J. Appl. Cryst. (1993). **26**, 47–60

Quality Control of Protein Models: Directional Atomic Contact Analysis

BY G. VRIEND AND C. SANDER

EMBL, Meyerhofstrasse 1, 6900 *Heidelberg, Germany*

(*Received* 28 *November* 1991; *accepted* 30 *July* 1992)

Abstract

Branden & Jones state, in *Nature*: 'Protein crystallography is an exacting trade, and the results may contain errors that are difficult to identify. It is the crystallographer's responsibility to make sure that incorrect protein structures do not reach the literature.' [Branden & Jones. (1990). *Nature* (*London*), **343**, 687–689.] One of several available methods of checking structures for correctness is the evaluation of atomic contacts. From an initial hypothesis that atom–atom interactions are the primary determinant of protein folding, any protein model can be tested for proper packing by the calculation of a contact quality index. The index is a measure of the agreement between the distributions of atoms around each residue fragment in the model and equivalent distributions derived from the database of known structures solved at high resolution. The better the agreement, the higher the contact quality index. This empirical test, which is independent of X-ray data, is applied to a series of successively refined crystal structures. In all cases, the model known or expected to be better (the one with the lower R factor) has a better contact quality index, indicating that this type of contact analysis can be used as an independent quality criterion during crystallographic refinement. Modelled proteins and predicted mutant structures can also be evaluated.

Introduction

Assume that, at some stage of density fitting, X-ray refinement, structure determination from two-dimensional NMR data or even protein design or structure prediction of mutants, a complete atomic model has been generated. We ask how such a model can be tested for correctness. Has the chain been correctly traced? Have all amino acid side chains been correctly placed?

In crystallography, the quality of a model is expressed by an R factor. With an R factor below 20.0, it can be safely assumed that the model is essentially correct. However, some recent experiences, summarized by Branden & Jones (1990), have made clear that at intermediate R factors, around $R = 25.0$ for example, there is no guarantee that the model is correct.

Many indicators exist for the quality of structures. A φ–ψ plot is often very informative and real-space R-factor plots can indicate bad spots. Here an attempt is made to evaluate the correctness of models by a purely empirical, *i.e.* non-physical, method based on generalizations of observations recorded in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977).

Contacts in known protein structures have been analysed before (Warme & Morgan, 1978; Levitt & Perutz, 1988; Singh & Thornton, 1985, 1990; Burley & Petsko, 1985, 1986; Crippen & Kuntz, 1978; Reid, Lindley & Thornton, 1985; Richmond & Richards, 1978; Tanaka & Scheraga, 1975; Miyazawa & Jernigan, 1985), but have not been satisfactorily used for predictive or analytical purposes. These studies have focused on the occurrence of residue–residue interactions, on the occurrence of atom–atom interactions or on the relative spatial orientation of residue–residue interactions. Here, as one partner of a contact event, a fragment of a residue without internal degrees of freedom has been taken and, as the other partner, an atom described by its chemical type. Each contact event is thus characterized by the fragment type, the atom type and the three-dimensional location of the atom relative to the local frame of the fragment. The resulting database-derived distributions are used for contact quality control by comparing them with the actual distributions in the protein structures being tested for correctness.

The method

Fragment types

The 20 amino acid types are divided into 80 fragments. These fragments are chosen so that they are as large as possible, but do not contain rotational degrees of freedom around dihedral angles. All hydrogen atoms are neglected. The largest fragment is the tryptophan double-ring system; the smallest contains only three atoms (*e.g.* the three consecutive atoms in the lysine side chain). Table 1 shows how this division into fragments is made.

0021-8898/93/010047-14$06.00

Table 1. *Atoms used per fragment*

The atoms listed are used to superimpose fragments on the master fragment. Primed atoms are the so-called unique atoms that are used to score this fragment.

N' $C^{\alpha\prime}$ C^{β}	All residues except Pro, Gly
N' $C^{\alpha\prime}$ C	Gly
N' $C^{\alpha\prime}$ $C^{\beta\prime}$ $C^{\gamma\prime}$ $C^{\delta\prime}$	Pro
C^{α} C' O'	All residues
C^{α} C $C^{\beta\prime}$	Ala
C^{α} $C^{\beta\prime}$ $S^{\gamma\prime}$	Cys
C^{α} $C^{\beta\prime}$ C^{γ} (1)	Asp Glu Phe His Lys Leu Met Asn Gln Arg Trp Tyr
$C^{\gamma\prime}$ $O^{\delta 1\prime}$ $O^{\delta 2\prime}$	Asp
C^{β} $C^{\gamma\prime}$ C^{δ}	Glu Lys Gln
$C^{\delta\prime}$ $O^{\varepsilon 1\prime}$ $O^{\varepsilon 2\prime}$	Glu
C^{β} $C^{\gamma\prime}$ $C^{\delta 1\prime}$ $C^{\delta 2\prime}$ $C^{\varepsilon 1\prime}$ $C^{\varepsilon 2\prime}$ $C^{\zeta\prime}$	Phe
C^{β} $C^{\gamma\prime}$ $N^{\delta 1\prime}$ $C^{\delta 2\prime}$ $C^{\varepsilon 1\prime}$ $N^{\varepsilon 2\prime}$	His
C^{α} $C^{\beta\prime}$ $C^{\gamma 2\prime}$	Ile
C^{β} $C^{\gamma 1\prime}$ $C^{\delta 1\prime}$	Ile
C^{γ} $C^{\delta\prime}$ C^{ε}	Lys
C^{δ} $C^{\varepsilon\prime}$ $N^{\zeta\prime}$	Lys
$C^{\gamma\prime}$ $C^{\delta 1\prime}$ $C^{\delta 2\prime}$	Leu
C^{β} $C^{\gamma\prime}$ S^{δ}	Met
C^{γ} $S^{\delta\prime}$ $C^{\varepsilon\prime}$	Met
$C^{\gamma\prime}$ $O^{\delta\prime}$ $N^{\delta 2\prime}$	Asn
$C^{\delta\prime}$ $O^{\varepsilon 1\prime}$ $N^{\varepsilon 2\prime}$	Gln
C^{β} $C^{\gamma\prime}$ $C^{\delta\prime}$	Arg
$N^{\varepsilon\prime}$ $C^{\zeta\prime}$ $N^{\eta 1\prime}$ $N^{\eta 2\prime}$	Arg
C^{α} $C^{\beta\prime}$ $O^{\gamma\prime}$	Ser
$C^{\beta\prime}$ $O^{\gamma 1\prime}$ $C^{\gamma 2\prime}$	Thr
$C^{\beta\prime}$ $C^{\gamma 1\prime}$ $C^{\gamma 2\prime}$	Val
C^{β} $C^{\gamma\prime}$ $C^{\delta 1\prime}$ $C^{\delta 2\prime}$ $N^{\varepsilon 1\prime}$ $C^{\varepsilon 2\prime}$ $C^{\varepsilon 3\prime}$ $C^{\zeta 2\prime}$ $C^{\zeta 3\prime}$ $C^{\eta 2\prime}$	Trp
C^{β} $C^{\gamma\prime}$ $C^{\delta 1\prime}$ $C^{\delta 2\prime}$ $C^{\varepsilon 1\prime}$ $C^{\varepsilon 2\prime}$ $C^{\zeta\prime}$ $O^{\eta\prime}$	Tyr

Atom types

The 20 amino acids consist of 167 atom types when all of the atoms in each residue are labeled separately. The atom types are grouped into 57 atom classes in order to increase the number of observable contacts per class as far as possible. For example, atoms $O^{\varepsilon 1}$, and $O^{\varepsilon 2}$ of Glu are grouped with $O^{\delta 1}$ and $O^{\delta 2}$ of Asp in one class of atoms, C^{γ} of Phe and Tyr make another, $N^{\delta 2}$ of Asn and $N^{\varepsilon 2}$ of Gln make a third, and so on. Any subdivision has a certain degree of arbitrariness to it and it is therefore hoped that the number of high-quality protein structures deposited in the PDB will increase rapidly so that all 167 atom types can be treated individually.

The 57 classes of atom types are: (1) N all except Pro, Gly; (2) N Pro; (3) N Gly; (4) C^{α} all except Pro, Gly; (5) C^{α} Pro; (6) C^{α} Gly; (7) C (OOH) all except Pro, Gly; (8) C (OOH) Pro; (9) C (OOH) Gly; (10) O all except Pro, Gly; (11) O Pro; (12) O Gly; (13) C^{β} Ala; (14) C^{β} Cys; (15) S^{γ} Cys; (16) C^{β} Asp, C^{β} Asn, C^{β} Glu, C^{β} Gln; (17) C^{γ} Asp, C^{δ} Glu; (18) $O^{\delta 1}$ Asp, $O^{\delta 2}$ Asp, $O^{\varepsilon 1}$ Glu, $O^{\varepsilon 2}$ Glu; (19) C^{γ}, Glu, C^{γ} Gln; (20) C^{β} Phe, C^{β} Tyr, C^{β} Trp; (21) C^{γ} Phe, C^{γ} Tyr; (22) $C^{\delta 1}$ Phe, $C^{\delta 2}$ Phe, $C^{\varepsilon 1}$ Phe, $C^{\varepsilon 2}$ Phe, C^{ζ} Phe, $C^{\delta 1}$ Tyr, $C^{\delta 2}$ Tyr, $C^{\varepsilon 3}$ Trp, $C^{\zeta 2}$ Trp, $C^{\zeta 3}$ Trp, $C^{\eta 2}$ Trp; (23) C^{β} His; (24) C^{γ} His; (25) $N^{\delta 1}$ His, $N^{\varepsilon 2}$ His; (26) $C^{\delta 2}$ His, $C^{\varepsilon 1}$ His; (27) C^{β} Ile, C^{γ} Leu, C^{β} Val; (28) $C^{\gamma 1}$ Ile; (29) $C^{\gamma 2}$ Ile, $C^{\delta 1}$ Ile, $C^{\delta 1}$ Leu, $C^{\delta 2}$ Leu, $C^{\gamma 1}$ Val, $C^{\gamma 2}$ Val; (30) C^{β} Lys, C^{γ} Lys, C^{δ} Lys, C^{β} Arg, C^{γ} Arg; (31) C^{ε} Lys; (32) N^{ζ} Lys; (33) C^{β} Leu; (34) C^{β} Met; (35) C^{γ} Met; (36) S^{δ} Met; (37) C^{ε} Met; (38) C^{γ} Asn, C^{δ} Gln; (39) $O^{\delta 1}$ Asn, $O^{\varepsilon 1}$ Gln; (40) $N^{\delta 2}$ Asn, $N^{\varepsilon 2}$ Gln; (41) C^{β} Pro, C^{γ} Pro; (42) C^{δ} Pro; (43) C^{δ} Arg; (44) N^{ε} Arg; (45) C^{ζ} Arg; (46) $N^{\eta 1}$ Arg, $N^{\eta 2}$ Arg; (47) C^{β} Ser, C^{β} Thr; (48) O^{γ} Ser, $O^{\gamma 1}$ Thr; (49) $C^{\gamma 2}$ Thr; (50) C^{γ} Trp; (51) $C^{\delta 1}$ Trp; (52) $C^{\delta 2}$ Trp; (53) $N^{\varepsilon 1}$ Trp; (54) $C^{\varepsilon 2}$ Trp; (55) $C^{\varepsilon 1}$ Tyr, $C^{\varepsilon 2}$ Tyr; (56) C^{ζ} Tyr; (57) O^{η} Tyr.

Fragment environment (box)

The contact probability densities for each atom class around each fragment type are calculated on grid points arranged in a cubic box around each fragment. Each of the 80 fragment types is placed in a standard orientation at the centre of a box of 16^3 cubically arranged grid points spaced at 1 Å intervals, with one box for every fragment–atom combination. The fragment at the centre of the box is called the 'master fragment'. There are $80 \times 57 = 4560$ boxes for helical residues and the same number for non-helical residues.

The box dimensions were chosen to be $16 \times 16 \times 16$ Å because these give the smallest box that can contain the largest fragment (the tryptophan double-ring system) and all its contacting atoms. Smaller

Table 2. *PDB and chain identifiers of the proteins used to determine the probability distributions*

1crn	2sec E	3gap A	1pp2 R	451c	2pka B
1ctf	3adk	9wga A	3app	3c2c	3rp2
1gcr	2ccy A	4dfr A	2sga	1nxb	6cha
1gp1 A	2cpp	4fxn	2gdl 0	1mcb	1ton
1hip	3est	1rnt	6ldh	2hhb A	1sgt
1hmq	2cyp	3icb	2act	2hhb B	2prk
4ins A	3grs	3cln	9pap	2alp	2ci2 I
1pcy	2gn5	5tnc	2aza A	4sgb I	4pfk A
1rn3	2lzm	2cab	2cdv	1tpp	1prc C
1ubq	3wrp	1bp2	4fd1	2pka A	1prc L

fragments could of course be placed in smaller boxes, but we decided to make all the boxes of equal size for simplicity of implementation of the algorithm.

Calculation of contact probability distributions in each box

The probabilities of occurrence of atoms of a certain class at the grid points around a fragment type are estimated by the following procedure:

(i) Loop over all proteins in a non-redundant database. The proteins chosen are unique in sequence (no pair has more than 50% identical amino acids after alignment), are solved to at least 2.5 Å resolution and have a crystallographic R factor better than 25%. Table 2 lists the PDB files used.

(ii) For each of these proteins, loop over all fragments in all residues.

(iii) For each fragment, determine the transformation needed to place it in the standard orientation at the centre of the grid box and apply this transformation to the fragment and its environment.

(iv) For every residue in contact with the fragment, take all the atoms, including those that do not contact the fragment directly. For every atom on the contacting residue, add a value to all grid points falling within the van der Waals radius of this atom. Two atoms are considered to be in contact if the distance between them is less than the sum of their van der Waals radii plus 1.5 Å. The value 1.5 Å was not fine-tuned; it corresponds roughly to the radius of a water molecule.

(v) Discretization of probability values: the accumulation of values at grid points that lie within the van der Waals radius of atoms in contacting residues is performed by the Voorintholt method (Voorintholt, Kosters, Vegter, Vriend & Hol, 1989). This method is fast and not very sensitive to shifts in grid origin. It uses a truncated inverted parabola as local envelope function. The value V added to the grid point at position x at a distance d from the centre of an atom with van der Waals radius R is

$$\begin{aligned} V(d) &= 1 - d^2/R^2 \quad \text{if } d < R, \\ V(d) &= 0 \qquad\qquad\ \ \text{if } d \geq R. \end{aligned} \tag{1}$$

(vi) The final accumulated intensity values at every grid point in a box are normalized by the frequency of occurrence of the fragment type in the database. Other normalizations, for instance by the probability of a fragment–atom contact, were tested and found not to affect significantly the results of the application of the method.

Therefore, $P(s, \text{fr}, \text{atm}, x)$ [see (2)], when integrated over the grid points that fall within the volume of an atom, represents an estimated probability for an atomic occurrence derived from the database. Specifically, given a fragment of type fr in a protein, $P(s, \text{fr}, \text{atm}, x)$ is proportional to the conditional probability of finding an atom of type atm at location x near fragment fr.

In summary: the probability distributions $P(s, \text{fr}, \text{atm}, x)$ are determined by a sum over all fr′ and atm′ in the database:

$$P(s, \text{fr}, \text{atm}, x) = [1/F(\text{fr}, s)] \sum_{\text{fr}'} \sum_{\text{atm}'} V[d(x, \text{atm})] \times D(s, \text{fr}, \text{atm}, s', \text{fr}', \text{atm}'), \tag{2}$$

where fr is the fragment type (one of 80 different ones), atm is the atom type (one of 57 different ones), x is the position in the box (one of 4096 grid points), $P(s, \text{fr}, \text{atm}, x)$ is the probability distribution for an atom of type atm in the environment of a fragment of type fr in a residue with a secondary structure of type s, $F(\text{fr}, s)$ is the frequency of occurrence of fragment fr in a residue with secondary structure s in the database, $V(d)$ is the local envelope function centred on atm [see (1)], $d(x, \text{atm})$ is the distance from atom atm to grid point x, fr′ is the fragment of a particular residue in a particular protein, atm′ is the atom in a particular residue in a particular protein, $D(s, \text{fr}, \text{atm}, s', \text{fr}', \text{atm}')$ is the delta function, $D = 1$ if s', fr′ and atm′ are of the same secondary structure, fragment and atom type as s, fr and atm and if atm is in a residue in contact with fr, otherwise $D = 0$.

The packing is distinctly different around residues in helical and non-helical secondary-structure elements, respectively, so two sets of distributions (boxes) were derived and used, labelled helical and non-helical. The amount of data available in the protein structure database does not allow further statistically significant subdivision of the data.

The probability values can be represented as density maps in a manner similar to that described by (Rosenfield, Swanson, Meyer, Carrell & Murray-Rust, 1984). Fig. 1 shows an example of observed probability distributions for the occurrence of positively charged nitrogen atoms around phenylalanine side chains contoured at an intermediate level. The first set is for the central phenylalanine in an α-helix, the second, for that in β-strand.

Application of contact probability densities to protein models

Having calculated from the database $P(s, \text{fr}, \text{atm}, x)$, the conditional probability of occurrence of atoms of type atm at position x near a fragment of type fr, one can evaluate any observed contact event $(s, \text{fr}, \text{atm}, x)$ in a protein model and ask: is the contact event in agreement with what is expected from the database analysis? In other words, is the observed distribution of contacts with a fragment consistent with the database distribution? Mathematically, there are many ways of comparing distributions: χ squared, difference entropy *etc.* A complete comparison includes points of both low and high probability. Here we are faced with the notorious statistical problem of small numbers for grid points with low contact values; this may be due either to genuine low probability for the event (statistically reliable values) or to a small number of observations in the database for the particular box (statistically unreliable values). We choose here to compare only the distributions at points of high probability, disregarding grid points with small probability values. Technically, this is done by using a multiplicative measure of comparison, *i.e.* the product of the test distribution and the database distribution. In this way, only contact events which are frequent in the database and actually observed in the test case contribute (positively) to the quality index. Events which are either infrequent in the database or not observed in the test case are disregarded.

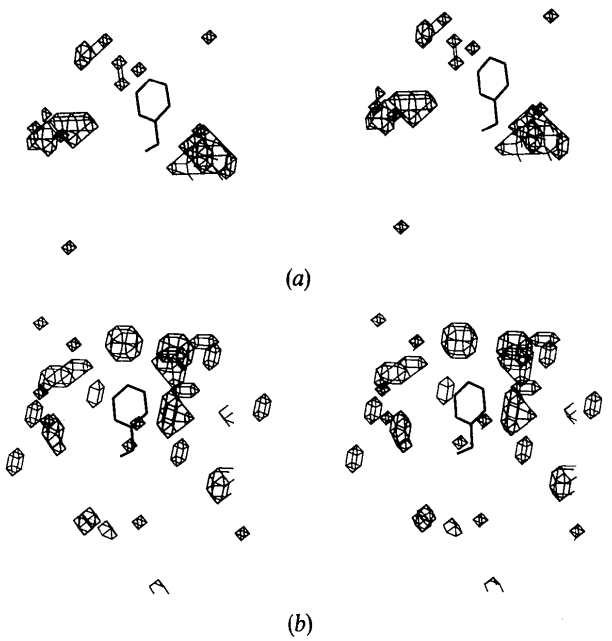


Fig. 1. Probability distribution for positively charged nitrogen atoms around the phenylalanine side chain, contoured at an intermediate level. Dotted lines are lysine N^{ζ}. Solid lines are arginine $N^{\eta 1}$ and $N^{\eta 2}$. (*a*) Helical phenylalanine. (*b*) Non-helical phenylalanine.

In detail, the empirical procedure for evaluating the quality of packing of a residue, a residue range or a whole protein is as follows:

(i) Loop over all fragments in the given residue (range).

(ii) For each fragment, determine the transformation to superpose this fragment onto the master fragment in the box of the appropriate type and apply this transformation to the fragment and its entire environment.

(iii) Determine all residues in the environment that make at least one atomic contact with the fragment.

(iv) Loop over all atoms in these contacting residues and determine the quality index $Q(\text{fr})$ for this fragment, defined as

$$Q(\text{fr}) = \sum_{\text{atm}} \sum_{x} A(\text{fr}) V[d(x, \text{atm})] P(s, \text{fr}, \text{atm}, x), \quad (3)$$

where fr is the fragment being evaluated, atm is the type of atom in the residue that makes a contact with fragment fr, x is the grid position around the fragment, $A(\text{fr})$ is the number of unique atoms in fragment fr, $V(d)$ is the envelope function [see (1)], $d(x, \text{atm})$ is the distance from the atom of type atm to the grid point x, $P(s, \text{fr}, \text{atm}, x)$ is the probability density for atom type atm in the environment of fragment type fr at grid position x. Fragment fr resides in a residue with secondary structure s.

Atoms that occur in more than one fragment contribute only to the $A(\text{fr})$ value of one fragment. The atoms that are used in each fragment are primed in Table 1. For points outside the box, $P(s, \text{fr}, \text{atm}, x)$ is zero.

(v) Sum the $Q(\text{fr})$ values over all fragments in the selected residue range.

All the atoms in contacting residues are considered, rather than only those atoms that make a contact with the fragment. In this way, the probability distributions also reflect the orientation of the contacting residues.

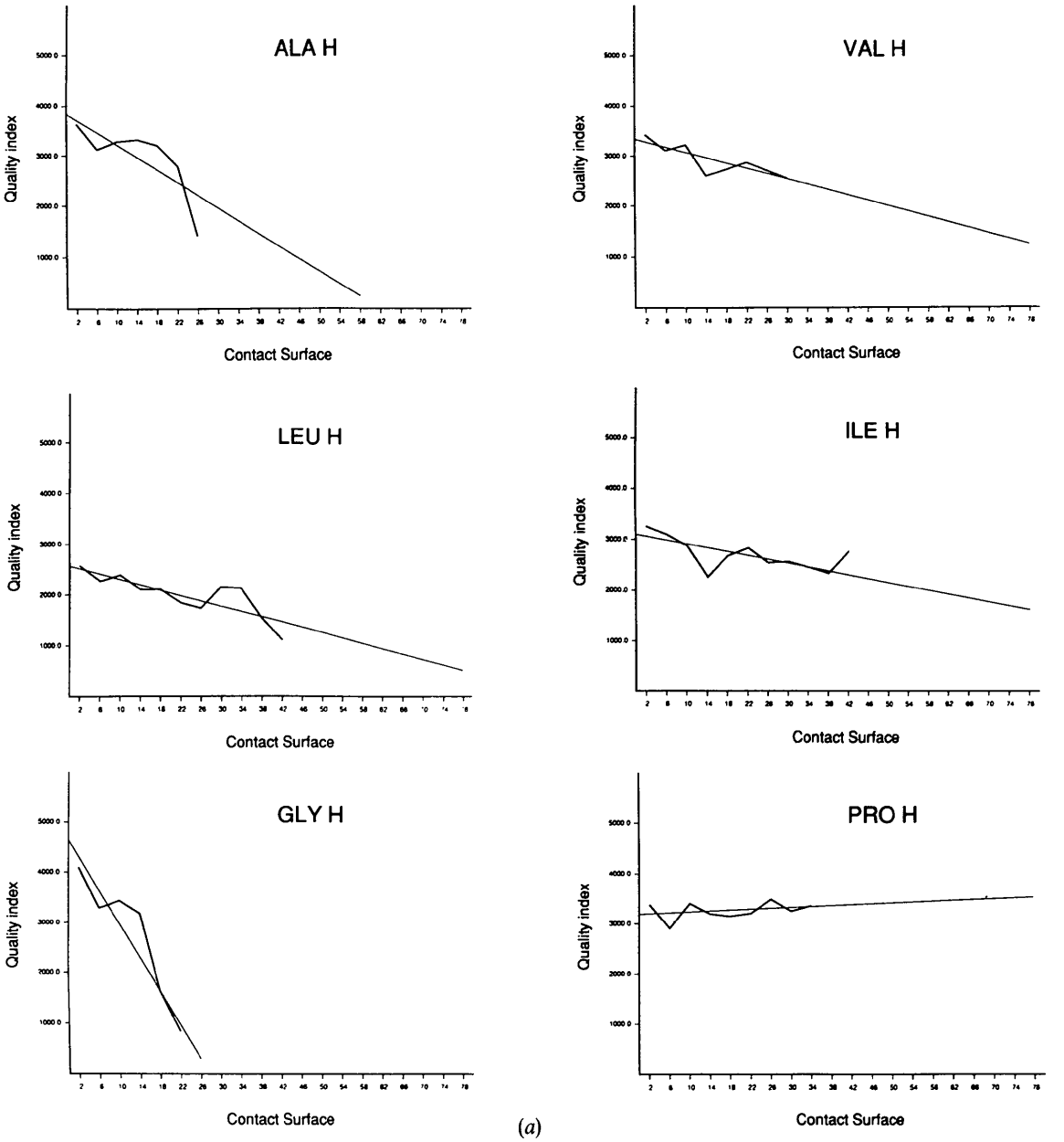
Interpretation and calibration of quality-index values

The quality index measures the agreement between the contacts observed in a particular protein structure and the contact probability distribution estimated from the database. Larger values mean that the observed contacts are more typically like those found in the database. When comparing, for example, two model structures with the same amino acid sequence, one model is interpreted to be 'better' if its quality index is higher.

When comparing different sequences or mutations in a given sequence, two problems should be taken into account. First, small residues make fewer contacts than large residues, so their quality-index

values can be lower even when packed perfectly. The solution to this problem is to compare the quality index for a residue with the average quality index for residues of the same type. Second, residues at the surface make fewer contacts than residues of the same type in the interior. If the surface contact area is complementary to the total number of contacts made by a residue (Colonna-Cesari & Sander, 1990), the second problem can be solved by determining the quality index as a function of the surface contact area and comparing it with the average quality index for residues of the same type and the same surface contact area. This provides an absolute scale for contact quality values of residue types.

To define this absolute scale, the evaluation procedure was applied to 154 proteins and the average quality index was determined as a function of residue type, secondary-structure type and surface contact area. Fig. 2 shows the un-normalized quality index as a function of the contact surface area for the 20 amino



(a)

Fig. 2. Average quality of residues as a function of their contact surface. The quality is given in arbitrary units. The straight lines indicate the best straight lines through these curves. The slope and abscissa cut-off values for these straight lines are given in Table 3. (*a*) Helical residues. (*b*) Non-helical residues.

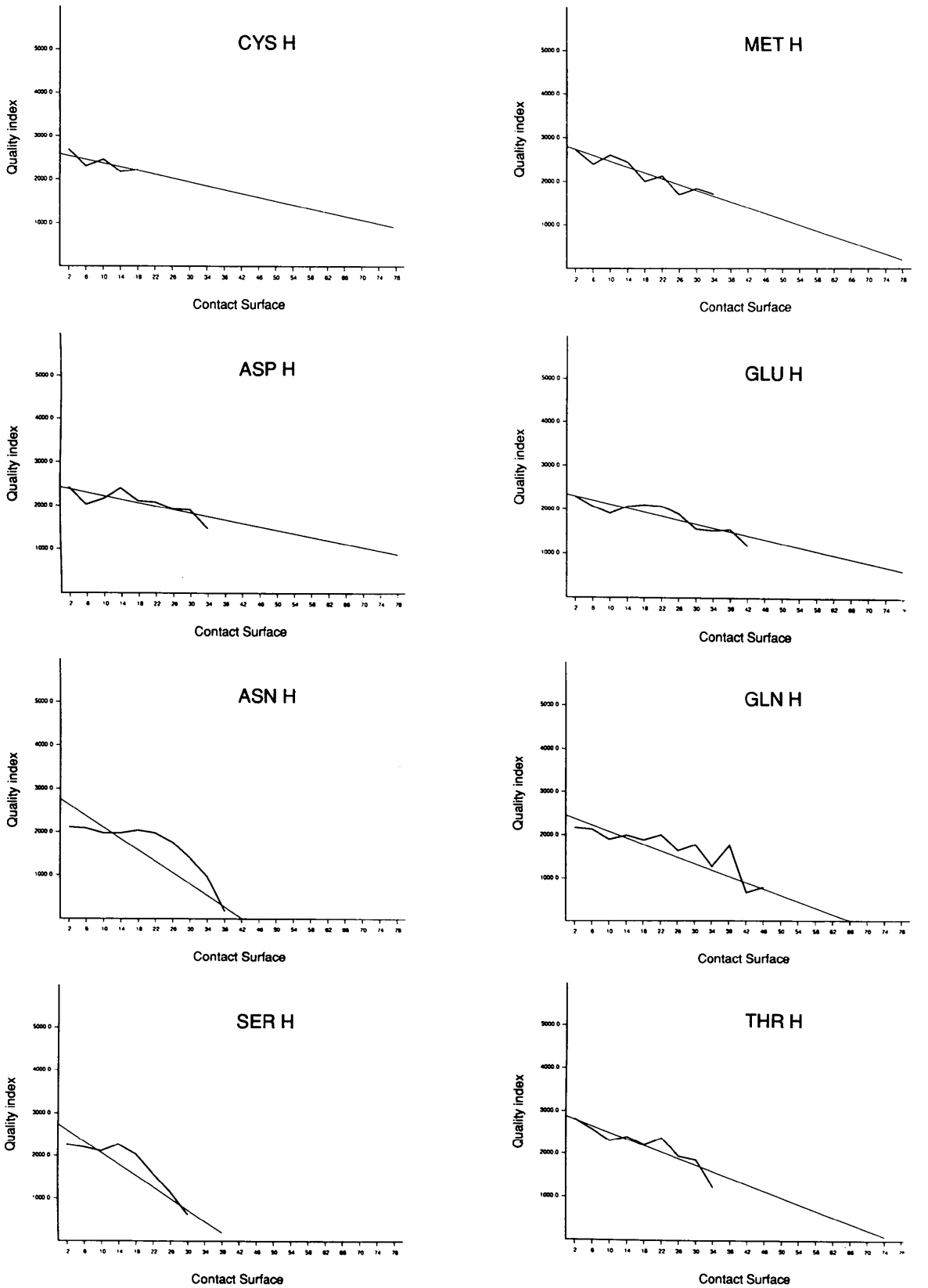

CYS H
MET H
ASP H
GLU H
ASN H
GLN H
SER H
THR H
Quality index
Contact Surface

Fig. 2(*a*) (*cont.* 1)

acids. Table 3 shows the slope, abscissa, cut-off and standard deviations determined from these curves. Quality-index values can then be given relative to the average value for this residue type with the same accessible contact surface in units of standard deviations. In this way, the quality-index values for different residues in different structural contexts are numerically comparable. If the normalized quality index is positive, the quality is higher than the average quality of the residues in the proteins of the test set. The interpretation of the shape of the curves shown in Fig. 2 is beyond the scope of this article.

The decision as to whether to use the raw quality index or the normalized quality index depends on the information required. For example, to determine the quality of mutations that are supposed to stabilize a protein by filling cavities, one would use the unnormalized quality index. For comparing properly folded and misfolded proteins or for evaluating the packing quality of an X-ray structure, the absolute scale is more appropriate.

Typically, the scaled quality index of residues ranges from -5σ to $+5\sigma$. In practice, a value of less than -5σ almost certainly means that something is

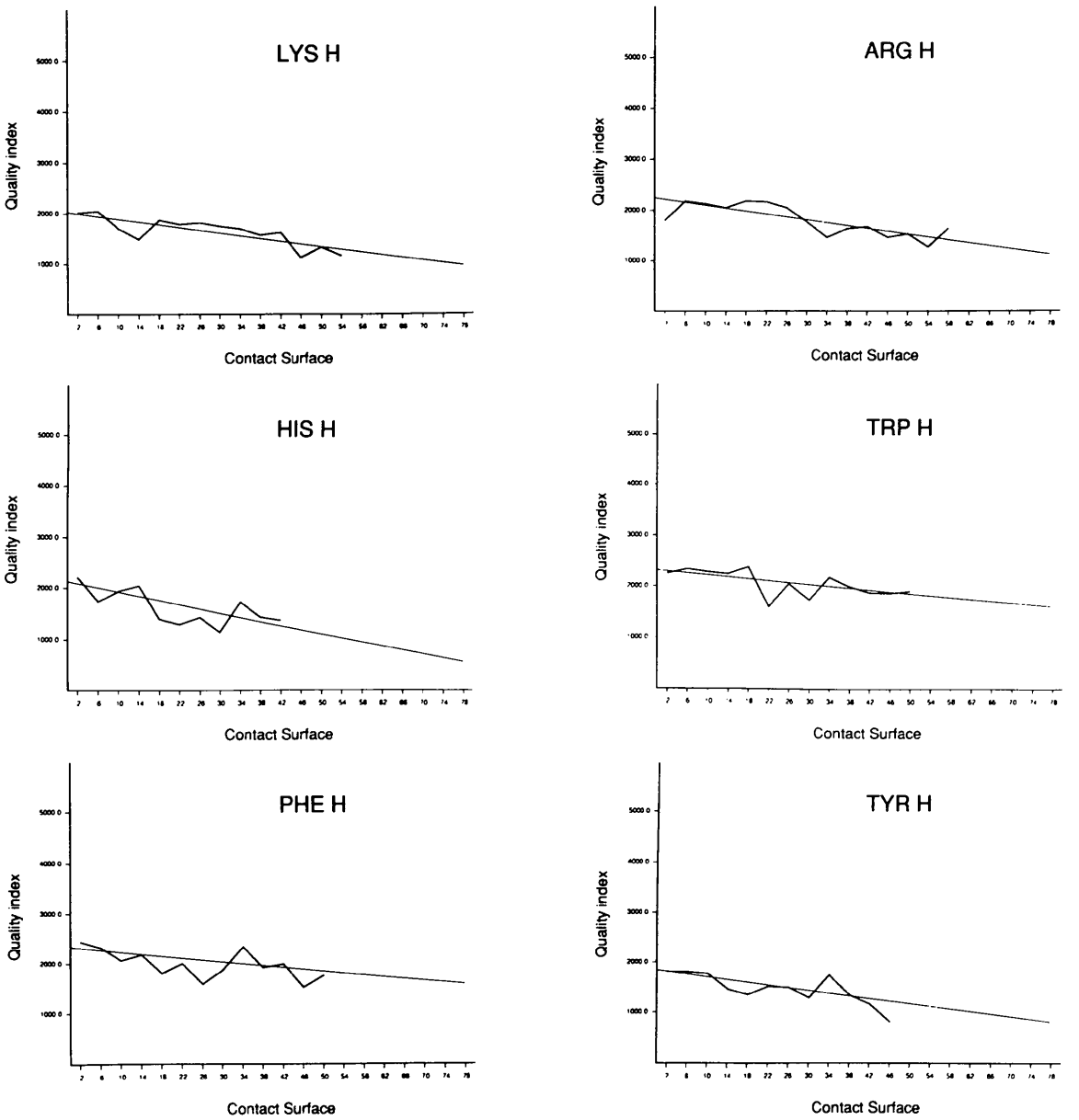


Fig. 2(*a*) (*cont.* 2)

'wrong'. In such cases, the residue is packed or built incorrectly or the residue has several contacts with a co-factor or is involved in crystal contacts without the other molecule(s) being present in the calculation.

What are the typical quality-index values for correct models, *e.g.* refined structures determined by X-ray crystallography? To answer this question, the absolute quality index was determined for 235 monomeric proteins for which the R factor and the resolution were available in the PDB. Fig. 3 shows a stereoplot of the quality index as a function of R factor and resolution. Most structures cluster around $R = 20.0$ and 2.0 Å resolution. After removal of the 16 structures that had the largest deviation from the centroid seen in Fig. 3, we find that the 'average monomeric structure' in the PDB has $R = 17.7$ with a standard of 2.3. The average resolution is 1.95 (34) Å. The average quality index is -0.59 with a standard deviation of 0.42.

What are the typical quality indices for incorrect structures? 26 misfolded structures were described by Holm & Sander (1992). At first glance, these models

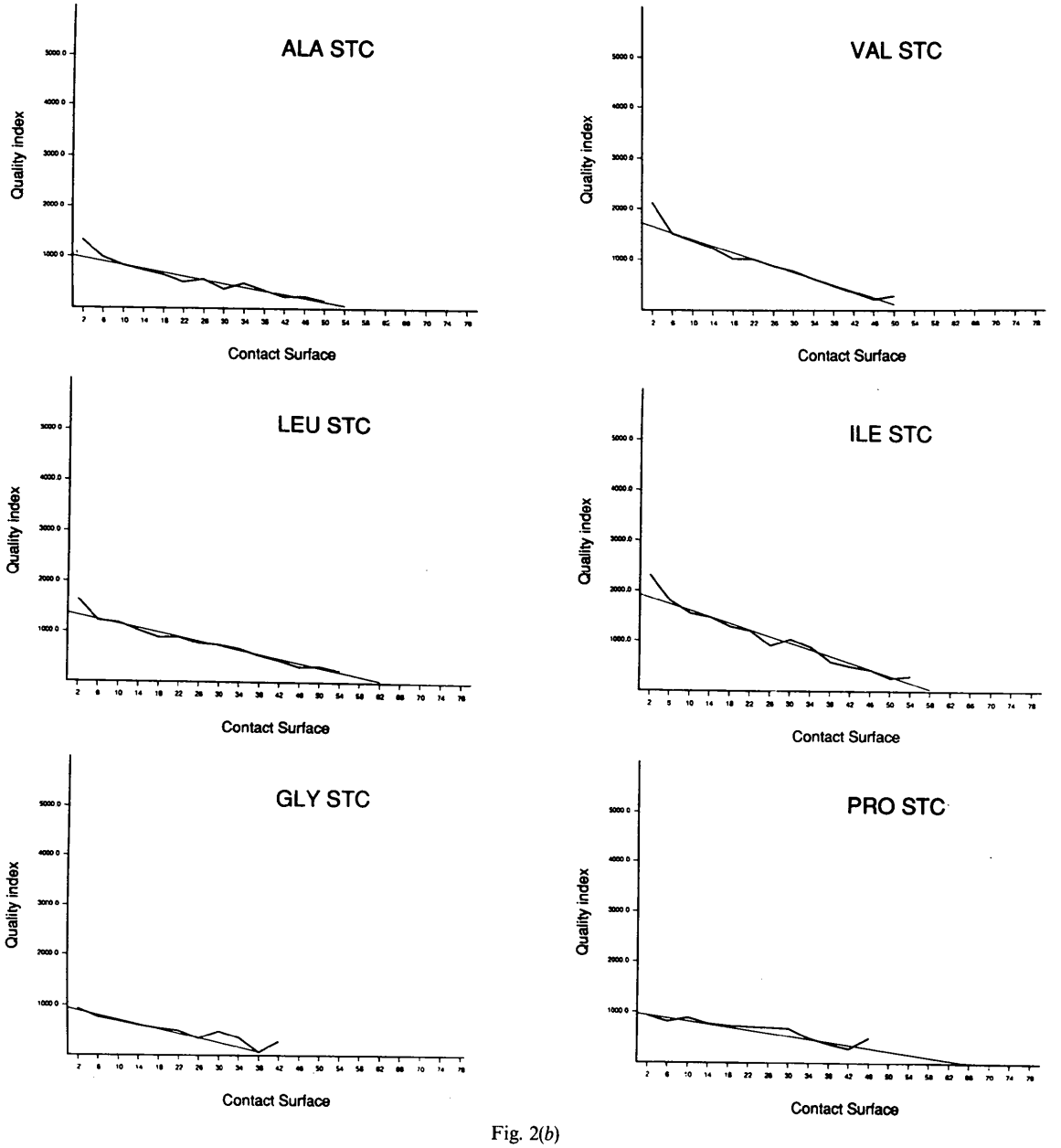


Fig. 2(*b*)

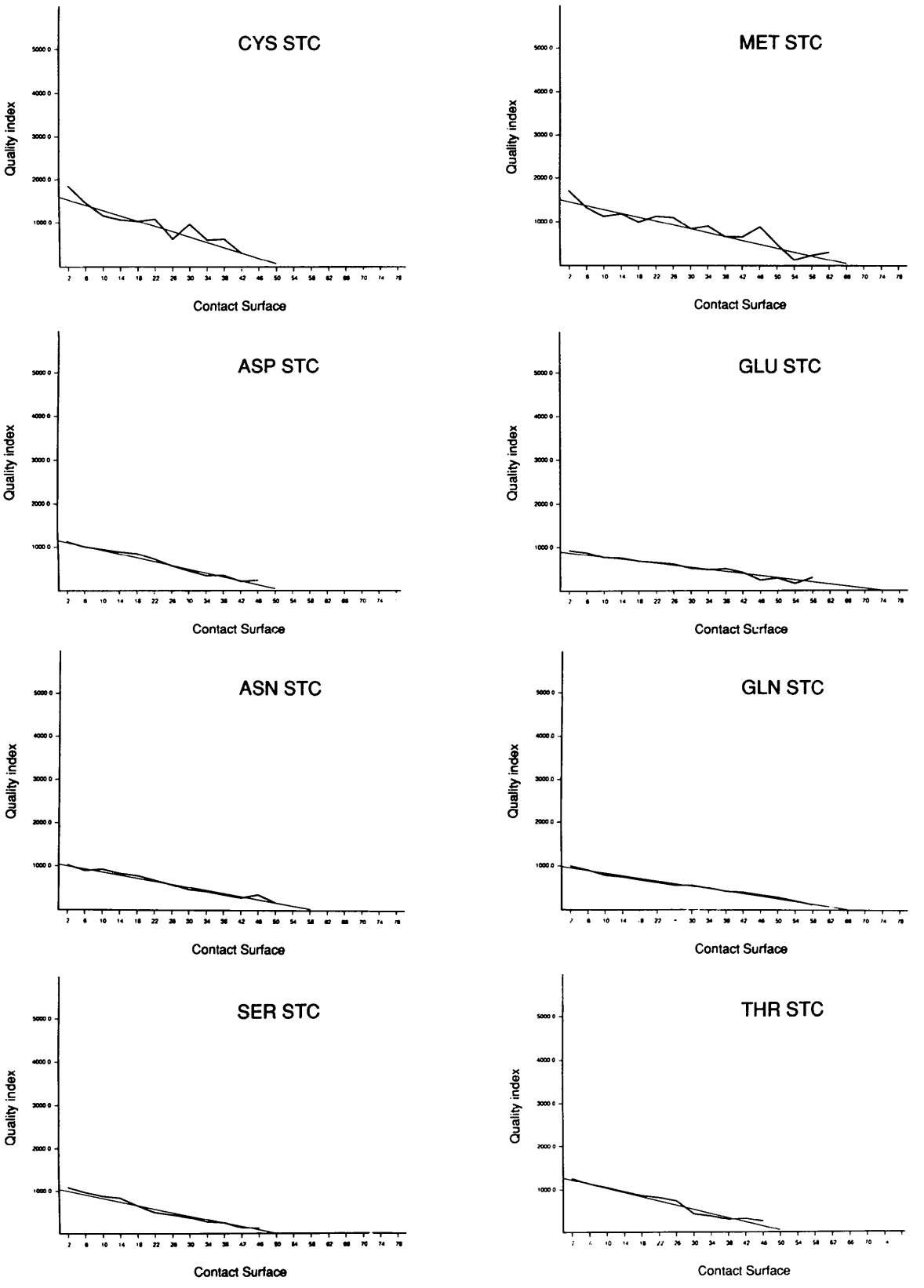

CYS STC
MET STC
ASP STC
GLU STC
ASN STC
GLN STC
SER STC
THR STC
Quality index
Contact Surface

Fig. 2(*b*) (*cont.* 1)

look perfectly normal but they are nevertheless wrong. The average quality index for these incorrect structures was −2.07 (33); after extensive energy minimization the average quality index became −1.80 (37). Only one of the 26 misfolded structures had a quality index that fell within two standard deviations of the average for 'correct' structures. We have derived the following two rules of thumb which are shown in Fig. 4:

(i) Any structure with quality index below −2.5 is very probably wrong.

(ii) Any structure with quality index below −1.2 should be treated with great caution.

While testing all the proteins in the PDB we discovered two crystal structures which are most likely not correct. We have informed the original authors of our findings.

Note added in proof: Since submission of this manuscript, one of the two structures has been independently determined by nuclear magnetic resonance spectroscopy, which has confirmed that the original crystal structure was incorrect.

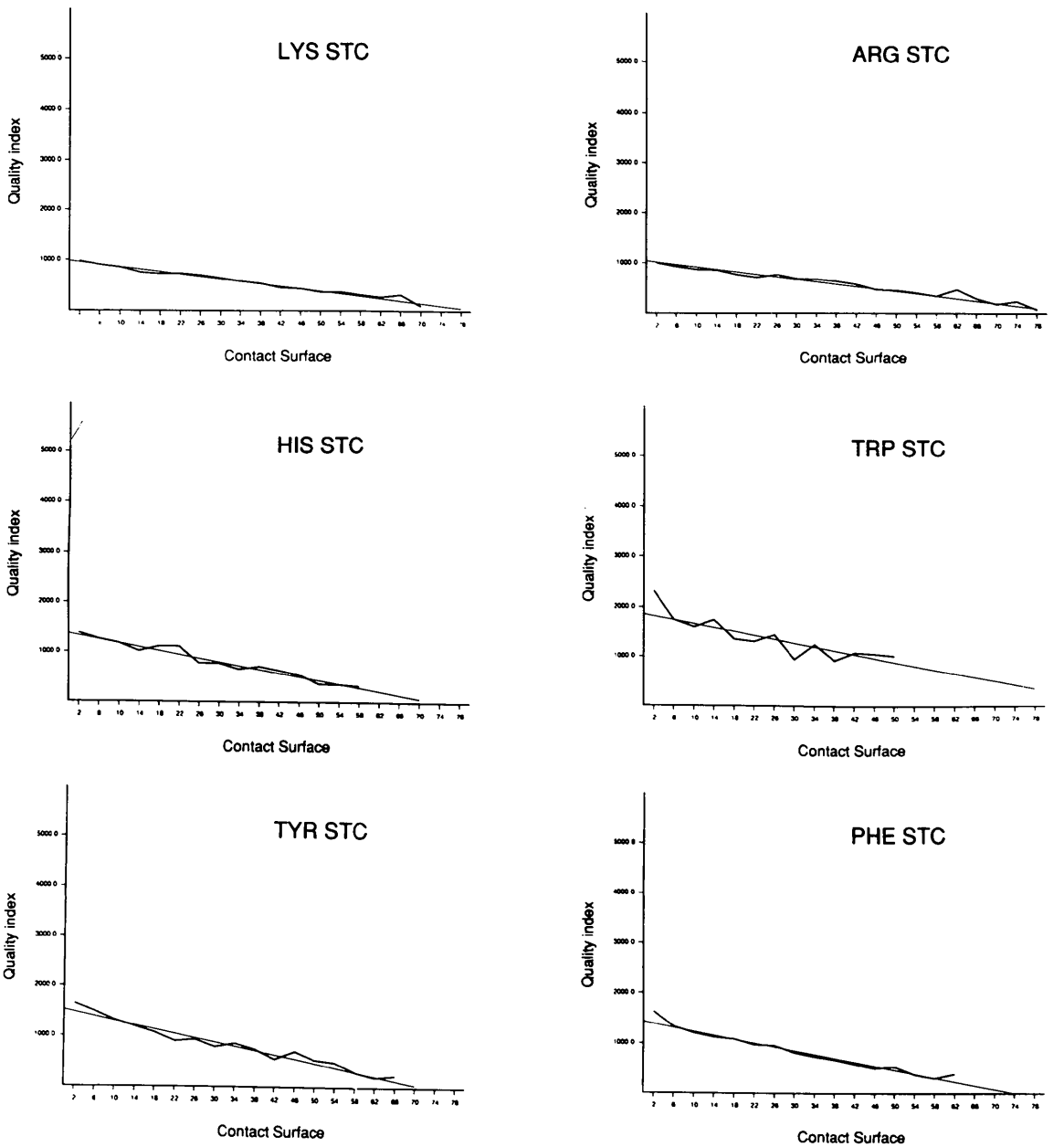


Fig. 2(*b*) (*cont.* 2)

Table 3. *Abscissa cut-off (C) and slope (S) for the averaged quality as a function of accessibility.*

The quality is given in arbitrary units. ΔC and ΔS are the standard deviations in C and S, respectively.

Amino acid	Helix				Sheet			
	C	ΔC	S	ΔS	C	ΔC	S	ΔS
Ala	3479	1318	−77	89	1009	275	−22	7
Cys	2284	948	−27	65	1582	304	−38	9
Asp	2330	701	−25	36	1133	276	−28	7
Glu	2364	560	−27	24	922	164	−15	4
Phe	2378	540	−12	18	1487	215	−24	5
Gly	4592	1313	−209	77	982	242	−29	7
His	2158	393	−25	16	1406	231	−23	5
Ile	3055	920	−24	41	1960	367	−40	9
Lys	2014	574	−17	18	989	164	−15	3
Leu	3580	730	−33	30	1403	221	−27	5
Met	3843	768	−42	40	1494	219	−27	4
Asn	2835	614	−82	19	1017	202	−22	5
Pro	3189	699	5	29	982	274	−18	9
Gln	2461	515	−46	17	991	149	−18	3
Arg	2310	461	−18	14	1019	139	−15	2
Ser	2787	754	−84	32	1012	247	−26	6
Thr	2940	863	−47	44	1278	314	−31	10
Val	3201	1189	−33	66	1808	368	−40	9
Trp	2313	436	−11	14	1904	381	−24	12
Tyr	1908	438	−17	17	1501	191	−26	3

Testing the method

Asymmetry in contact distributions

The method relies on the hypothesis that residue–residue contacts are distributed asymmetrically in globular proteins. The asymmetry in residue–residue contact distributions has been assessed using the atomic contact option of the relational protein structure database module of the program *WHAT IF* (Vriend, 1990). The database module allows database searches for contact distributions in terms of residues, fragments of residues or individual atoms according to chemical type and spatial arrangement. Contact distributions around one master fragment can be displayed or, optionally, probability density distributions can be calculated and visualized.

Two examples of contact distributions are shown in Fig. 5. Figs. 5(*a*) and (*b*) show the distribution of

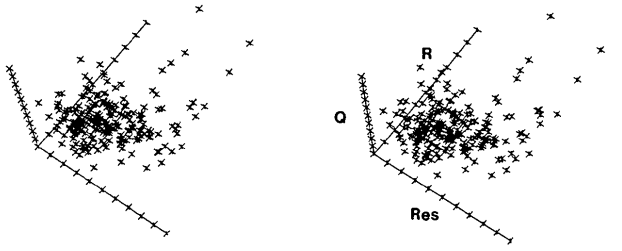


Fig. 3. Quality index (Q) as a function of resolution (Res) and R factor (R) (stereo representation). Each cross represents one of the 235 momomeric proteins mentioned in the text. Resolution range is 1.2–3.2 Å. R ranges from 9.0 to 35.0. Q ranges from −3.3 to +0.3.

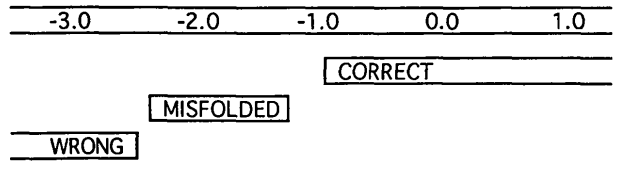


Fig. 4. 'Rule of thumb scale' for the interpretation of quality indices.

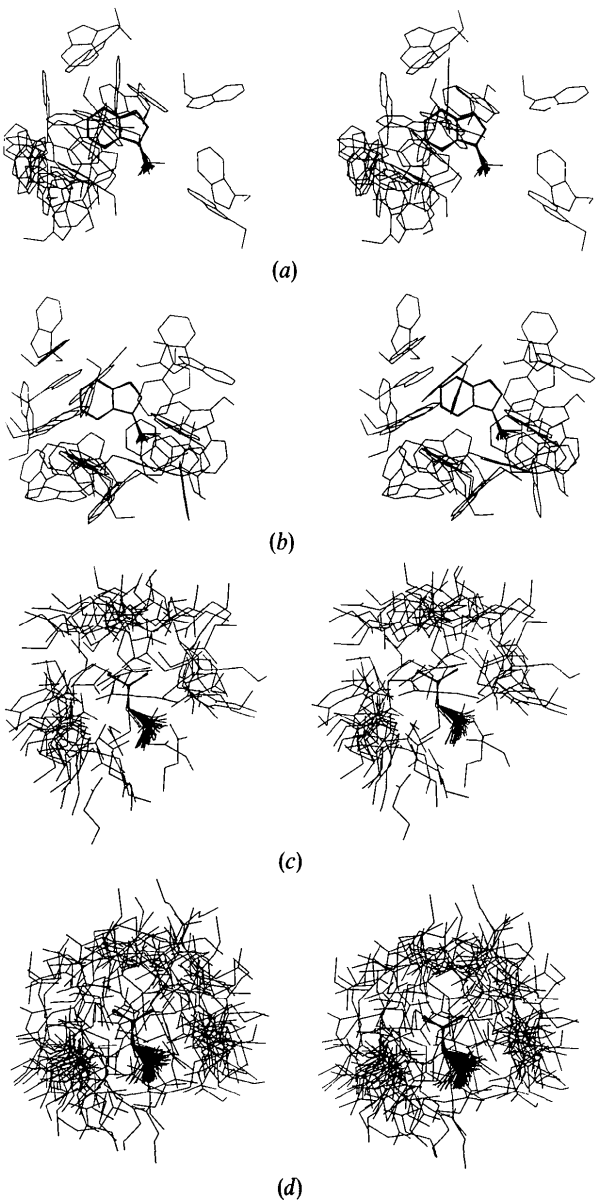


Fig. 5. (*a*) Distribution of contacting tryptophan side chains around a central helical tryptophan side chain. (*b*) As (*a*) for a non-helical central tryptophan side chain. (*c*) Distribution of aspartic acid and glutamic acid side chains around a cental C^{δ}, N^{ε}, C^{ζ}, $N^{\eta 1}$, $N^{\eta 2}$ arginine side-chain fragment. The central arginine is in a helix. (*d*) As (*c*) for arginines that are non-helical.

tryptophan-sidechain–tryptophan-sidechain contacts where the central (master) tryptophans are helical and non-helical, respectively. The clusters are asymmetric and distinctly different in these two cases. Figs. 5(*c*) and (*d*) show the packing of aspartic acid side chains around the outer fragment of the arginine side chain. The asymmetry is not as pronounced as in the previous example. However, it is worth noting that when both of the two aspartic acid oxygen atoms contact the arginine fragment, they tend to make these contacts with the N^{ε} and $N^{\eta 1}$ atoms of arginine. A detailed analysis of the asymmetry of residue contacts can be found elsewhere (Singh & Thornton, 1990). The small differences between our observations and the observations by Singh & Thornton are due to a different choice of proteins used to build the contact event database, to a different definition of a contact event and to different distance cut-offs and related program parameters.

These examples, together with the many examples in the literature (Levitt & Perutz, 1988; Singh & Thornton, 1985, 1990; Burley & Petsko, 1985, 1986; Reid, Lindley & Thornton, 1985), show that there is a sufficiently strong asymmetry in residue–residue contacts to make the method described here a good candidate for the evaluation of protein structures.

Quality as a function of the crystallographic R factor

Consistency of a model structure with crystallographic data is the primary objective measure of the quality of a model. In a blind test, we have determined the correlation between the contact quality index, which is based on database analysis, and the *R* factor, which is based on experiment. We tested the correlation with three series of successively refined structures kindly made available to us by crystallographers: triose phosphate isomerase (TIM) (Wierenga *et al.*, 1990), colicin (Postma, Parker & Tsernoglou, 1989; Parker, Pattus, Tucker & Tsernoglou, 1989) and thermitase (Gros, Betzel, Dauter, Wilson & Hol, 1989; Fujinaga, Gros & van Gunsteren, 1989).

In the case of TIM, the correlation between the quality index and the *R* factor is almost linear (see Fig. 6*a*). The two steps where the rise in quality index is lower than average (from $R = 23.2$ to $R = 22.5$ and from $R = 21.5$ to $R = 20.5$) correspond either to a change of refinement program (*GROMOS* → *PROLSQ* and *PROLSQ* → *TNT*, respectively) or to considerable rebuilding and addition of water molecules at these stages (R. Wierenga, personal communication). The resulting changes in the *R* factor are not reflected in the quality index, *i.e.* external intervention in the refinement process changed the *R* factor in a discontinuous fashion while the quality of packing was not significantly altered. By the time the *R* factor had reached 19.0, the TIM refinement mainly consisted of improving the positions and *B* factors of water atoms. As the quality evaluation method does not yet take water molecules into account, this is not reflected in the quality index, but does decrease the *R* factor. The remarkable result is that the linear relationship is broken only when there are known discontinuities in the refinement process.

The series of refined colicin structures shows a similar behaviour (Fig. 6*b*). The dip at $R = 23$ is caused by major manual rebuilding after $R = 25$, involving the shift of a large helix by one residue (M. Parker, personal communication). While during automatic refinement local packing is gradually improved as the *R* factor is decreased, manual rebuilding, which is normally done to remove large discrepancies between model and electron density, may actually make local contacts worse. These are

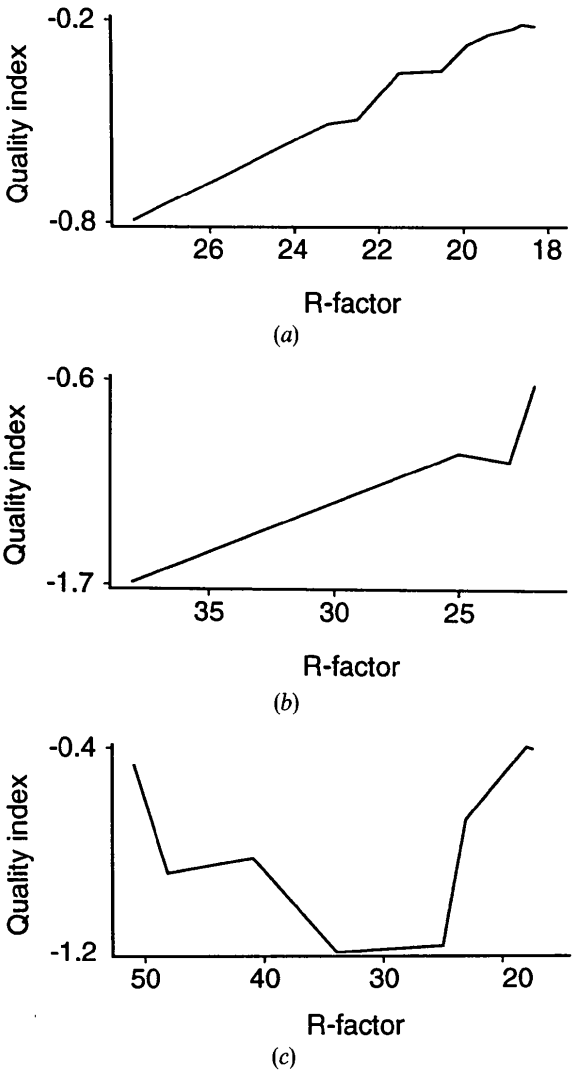


Fig. 6. Quality index as function of *R* factor for (*a*) TIM, (*b*) colicin, (*c*) thermitase–eglin complex.

later improved by the refinement program as the R factor is further improved.

A more complicated refinement course was taken for the thermitase–eglin complex (Fig. 6*c*). Because thermitase (TRM) is homologous (44% identical residues) to subtilisin Carlsberg (SCB), the latter structure was used as a starting model. The SCB structure was used to make a model of the TRM structure, initially without dealing with the difficult insertions and deletions. This model was extensively energy minimized and subsequently used as the starting point for the refinement of TRM. The starting model therefore partly had the contact quality of the known SCB structure and partly the contact quality of a well energy-minimized model, but had the (initially bad) R factor of the thermitase structure. The molecular-dynamics refinement procedure thereafter had to move atoms and residues from positions which were energetically favourable, but not in agreement with the X-ray data, to other energetically favourable positions, which were in agreement with the X-ray data. It is very unlikely that this path involves only perfectly packed intermediate states. Rather, as the packing is readjusted, the contact quality index initially decreases. In other words, the initial decrease and subsequent increase of the quality index reflects the transition from an initially well packed model to a significantly altered model *via* badly packed intermediates.

At $R = 41$ additional X-ray data were included (from 3.0 to 2.5 Å resolution) and several previously omitted residues were included in the model. The addition of new residues has been corrected for in Fig. 6(*c*) and therefore the increase in quality is mainly the result of the inclusion of higher-resolution data. It is likely that those short-range contacts that were already close to correctness at this stage of refinement became better because of the added high-resolution data.

Our interpretation agrees with what Fujinaga *et al.* (1989) state: 'The use of lower-resolution data allows greater conformational transitions to occur, whereas the inclusion of high-resolution data results in a more accurate structure'. In addition, we can state that it might have been more efficient to delay the inclusion of the higher-resolution data to a later stage of refinement because at an R value of around 30 the quality index is still decreasing, indicating that the refinement process is still making the structure look less like the SCB structure rather than fine tuning the final TRM coordinates.

It is interesting to speculate that the quality-index evaluation may be useful in the optimization of refinement strategies. For example, in molecular dynamics refinement, one might want to reduce the short-range interatomic force constants for some time after manual rebuilding until the quality index levels off. Whether or not it will be technically feasible to use the quality index as an additional term in refinement is an open question.

Discussion

The method has been shown to be a useful tool for the evaluation of protein structures. One should, however, keep in mind that there are circumstances where its use is not justified.

Crystal contacts have to be taken into account. The quality-index values for TIM were all determined for one TIM dimer, without its neighbours in the crystal being present. Of the eleven residues in TIM that had a relative quality index less than -4.0, four were involved in crystal packing.

An equivalent problem is seen in the case of multimeric proteins. An example is insulin: the average relative quality index for the A chain in the absence of the B chain is 0.491. The B chain alone has an index of -0.838. When these two chains are evaluated together, a much better relative quality index of 0.566 is obtained.

The conclusion to be drawn from these examples is that one should always carefully evaluate the entire packing environment.

There are several obvious improvements that could be made in the future. The first and most important problem to be solved is that of water contacts. There are two forms of water to be taken into account. Bulk water is adequately taken into account when the quality index for a residue is compared with the average quality index for this residue type as a function of the accessible contact surface area. The tightly bound water molecules that are seen in all protein structures are asymmetric in their distributions around residues (Thanki, Thornton & Goodfellow, 1988). However, there are too few entries in the Brookhaven protein database with reliable water positions for statistically reliable calculation of quality control boxes for water. A distinct asymmetry in the way calcium ions are bound to proteins, for example, has also been observed (Chakrabarti, 1989). However, here also the limited amount of data prevents statistical evaluation.

Another refinement which could be made is the inclusion of co-factors. For co-factors the available data is also insufficient. A possible improvement could be to treat all co-factors on the same basis as amino acids by putting every co-factor atom into one of the 57 atom classes. At present, however, co-factors are totally neglected.

Finally, the statistical evaluation of contacts, including solvent, could perhaps be formulated in terms of conditional probabilities so that absolute values become available without the need for an empirical normalization procedure.

Concluding remarks

The distribution of atomic contacts in the protein structure database has been analysed. The distribution of contacts has been summarized in terms of probability densities in the neighbourhood of residue fragment types. A method has been devised for the evaluation of protein structures. The method appears very useful in several ways.

(i) The contact distribution boxes can be used in analysing and understanding preferential atomic interactions in considerable detail.

(ii) The quality index can be used to assess the overall quality of crystallographic or theoretical models and may be used as a progress indicator in refinement procedures.

(iii) The contact probability distributions can be used as constructive tools for placing side chains in preferential orientations during model building, either by manual interaction with a graphics device or by an automatic optimization procedure.

(iv) By trying all 20 possible residues in all their allowed rotamers at a particular position in the protein interior, one can estimate how to improve thermostability by improving packing. Such a procedure has been applied successfully in several cases (see, for instance, Eijsink, Vriend, Van den Burg, Venema & Stulp, 1990).

The method described here is implemented as part of the molecular modeling and drug-design program *WHAT IF*. This program is available from one of the authors (GV) for a minimal fee. A list of quality indices for monomeric PDB entries is available. Upon request, we offer to provide a detailed quality-index report for experimental structures about to be deposited in the PDB. Send requests to VRIEND@EMBL-Heidelberg.DE.

Many people at EMBL have contributed to this project. Anna Tramontano's stimulation of this project deserves special recognition. Mike Parker, Piet Gros and Rik Wierenga kindly provided X-ray data. We are especially grateful to the many crystallographers who have made the coordinates of protein structures available to the scientific community by depositing them in the Protein Data Bank.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Branden, C. I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

Burley, S. K. & Petsko, G. A. (1985). *Science*, **229**, 23–28.

Burley, S. K. & Petsko, G. A. (1986) *FEBS Lett.* **203**, 139–143.

Chakrabarti, P. (1989). *Biochemistry*, **28**, 6081–6085.

Colonna-Cesari, F. & Sander, C. (1990). *Biophys. J.* **57**, 1103–1107.

Crippen, G. M. & Kuntz, I. D. (1978). *Int. J. Pept. Protein Res.* **12**, 47–56.

Eijsink, V. G. H., Vriend, G., Van den Burg, B., Venema, G. & Stulp, B. K. (1990). *Protein Eng.* **4**, 99–104.

Fujinaga, M., Gros, P. & van Gunsteren, W. F. (1989). *J. Appl. Cryst.* **22**, 1–8.

Gros, P., Betzel, Ch., Dauter, Z., Wilson, K. S. & Hol, W. G. J. (1989) *J. Mol. Biol.* **210**, 347–367.

Holm, L. & Sander, C. (1992) *J. Mol. Biol.* **225**, 93–105.

Levitt, M. & Perutz, M. F. (1988). *J. Mol. Biol.* **201**, 751–754.

Miyazawa, S. & Jernigan, R. L. (1985). *Macromolecules*, **18**, 534–552.

Parker, M. W., Pattus, F., Tucker, A. D. & Tsernoglou, D. (1989) *Nature (London)*, **337**, 93–96.

Postma, J. P. M., Parker, M. W. & Tsernoglou, D. (1989). *Acta Cryst.* A**45**, 471–477.

Reid, K. S. C., Lindley, P. F. & Thornton, J. M. (1985). *FEBS Lett.* **190**, 209–213.

Richmond, T. J. & Richards, F. M. (1978). *J. Mol. Biol.* **119**, 537–555.

Rosenfield, R. E., Swanson, S. M., Meyer, E. F., Carrell, H. L. & Murray-Rust, P. (1984). *J. Mol. Graph.* **2**, 43–46.

Singh, J. & Thornton, J. M. (1985). *FEBS Lett.* **191**, 1–6.

Singh, J. & Thornton, J. M. (1990). *J. Mol. Biol.* **211**, 595–615.

Tanaka, S. & Scheraga, H. A. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 3802–3806.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1988). *J. Mol. Biol.* **202**, 637–657.

Voorintholt, R., Kosters, M. T., Vegter, G., Vriend, G. & Hol, W. G. J. (1989). *J. Mol. Graph.* **7**, 243–245.

Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.

Warme, P. K. & Morgan, R. S. (1978). *J. Mol. Biol.* **118**, 273–287, 289–304.

Wierenga, R. K., Noble, M. E. M., Postma, J. P. M., Groenendijk, H., Kalk, K. H., Hol, W. G. J. & Opperdoes, F. R. (1990). *Proteins*, **10**, 33–49.