, 10 S-S bridges), wheat germ agglutinin (3WGA, see above), ovomucoid inhibitor (1OVO, fragment) and melittin (1MLT, see above). The large SURFMW values for the multi-haem cytochrome C3 protein (2CDV),

Table I. Reference surface values for amino acids: maximal accessible surface area, SURF, and polar surface area, POLSURF, in an extended structure

Residue	Main-chain		Side-chain		Residue		Residue
(X)	SURF	POLSURF	SURF	POLSURF	SURF	POLSURF	SURF (#)
Ala	43.6	13.9	62.7	2.6	106.3	16.5	115
Arg	40.1	13.5	199.3	45.5	239.5	59.0	225
Asn	40.5	13.8	109.2	28.9	149.8	42.7	160
Asp	41.1	13.8	108.4	30.4	149.5	44.3	150
Cys	41.1	13.8	97.4	1.9	138.5	15.7	135
Glu	42.0	13.8	140.8	37.5	182.8	51.3	190
Gln	43.4	14.1	143.1	31.0	186.6	45.2	180
Gly	83.6	19.6	0.0	0.0	83.6	19.6	75
His	41.1	13.8	141.1	13.8	182.2	27.6	195
Ile	39.7	13.4	131.8	3.8	171.5	17.2	175
Leu	40.2	12.9	123.4	3.7	163.6	16.7	170
Lys	40.4	13.5	160.5	15.0	200.8	28.5	200
Met	42.1	13.9	151.5	7.6	193.7	21.5	185
Phe	40.5	13.5	160.4	2.2	200.9	15.7	210
Pro	39.6	11.4	96.3	1.8	135.9	13.2	145
Ser	42.6	14.0	80.7	11.3	123.1	25.2	115
Thr	41.3	13.8	100.4	9.7	141.7	23.5	140
Trp	40.5	13.5	204.6	10.7	245.4	24.7	255
Tyr	40.5	13.5	171.9	14.7	212.4	28.2	230
Val	39.5	13.5	108.9	3.2	148.4	16.7	155
CSS	40.7	13.5	42.0	0.9	82.7	14.4	–

These values are needed to calculate maximal and internal surfaces and maximal and internal polar fractions of a protein structure. SURF is in units of Å^2 and POLSURF in eÅ^2, as defined in the text. Extended structure for a residue X is defined as a β-strand-like structure Gly-Gly-X-Gly-Gly. See Figure 1 for method of calculation. CSS = CYS in S-S bridges (1/2 cystine). (#) values from Lee and Richards (1971).

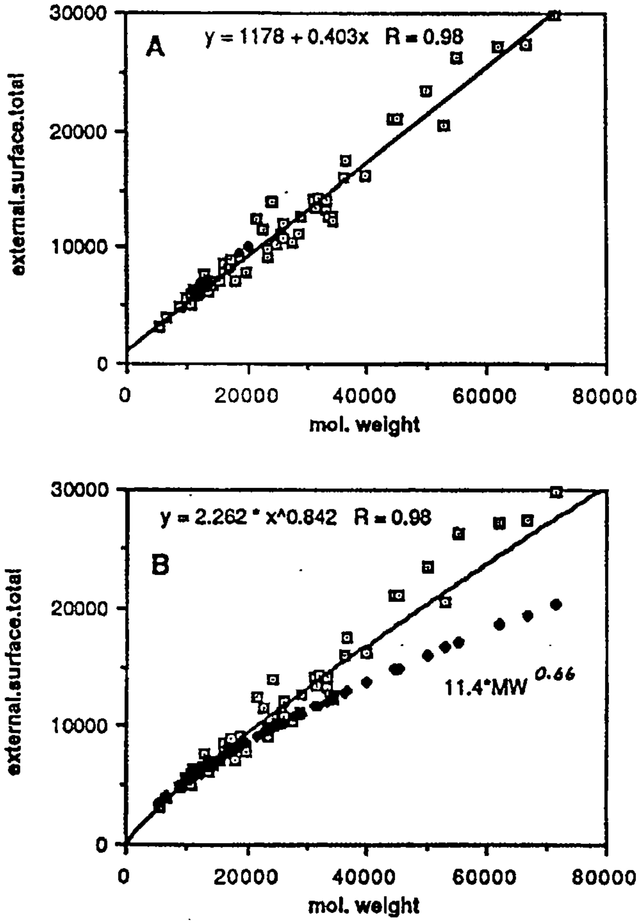

Fig. 1. Analysis of the empirical relationship between molecular weight and external surface area of globular proteins. Each dotted square (⊡) represents a protein of known 3-D structure. The linear fit with additive offset and exponent 1.0 (A) and the exponential fit with zero offset and exponent 0.84 (B) are equally good with a coefficient of regression r = 0.98. For simplicity, we work with the linear fit. The purely linear model is useful in defining characteristic quantities that are independent of molecular weight. Black diamonds (◆) in (A) show that designed (Sander, 1987) and misfolded (Novotny *et al.*, 1984) proteins do not have abnormal values of total exposed surface..

Exponential versus linear fit: Surface and weight of geometrical spheres are related by an exponent of 2/3 [black diamonds (◆) in (B)]. For globular proteins this simple geometrical fit breaks down at larger molecular weights. There are several ways to improve the fit. (i) Use the exponent of MW as a parameter but still force the fit to include the origin at MW=0, SURF=0; this yields SURF ≈ MW $^{0.84}$ (B) with an exponent significantly larger than the purely geometrical value of 2/3 (Janin, 1976; Teller, 1976) and similar to the exponent found by Miller *et al.* (1987a,b), 0.73 and 0.76. (ii) In an exponential fit, ignore the origin, as it is not a data point and allow a non-zero surface at MW=0; this yields SURF = 1200 + 0.537*MW$^{0.97}$, with an exponent close to 1.0, (iii) Linear fit: SURF = 1178 + 0.403*MW. The linear fit is the simplest and has the same coefficient of linear regression as the exponential fits (i) and (ii). The non-zero offset reflects the fact that a small (non-cyclic) peptide cannot fold back onto itself in a compact shape until a minimum length of approximately 10–15 residues is reached. No matter which fit is used, there is at larger molecular weights an apparent excess of protein surface over the geometrical minimum of SURF ≈ MW$^{0.67}$. The excess may be interpretable in terms of the roughness of protein surfaces, the restricted topology of supersecondary structure, domain organization and the like (Gates, 1979).

Method of surface calculation: the external surface of an atom, also called the solvent accessible surface, is calculated by numerical integration of small surface patches over 320 isotropically distributed points (Shrake and Rupley, 1973; Kabsch and Sander, 1983); the points are on the surface of a sphere centred at the atom with radius $r_{atom} + r_{water}$, where r_{atom} = 1.55 for O, 1.65 for N, 1.80 for C, 1.90 for S and r_{water} = 1.4 Å. A point contributes to the external surface if a water molecule centered at it does not touch any other protein atom. Hydrogen atoms are included in the heavier atom they are attached to (united atom approximation), except for the hydrogen of the peptide NH group. The reference value, the external surface of an unfolded protein, is calculated simply as the sum over the surfaces of individual amino acids X embedded in the pentapeptide Gly-Gly-X-Gly-Gly in an extended conformation (Table I). Surfaces are reported in Å^2.

Data base: 64 proteins of known structure:

155C, 156B, 1ABP, 1APR, 1AZU, 1BP2, 1CAC, 1CPV,
1ECD, 1EST, 1FAB, 1FDX, 1GPD, 1HIP, 1INS, 1LH1,
1LHB, 1LZM, 1MBN, 1MBS, 1OVO, 1P2P, 1PCY, 1PTN,
1PYP, 1REI, 1RNS, 1SBT, 1TIM, 2ACT, 2ADK, 2ALP,
2APE, 2APP, 2B5C, 2CAB, 2GCH, 2GRS, 2PAB, 2PGK,
2RHE, 2SBV, 2SGA, 2SNS, 2SOD, 2SSI, 351C, 3BP2,
3C2C, 3CNA, 3CYT, 3FXC, 3FXN, 3PGK, 3PGM, 3TLN,
4ADH, 4ATC, 4DFR, 4LDH, 4PTI, 5CPA, 7LYZ, 8PAP.

with four-letter identifiers from the Protein Data Bank (Bernstein *et al.*, 1977). If the data set contains an oligomer, external and internal surface areas are calculated for the oligomer as a whole.

0.589, is due to the large binding crevice and is reduced to a normal value, 0.418, when the haems are taken into account as excluding solvent (Table III). The same is true for a structurally homologous multihaem protein, 1CY3 (data not shown).

As Miller *et al.* (1987a) have pointed out, inaccuracies in atomic positions tend to increase the external surface, e.g. by 5% for two lysozyme data sets at 2.4 and 1.7 Å resolution. More dramatically, the protein structure with the largest abnormal value of external surface (SURFMW=0.616) in the database and thus most obviously identified as questionable or unusual is a ferredoxin structure reported in 1981 (data set 2FD1) and recently shown to be incorrect by an independent crystallographic experiment (Stout *et al.*, 1988). Excluding 2FD1, 1OVO, 1MLT, 1PFC, 2ABX, 2CDV, 3WGA, 1CY3 and 2FD1, we define as normal the range 0.320<SURFMW≤0.520 Å^2/dalton. Contrary to the earlier versions (Novotny *et al.*, 1984), the refined data sets of misfolded proteins have acceptable values of external surface, showing that the CONGEN algorithm (Bruccoleri and Karplus, 1987) has 'improved' the external surface of these data sets.

Polar fraction

The most sensitive criteria we have been able to find so far are motivated by the classical observation that the interior of globular proteins tends to be hydrophobic. From an atomic point of view, it is important to distinguish between side chains, which vary in polarity, and backbone, which has repeating polar NH and CO groups, and to distinguish between different parts of large side chains, e.g. the non-polar stem of Lys and its polar head.

Although many quantitative measures of hydrophobicity exist for entire residues, there are few on the atomic level. The measure we use, polar surface or POLSURF, reflects the idea that the contribution of individual atoms to the polar nature of a surface is related to the partial charge of an atom; for amino acid side-chains POLSURF was shown by one of us to correlate well with the solvation energy of amino acid side chains (Frömmel, 1984). POLSURF is qualitatively similar to Chothia's (1976) definition of polar surface, but different in detail.

POLSURF is defined as the weighted sum over *absolute* values of partial atomic charges, with the magnitude of atomic surface area as weight. The polar fraction, POLFRAC, is the corresponding normalized quantity.

Table II. Characteristic quantities of globular proteins. Normal range of values in 128 reference proteins of known 3-D structure

A. Quantities independent of 3-D structure: MWRES, molecular weight per residue; POLFRAC_MAX, maximum polar fraction

	Low	Mean	High	Units
MWRES	97.000	109.000	118.000	dalton/res
POLFRAC_MAX	0.160	0.173	0.185	e

B. Quantities dependent on 3-D structure: SURFMW external, external surface per molecular weight; POLFRAC internal, internal polar fraction; POLFRAC external, external polar fraction; POLFRAC SC_int, internal polar fraction for side chains; POLFRAC SC_ext, external polar fraction for side chains

	Low	Mean	High	Units
SURFMW external	0.320	0.400	0.520	Å^2/dalton
POLFRAC internal	0.158	0.172	0.194	e
POLFRAC external	0.144	0.172	0.204	e
POLFRAC SC_int	0.074	0.098	0.122	e
POLFRAC SC_ext	0.115	0.143	0.183	e

The cut-off limits 'low' and 'high' were chosen intuitively by inspection of structures with very large or very small values and identifying them as outliers if their structures appeared atypical of water-soluble, intact globular proteins. The atypical structures are listed in Table III. The 128 known 3-D protein structures available to us from the Protein Data Bank (Bernstein *et al.*, 1977) were:

451C 155C 156B 1ABP 2ABX 2ACT nADH 2ALP 4APE nAPE 2APP nAPR
nATC 2AZA 1AZU nBP2 3C2C 2CAB 1CAC nCAT 1CC5 1CCR 2CCY 2CDV
2CGA nCHA 1CHG 1CN1 2CNA nCPA 2CPP nCPV 1CRN 1CTF nCTS 1CTX
1CY3 1CYC 2CYP nCYT 4DFR 2DHB 2EBX 1ECA 1ECD 1ECN 1ECO nEST
1FB4 1FBJ nFC2 2FD1 1FDH 1FDX 1FX1 nFXN 2GCH 1GCN 1GCR 2GN5
1GP1 nGPD 2GRS 1HBS nHCO 1HDS 2HFL nHHB 1HHO 1HIP 1HKG 1HMG
1HMQ 1IG2 nINS 2KAI nLDH nLDX 2LH1 2LYM nLYZ 1LZ1 2LXM 1LXT
nMBN 1MBO 1MBS 1MCP 1MEV 1MLT 1NTP 1NXB nOVO 1P2P 2PAB nPAD
9PAP nPCY 1PFC 1PP2 1PPD 1PPT 3PTB 2PTC nPT1 nPTN 1PYP 1REI
1RHD 2RHE 2RHV 1RN3 1RNS 1RNT 3RP2 nRSA 1RSM nRXN 1SBT 4SBV
2SGA 3SGB 1SGC 1SN3 2SNS 2TAA 1TGB 2TGP 1TGS 1TGT 1TIM nTLN
1TON 1TPA 3TPI 1TPO 1UBQ 3WGA 2YHX

where the use of several datasets per protein XXX is indicated by *n*XXX.

Table III. Characteristic quantities of globular proteins. Unusually high or low values in crystal and model 3-D structures

	MWRES	POLFRAC	SURFMW		POLFRAC		
		MAX	ext	int	ext	SC_int	SC_ext
Normal range deduced from database of globular proteins							
Low	97.0	0.160	0.320	0.158	0.144	0.074	0.115
Mean	109.0	0.173	0.400	0.172	0.172	0.098	0.143
High	118.0	0.185	0.520	0.194	0.204	0.122	0.183
Crystal structures with unusually high (+) or low (−) values							
1GCR	120.4+	0.180	0.356	0.174	0.195	0.109	0.178
1GCN	120.0+	0.180	0.647+	0.218+	0.164	0.079	0.144
3WGA	97.5	0.194+	0.524+	0.215+	0.167	0.125+	0.131
1MLT	109.2	0.155−	0.566+	0.185	0.128−	0.077	0.104−
2FD1	113.7	0.173	0.616+	0.183	0.164	0.118	0.122
1OVO	107.9	0.175	0.529+	0.179	0.171	0.103	0.128
1PFC	112.2	0.168	0.558+	0.168	0.169	0.088	0.137
2ABX	107.8	0.169	0.539+	0.182	0.153	0.103	0.110−
2CDV	107.7	0.167	0.589+	0.189	0.142−	0.107	0.104−
2CDVh	107.7	0.167	0.418	0.173	0.156	0.097	0.118
1PPT	117.7	0.182	0.515	0.195+	0.173	0.099	0.154
1NXB	110.6	0.182	0.415	0.198+	0.161	0.129+	0.123
2EBX	110.6	0.182	0.434	0.195+	0.166	0.122	0.131
1CRN	102.9	0.166	0.382	0.190	0.139−	0.092	0.100
1FX1	106.7	0.183	0.384	0.168	0.213+	0.092	0.195
Misfolded structures							
misHM	117.9+	0.169	0.401	0.189	0.132−	0.126+	0.100−
misIG	107.7	0.172	0.465	0.196+	0.137−	0.119	0.107−
Designed structures							
BEAL	108.1	0.178	0.451	0.169	0.192	0.083	0.175
BUND	106.7	0.200	0.385	0.214+	0.171	0.152+	0.150
BABA	103.9	0.177	0.444	0.173	0.184	0.088	0.163
TINY	105.3	0.174	0.439	0.170	0.181	0.089	0.159
FXNI	109.7	0.182	0.424	0.179	0.188	0.108	0.161

+, unusually high value; −, unusually low value; misHM, misfolded haemerythrin (CONGEN improved version); misIG, misfolded immunoglobulin (CONGEN improved version); 1GCR, gamma crystallin; 1GCN, glucagon; 3WGA, wheat germ agglutinin; 1MLT, melittin; 2FD1, ferredoxin; 1OVO, ovomucoid inhibitor; 1PFC, IG fragment; 2ABX, bungarotoxin; 2CDV, cytochrome C3 (−4 haems); 2CDVh, cytochrome C3 (+4 haems); 1CY3, cytochrome C3; 1PPT, pancreatic polypeptide; 1NXB, neurotoxin; 2EBX, erabutoxin; 1CRN, crambin; 1FX1, flavodoxin.

$$\mathrm{POLSURF} = \sum_{i=1}^{\mathrm{NATOM}} \left[\, |q(i)| \cdot a(i) \,\right]$$

$$\mathrm{POLFRAC} = \sum_{i=1}^{\mathrm{NATOM}} \left[\, |q(i)| \cdot a(i) \,\right] \Big/ \sum_{i=1}^{\mathrm{NATOM}} \left[\, a(i) \,\right]$$

where atom *i* has partial charge *q*(*i*) [values for united atoms as in Frömmel (1984) and Momany *et al.* (1974)] and surface area *a*(*i*) [calculated as in Kabsch and Sander (1983)]. Mathematically, the polar fraction is like the expectation value of the *absolute* value of partial charges, if the atomic surfaces *a*(*i*) are interpreted as statistical weights. The concept of polar fraction is similar to that of surface charge density. We use atomic units for partial charge (e) and Å^2 for surface area, so POLFRAC is in units of e.

Here, the atomic surface area *a*(*i*) can either be the external (solvent accessible) area or the internal (buried) surface area. The internal surface area reflects the extent to which parts of the protein cover other parts in the folded conformation of the protein, excluding solvent, and is defined as the difference between the external surface area of the extended and of the folded form. The atomic surface area can be averaged over all atoms, or over side-chain atoms and backbone atoms separately. The corresponding six characteristic quantities are: external/internal polar fraction for side chains/backbone/all atoms. Four of these quantities (e.g. the ones in Table II) carry non-redundant information.

The actual values of POLFRAC for natural proteins are approximately independent of molecular weight. This statement is true both for the external and internal polar fraction, averaged over all atoms or over side chains only; the relative fluctuation around the mean value is larger for smaller molecular weights (e.g. Figure 2B and C). The independence of molecular weight suggest that the polar fraction approximates a physical invariant of folded globular proteins, possibly related to solubility requirements in water. Miller *et al.* (1987a) have made similar observations in terms of differently defined quantities.

Interestingly, the protein exterior, the protein interior and a maximally unfolded chain are very similar in polarity on average:

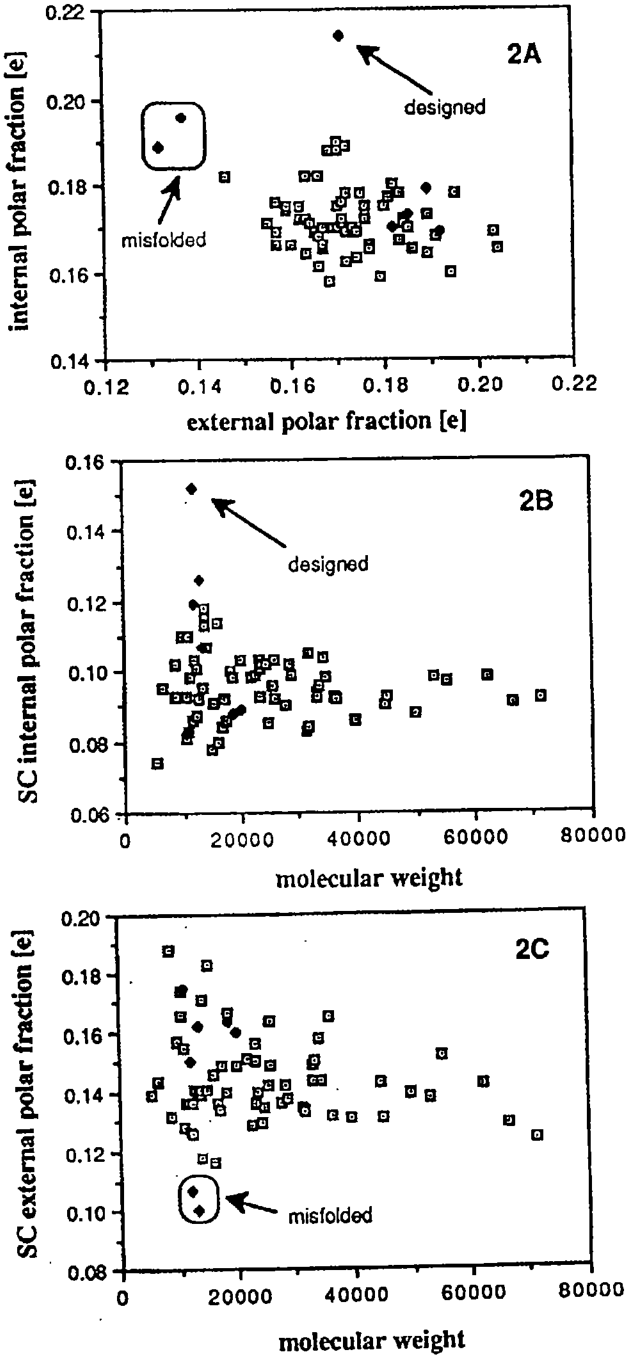

Fig. 2. Testing hypothetical structures by polarity characteristics. A misfolded (Novotny *et.al.*, 1984) or designed (Sander, 1987) structure [black diamonds (♦)] is regarded as questionable if its characteristic values fall outside of the cloud of reference values [64 proteins, dotted squares (⊡)] in the *n*-dimensional space of *n* characteristic values. In practice, 1- and 2-D projections are used, as in (A), (B) and (C). (A) Values of external polar fraction and internal polar fraction (side-chain plus main-chain atoms) scatter around their (common!) mean value 0.172 e. Three of the seven [♦] hypothetical structures are identified as questionable, four others are identified as normal. (B) The side-chain (SC) internal polar fraction clearly filters out one inadequate design, BUND, on the basis of an excessively polar interior. (C) The side-chain (SC) external polar fraction clearly identifies the two misfolded proteins (Novotny *et al.*, 1984) on the basis of an insufficiently polar external surface. For definition of polar fraction see text. The misfolded proteins are: the sequence of haemerythrin, a four-helix bundle protein, placed into the three dimensional structure of the *x*-chain of the mouse immunoglobulin variable domain, a structure with two four-stranded β-sheets; and *vice versa* (Novotny *et al.*, 1984); the data sets were refined using the CONGEN algorithm for generating conformations (Bruccoleri and Karplus, 1987). The designed (*de novo*) proteins are from the September 1986 EMBO course *Protein Design on Computers* (Sander, 1987): bundle (BUND), betaalphacin (BEAL), baburellin (BABA), tiny tim (TINY), and idealized flavodoxin (FXNI). A detailed description of the design exercises is available on request from the BIOcomputing Secretary, EMBL, 6900 Heidelberg, FRG. The coordinates are available on electronic mail from the bitnet file server NETSERV@EMBL. Send an email message containing the string 'help' or the string 'SEND $PDB:xxxx.BRK__MOD', where xxxx = BUND etc. is the four-letter protein identifier, PDB stands for Protein Data Bank, BRK for Brookhaven and MOD for model.

the mean values of the polar fraction (averaged over all atoms and all proteins) is 0.172 e for all three, with a scatter of about ±10% for the internal and ±20% for the external polar fraction (Figure 2A). To the extent that this statement is true for individual proteins, there is no net fractionation of polarity (as quantified here) in the folding process. A literally interpreted oil-drop model of proteins appears not be valid: the protein interior is approximately as polar as the exterior, due to the considerable concentration of polar peptide units not accessible to solvent in helices and sheets (see also Lee and Richards, 1971; Chothia, 1976; Richards, 1977; Rose *et al.*, 1985; Miller *et al.*, 1987a). However, the classical notion of a non-polar protein interior is reflected in the side-chain averages: the protein exterior has a much larger (0.143 e) average side-chain polar fraction than does the interior (0.098 e) (Table II). Interestingly, Miller *et al.* (1987a) in addition have pointed out that the protein interior is particularly strongly depleted in charged side chains.

In the database of known structures we find several unusually high and low values of polar fraction (last four columns of Table III): a flavodoxin (data set 1FX1, original authors explicitly state that the structure was manually built and not regularized nor refined), melittin (1MLT, see above), crambin (1CRN, small 46-residue plant seed protein with three S-S bridges), α-bungarotoxin (2ABX, see above), sea snake erabutoxin (1NXB and 2EBX, four S-S bridges in 62 residues), avian pancreatic polypeptide (1PPT, small 36-residue helix-loop peptide), multi-haem cytochrome C3 (2CDV, see above) and wheat germ agglutinin (3WGA, see above). Excluding these proteins, we define as normal the ranges of the four types of polar fraction values given in Table II.

Evaluation of hypothetical structures by polarity criteria

Ideally, characteristic quantities not only have a clear physical meaning but are also useful in practice. The available testing ground consists of deliberately misfolded proteins (Novotny *et al.*, 1984), which are clearly wrong, and of proteins designed *de novo* (Sander, 1987), which may or may not be wrong.

When the polarity values were calculated for these hypothetical proteins, the misfolded proteins have too few polar side-chain atoms exposed on the surface, with a side-chain external polar fraction of 0.100–0.107 e, compared to the average of 0.143 e and a lower bound in the database of 0.115 e (Figure 2C). The misfolded haemerythrin has too many polar side-chain atoms in the interior, with a side-chain internal polar fraction of 0.126 e, compared to the average of 0.098 e and an upper bound in the database of 0.122 e (Figure 2B).

As a realistic test of the usefulness of the polarity criteria applied to new protein designs, consider five protein sequence-structure pairs designed *de novo* as an exercise during an EMBO course in 1986 (Sander, 1987). These designs looked reasonable from many points of view (visual examination, secondary structure preferences, energy minimization, etc.). However, applying our criteria, one of them can be rejected because of an excessively high internal polar fraction of side chains: POLFRAC (side chains, internal) = 0.152 e compared to the average of about 0.098 e and an upper limit in the database of 0.122 e (Figure 2B). Clearly, the designed sequence has very little chance of leading to successful experimentation, unless modified.

These results show that our criteria are a first step toward theoretical screening of sequence-structure designs, of 3-D structures model-built by homology and even of experimental structures. However, as the approach here is based on empirical and statistical analysis of the database, which may not be fully representative of the range of physically permissible structures, we are only able to identify a hypothetical structure as questionable and not yet able to definitively prove or disprove the validity of a design.

Other recent attempts to evaluate misfolded proteins are based on the notion of hydrophobicity. Eisenberg and McLachlan (1986) use the relationship between accessible surface area and energy of transfer from water or organic solvent to estimate the hydrophobic stabilization of proteins. They obtain lower than normal hydrophobic stabilization energy for the misfolded proteins (refined data sets) of Novotny *et al.* (1984) compared to their natural counterparts, but provide no systematic criterion applicable to protein design. Bryant and Amzel (1987) count hydrophobic neighbour contacts in protein structures. Hydrophobic contact counts are approximately invariant for different proteins, but the criterion based on them is less discriminating than the one used here and cannot easily be extrapolated to higher molecular weights.

Conclusion

Our conclusion is that characteristic quantities with sufficiently narrow scatter around their canonical value can be used to improve protein design. The normal range of values can be used as a filter, identifying hypothetical protein structures as questionable when their value of the characteristic quantity falls outside of the range of values in the database of reference proteins. The mean value of the characteristic quantity can be used as a constructive criterion: bringing a protein closer to the mean value improves the chances of a successful design.

We expect continued progress in the search for physically diverse and more refined criteria and further quantitative and qualitative improvement in the database of known structures. A computer program, POL__DIAGNOSTICS__88, capable of applying the current set of criteria to a data set of protein coordinates (all atoms), is available on request (academic or other license agreement).

Acknowledgements

We thank R.E.Bruccoleri and J.Novotny for their refined (using the CONGEN algorithm by Bruccoleri and Karplus, 1987) data sets of two misfolded proteins and Arthur Lesk for making us aware of the new ferredoxin structure (Stout *et al.*, 1988). We are deeply grateful to the crystallographers who have made protein coordinates available through the Protein Data Bank.

Note added in proof

J.Novotny, A.A.Rashin and R.E.Bruccoleri [*Proteins*, 4, 19–30 (1988)] report successful discrimination between misfolded and native structures using the following criteria: (i) solvent-exposed side-chain non-polar surface, (ii) number of buried ionizable groups and (iii) empirical free-energy functions that incorporate solvent effects.

References

Bernstein,F.C. *et al.* (1977) *J. Mol. Biol.*, **112**, 535–542.
Biernat,J., Hasselmann,H., Hofer,B., Kennedy,B. and Koester,H. (1987) *Prot. Eng.*, **1**, 345–351.
Bryant,S.H. and Amzel,L.M. (1987) *Int. J. Peptide Protein Res.*, **29**, 46–52.
Bruccoleri,R.E. and Karplus,M. (1987) *Biopolymers*, **26**, 137–168.
Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–12.
Eisenberg,D. and McLachlan,A.D. (1986) *Nature*, **319**, 199–203.
Frömmel,C. (1984) *J. Theor. Biol.*, **111**, 247–260.
Gates,R.E. (1979) *J. Mol. Biol.*, **127**, 345–351.
Janin,J. (1976) *J. Mol. Biol.*, **105**, 13–14.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Klein,P., Kanehisa,M. and DeLisi,C. (1984) *Biochim. Biophys. Acta*, **787**, 221–226.
Lee,B. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
Momany,F.A., McGuire,R.F., Burgess,A.W. and Scheraga,H.A. (1974) *J. Phys. Chem.*, **79**, 2361–2381.
Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987a) *J. Mol. Biol.*, **196**, 641–656.
Miller,S., Lesk,A.M., Janin,J. and Chothia,C. (1987b) *Nature*, **328**, 834–836.
Novotny,J., Bruccoleri,R.E. and Karplus,M. (1984) *J. Mol. Biol.*, **177**, 787–818.
Richards,F.M. (1977) *Ann. Rev. Biophys. Bioeng.*, **6**, 151.
Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) *Science*, **229**, 834–838.
Sander,C. (ed.) (1987) EMBL BIOcomputing Technical Document 1.
Shrake,A. and Rupley,J.A. (1973) *J. Mol. Biol.*, **79**, 351–371.
Stout,G.H., Turley,S., Sieker,L.C. and Jensen,L.H. (1988) *PNAS*, **85**, 1020–1022.
Teller,D.C. (1976) *Nature*, **260**, 729–731.

Received on September 20, 1988; revised on November 9, 1988