J. Mol. Biol. (1997) **268**, 678–685

JMB

Contact Area Difference (CAD): A Robust Measure to Evaluate Accuracy of Protein Models

Ruben A. Abagyan* and Maxim M. Totrov

The Skirball Institute of Biomolecular Medicine Biochemistry Department, NYU Medical Centre, 540 1st Avenue, New York, NY, 10016, USA

*Corresponding author

A simple unified measure to evaluate the accuracy of three-dimensional atomic protein models is proposed. This measure is a normalized sum of absolute differences of residue-residue contact surface areas calculated for a reference structure and a model. It employs more rigorous quantitative evaluation of a contact than previously used contact measures. We argue that the contact area difference (CAD) number is a robust single measure to evaluate protein structure predictions in a wide range of model accuracies, from *ab initio* and threading models to models by homology, since it reflects both backbone topology and side-chain packing, is smooth, continuous and threshold-free, is not sensitive to typical crystallographic errors and ambiguities, adequately penalizes domain and/or secondary structure rearrangements and protein plasticity, and has consistent linear and matrix representations for more detailed analysis. The CAD quality of crystallographic structures, NMR structures, models by homology, and unfolded and misfolded structures is evaluated. It is shown that the CAD number discriminates between models better than Cartesian root-mean-square deviation (cRMSD). Structural variability of the NMR structures was found to be three times larger than deformations of crystallographic structures in different packing environments.

Keywords: structure evaluation; modeling by homology; protein structure prediction; loop modeling; side-chain placement

Introduction

Fair and sensitive evaluation of three-dimensional (3D) protein models is essential for comparisons of different modeling and experimental techniques, and for the detection and promotion of the real improvements in structure prediction technology. Comparison of several structural models of the same polypeptide is somewhat special and is different from other types of structure comparisons between homologous or topologically related proteins. Several measures are commonly used to evaluate the difference between a reference structure and a model: (1) the root-mean-square-deviation of Cartesian coordinates of selected atoms (cRMSD), usually all heavy atoms, the backbone or C^{α} atoms; (2) the root-mean-square-deviation of selected inter-atomic distances (dRMSD); (3) the root-mean-square difference of selected torsion angles (aRMSD). The cRMSD is currently the most popular measure and it has been used for evaluation of protein conformations submitted to the comparative modeling section of the First Meeting on the Critical Assessment of Techniques for Structure Prediction (further referred to as CASP1; Moult *et al.*, 1995; Lemer *et al.*, 1995; Mosimann *et al.*, 1995).

All the above measures perform reasonably well only in the immediate vicinity of the correct conformation, however they do not allow a reasonable ranking of partially correct models, and are not based on any physical model of protein structure. The alternative class of measures is based on inter-residue contacts represented by a two-dimensional map (Phillips, 1970; Nishikawa & Ooi, 1974; Rossman & Liljas, 1974; Kuntz, 1975, etc.). An inter-residue contact was typically defined by the C^{α}–C^{α} or C^{β}–C^{β} distances with the distance cutoff ranging from about 7 to 12 Å (e.g. Braun, 1983). Godzik & Sander (1989) defined two grades of contacts by the shortest distance $D_{\min}$ between atoms of two residues: a strong contact with $D_{\min} < 3.75$ Å, a weak contact with $D_{\min} < 5$ Å. Analysis of inter-residue contacts was used pre-

Abbreviations used: 3D, three-dimensional; cRMSD, Cartesian root-mean-square deviation; CASP1, First Meeting on the Critical Assessment of Techniques for Structure Prediction; CAD, contact area difference; PDB, Protein Data Bank.

0022–2836/97/170678–08 $25.00/0/mb970994

viously to evaluate predictions of residue contacts (Goebel *et al*., 1994), conservation of side-chain to side-chain interactions between proteins with similar folds (Russell & Barton, 1994) and to evaluate the fraction of native contacts in simplified protein models (Skolnick & Kolinski, 1990; Shakhnovich *et al*., 1991; Guo & Thirumalai, 1995).

Evaluation of contact strength *via* interatomic distance is convenient and can be applied to protein models with different amino acid sequences. However in such a measure information about side-chain packing is essentially lost. To make the contact measure sensitive not only to the overall fold but also to fine details of side-chain arrangements, one needs a more accurate definition of the contact strength. Here we propose to characterize the strength of each residue-residue interaction by the contact surface area, add up all absolute differences between elements of two contact matrixes to generate a single measure of difference between them, and normalize the sum to 0 to 100% range. This measure further referred to as CAD has the following properties: (1) one number represents model accuracy from 0 (identical structures) to 100% (unrelated); (2) the measure is continuous and threshold-free; (3) it works in a wide range of model accuracies; (4) it adequately penalizes domain, fragment and side-chain rearrangements; (5) it captures essential geometrical characteristics which are related to protein stability; (6) it has consistent vector and matrix representations for detailed analysis; Here we also analyze the contact area differences between crystallographic structures in different crystal environments (reviewed recently by Kleywegt, 1996), between different possible models resulting from NMR structure determination, between models by homology and the correct structures.

The CAD number

Each two residues i and j in a protein structure can be characterized by certain contact area A_{ij} (Figure 1). To calculate this area one has to create a surface around isolated residue i by rolling a probe of radius R over the van der Waals surface of the residue atoms and tracing the center of the probe. The next step is to find which part of this surface is occluded by van der Waals surfaces of atoms of residue j. The resulting matrix $\{Aij\}$ is not exactly symmetric, i.e. A_{ij} is only approximately equal to A_{ji}. One may eliminate small differences between A_{ij} and A_{ji} by replacing them with their mean value: 0.5 $(A_{ij} + A_{ji})$. The diagonal values A_{ii} are set to zero.

The contact area matrixes can be calculated for a reference model (A^R) and a trial model (A^M). The non-zero elements of the difference matrix:

$$\Delta A_{ij} = (A^R_{ij} - A^M_{ij})$$

contain information about the wrongly predicted residue contacts. The local non-zero elements ΔA_{ij}

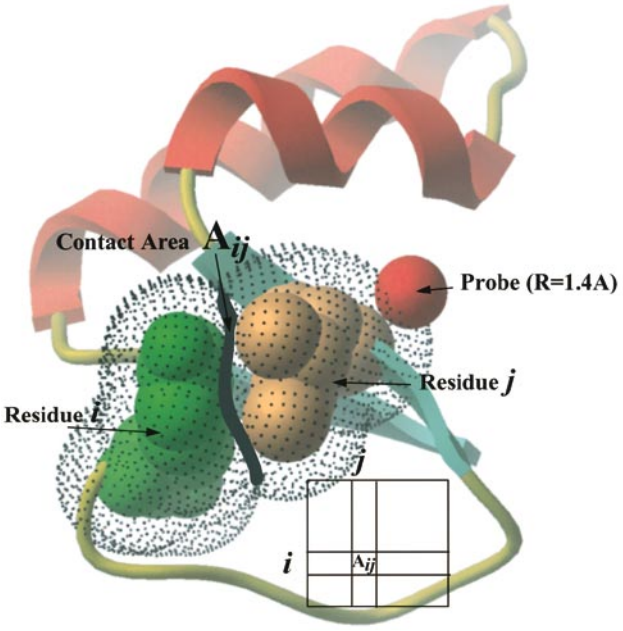

Figure 1. The surface area of contact between two residues i and j. Absolute differences between the A_{ij} values for two different models are accumulated in the CAD number.

$(j = i + 1,\ i + 2,\ i + 3,\ i + 4)$ contain information about secondary structure and local geometry, whereas elements with $i - j > 4$ correspond to differences in mutual arrangements of the structural elements.

The difference matrix $\{\Delta A_{ij}\}$ is the most detailed representation of the contact area error. It contains zeros for non-interacting residue pairs and correctly predicted contacts, positive values for missed or underestimated contacts and negative values for wrongly formed contacts. This matrix can be nicely represented in a graphical form with the color or shading dependent on the contact error.

The second representation is a profile of absolute contact errors for each residue:

$$\Delta A_i = \sum_j |(A^R_{ij} - A^M_{ij})|$$

The third and final representation is a single CAD number of the total unnormalized contact error:

$$\Delta A = \sum_{i,j} |(A^R_{ij} - A^M_{ij})|$$

There may be two extensions making the three CAD measures more useful. First, one can use residue weights W_i to reduce contribution from residues with high temperature factors B_i (in this work we used weights equal to 1, because for most of the structures B-factors were not available). Second, one can use a normalizing factor to make the CAD number independent of the protein size, shape and amino acid content. We propose a nor-

malizing factor which converts the CAD number into a relative structural difference measure in the range from 0 to 100%. We may evaluate the worst total difference ΔA_{worst} as proportional to the weighted average of elements of both matrixes. The normalized CAD number can be expressed as follows:

$$\text{CAD} = \frac{100 \sum_{i,j} W_i W_j |(A_{ij}^R - A_{ij}^M)|}{A_{\text{worst}}} \qquad (1)$$

where:

$$A_{\text{worst}} = C \sum_{i,j} W_i W_j \tfrac{1}{2}(A_{ij}^R + A_{ij}^M)$$

and:

$$W_i = \exp\left(-\frac{B_i}{B_{\text{std}}}\right)$$

W_i is the weight of residue i calculated *via* average residue temperature factor B_i and standard parameter B_{std}. Coefficient C would be equal to 2.0 if all the contacts could be rearranged in a random structure. It would mean that every correct contact area becomes zero. However, because of the covalent structure constraints the contacts between neighboring residues may be altered only to a certain extent. We found that $C = 0.9$ provides a reasonable normalization so that extended or scrambled random structures have close to 100% CAD difference from the reference structure (Figure 2(a)). A reasonable value of the B_{std} parameter is 20. Using larger values for B_{std} diminishes the difference between residues with high and low temperature factors.

The proposed structure similarity measure (CAD) is symmetric and can be used as the pairwise distance between n models for principal component analysis and clustering.

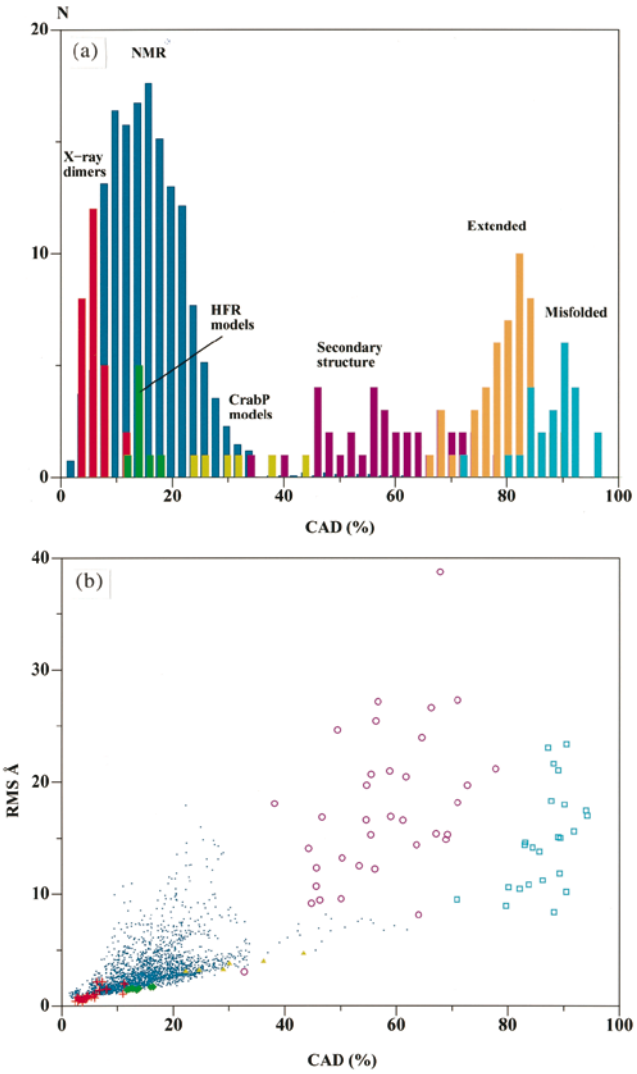

Figure 2. (a) A histogram of the CAD values and (b) the cRMSD *versus* CAD dependence for six sets of pairwise structure comparisons: PDB domains related by non-crystallographic symmetry and solved by X-ray crystallography (red), pairs of the NMR models submitted in one PDB entry (blue), models built by homology for the HPR protein (green), models by homology for CRABP (yellow), unfolded models with preserved secondary structure (violet), extended polypeptide chains (brown), deliberately misfolded models (dark green). The N scale for the NMR set have been divided by 20 to show the distribution on the same plot.

Application and calibration of the CAD number

In the following section we show how the CAD number was applied to various classes of structural deformations.

Accuracy of the structures solved by X-ray crystallography

Protein molecules have a certain degree of plasticity which can be revealed by comparison of protein structures in different crystal environments. Such comparisons can be made if the same protein is crystallized in several forms or if two molecules in the same crystal are related by a non-crystallographic symmetry. The average magnitude of structural variations due to crystal packing can be considered as a measure of ''accuracy'' of structure determination with respect to the protein molecule in solution, and largely constitutes the answer to a question: ''What is the effect of crystal packing on the structure?''.

We collected 27 pairs of structures related by non-crystallographical symmetry (see Materials and Methods) and calculated the CAD number and the backbone cRMSD for each pair (Figure 2(b)). The two measures correlate quite well. The average CAD number is about 5%.

Accuracy of the structures solved by NMR

A distribution of the pairwise CAD distances between models in a set of 121 NMR structures is shown in Figure 2(a). Typically, each NMR entry contains between 10 and 30 models. The distribution shows that the average CAD error is far beyond the accuracy of the X-ray structures. The distribution of the CAD number is characterized by the mean of 15.2%, standard deviation of 7.2%, and the CAD number reaches values as high as 60%, which demonstrates a wide variation of structural quality of the deposited NMR models. The CAD-cRMSD dependence (Figure 2(b)) shows structures of high inter-model cRMSD but relatively small CAD (16 Å and 22%, respectively, for the 1aaf structure). At the other extreme one can find proteins which definitely look unstructured from the CAD point of view (e.g. CAD of 58% and cRMSD of 6 Å on the average for the 1tiv PDB-entry).

Accuracy of models built by homology

We analyzed deviations for six models of cellular retinoic acid binding protein (CRABP) and eight models of histidine-containing phosphocarrier (HPR) built for the last CASP1 competition (Moult *et al*., 1995). The closest known templates used to build models of CRABP and HPR had 42% and 46% sequence identity, respectively. The resulting CAD accuracies were 22 to 43% for CRABP and 11 to 17% for HPR. Interestingly, the simple-minded models gave 36% for CRABP and 12% for HPR. Four CRABP models were better than the simple-minded model and one model was worse, while for the HPR protein the pattern was reversed: three models were better than the simple-minded model and four models were worse. The HPR models were based on an alignment to a known 3D template without any insertions and deletions and the deviations were mainly due to the quality of the side-chain prediction. It explains why all the CAD values are quite low (between 15 and 20%). However, it is interesting that a simple-minded model which inherits side-chain conformation for all residues with identical counterparts in the alignment and sets the most frequent rotamer to the ''non-identical'' residues has a CAD number of 16%, better than most of the models. The CRABP models were more difficult to build since they contained insertions and deletions, and the simple sequence alignment mistakenly places an insertion in the middle of the secondary structure element. The fragment misplaced in the initial alignment increases the CAD number by more than 10%: the misaligned models resulted in the CAD numbers of 36 and 43%, while other models ranged from 22 to 28%. The CAD number demonstrates a much better discrimination between models than RMSD.

Accuracy of the unfolded and misfolded models

Protein models which were deliberately misfolded or unfolded with preservation of the secondary structure are shown in Figure 2(a) and (b). The deliberately misfolded proteins were generated by Chris Sander and Lisa Holm by swapping side-chains between two proteins of similar length and optimizing them with a Monte Carlo algorithm (Holm & Sander, 1992). The chosen normalization (equation (1)) places them near the 100% wrong mark. In principle, it is possible to get a CAD number even higher than 100%, since the normalization takes into account only a limited set of the CAD numbers for unfolded/misfolded conformations. Models with totally extended backbones and random side-chain conformations get similar CAD scores (near 100%).

Discrimination ability of CAD and cRMSD

The previous four sections describe comparison sets which we will refer to as X-ray dimers, NMR models, HPR models, and preserved secondary structure and misfolded conformations. Intuitively, we know that the structural quality between classes is different and a good comparison measure should be able to separate them. Figure 2(a) and (b) show that the CAD measure separates most of the sets while the cRMSD does not. To compare the discrimination ability of CAD and cRMSD with respect to the various ''structural errors'', we evaluated overlap number ω (in per cent) between sets X_i ($I = 1, N_X$) and Y_j ($j = 1, N_Y$) as:

$$\omega = 100\left(1. - \left|\frac{1}{N_X N_Y}\left(\sum_{i=1,N_X}\sum_{j=1,N_Y}\delta(X_i, Y_j)\right)\right|\right) \quad (2)$$

where $\delta(X_i, Y_j)$ equals 1 if X_i is greater than Y_j and zero if X_i is equal to Y_j and -1 otherwise. This form of equation (2) relates the overlap to a probability that an element of the second set is greater than an element of the first set. Table 1 shows the overlap between four pairs of sets. In all cases the CAD measure allows us to discriminate between sets better than the cRMSD and appears to correspond to the intuitive evaluation of the amount of correct structural content. For example, the set with preserved secondary structure is clearly separated from the set of totally misfolded structures if evaluated with the CAD measure (overlap is only 1.1%), but not if evaluated with cRMSD (overlap 76%). More importantly, the CAD calculation clearly reveals the low quality of models built by homology as compared with the accuracy of crystallographic structure determination (overlap with the crystallographic set is 0%), while the RMSD calculation puts these models in the middle of the set of crystallographic dimers (overlaps of 40 and 25% for the backbone and all atom cRMSD,

Table 1. Discrimination abilities of the CAD number and cRMSD for four pairs of comparison sets

First set	Second set	Overlap of CADs (%)	Overlap of cRMSDs (%)
X-ray dimers	HPR models	0	40
X-ray dimers	NMR models	12.5	16
NMR models	Preserved secondary structure	0.5	3.2
Preserved secondary structure	Misfolded conformations	1.1	76

The overlap number was calculated according to equation (2).

respectively) suggesting overoptimism in the evaluation of the comparative models.

Dependence of CAD on side-chain rearrangements

The contact measures based on C^{α}–C^{α} or C^{β}–C^{β} distances are not sensitive to side-chain rearrangements and therefore cannot be used to evaluate models by homology. We generated a stack (Abagyan & Argos, 1992) of ten low-energy side-chain conformations by using an energy-minimized model of interleukin-1β as a starting structure and performing the biased probability sampling (Abagyan & Totrov, 1994) of the side-chain torsion angles. The polypeptide backbone and therefore all the C^{α}–C^{α} and C^{β}–C^{β} distances were kept unchanged. The CAD numbers for pairs of generated conformations were evaluated. The average CAD measure between these pairs of conformations was 12% and with the standard deviation of 2%. These CAD values are much larger than the average structural changes due to different crystallographic packing (see above).

Deviation measure to rank docking solutions

To rank different conformations of a ligand of N atoms ($i = 1,\ N$) docked to the receptor with respect to the known correct solution ($i' = 1,\ N$) one may use an RDE (relative displacement error) measure which is related to the CAD measure, but is much easier to calculate:

$$RelativeDisplacementError =$$

$$100\left(1 - \frac{L}{N}\left(\sum_{i=1,N}\frac{1}{L + D_{ii'}}\right)\right),$$

where L is the scale parameter, N is the number of ligand atoms and $D_{ii'}$ is the deviation of the model atom i from the corresponding atom i' in the reference structure. The scale parameter defines the accuracy scale. Values of L between 1.5 and 3 Å are reasonable, since at these distances specific interactions of ligand atoms with the receptor atoms are significantly reduced and possibly replaced by different interactions. The above formula has the following properties: if all the deviations are 0, RDE is 0%, if deviations are equal to L, RDE is about 50%, the same result may be achieved if half of the ligand atoms are predicted correctly (or deviate by much less than L), while the other half deviate by much more than L.

Discussion

The main motivation of this work was the inability of the commonly used RMSD measures to rank partially correct models. Ideally such ranking should not depend on rearrangements in parts of the structure which are obviously wrong or structurally unimportant (i.e. long exposed side-chains). It is this dependence that plagues the cRMSD, dRMSD and aRMSD measures in which contributions from the incorrect parts dominate, and are additionally emphasized by squaring each deviation, which then tend to oscillate strongly rather than converge. In simple words, we usually want to know how good is the structure, rather than measure how bad are the bad parts.

The contact or distance plots (Phillips, 1970; Nishikawa & Ooi, 1974; Rossmann & Liljas, 1974; Kuntz, 1975; Braun, 1983; Godzik & Sander, 1989; Skolnick & Kolinski, 1990; Shakhnovich *et al*., 1991; Guo & Thirumalai, 1995; and others) of two structures can be used to design a comparison measure which estimates the number of common contacts. However, previous definitions ignored the details of side-chain packing. The proposed measure is a normalized sum of absolute differences of the residue-residue contact surface areas calculated for two protein models. Quantitative and accurate evaluation of the residue-residue contact distinguishes the CAD measure from the distance-based contact difference maps and makes the measure sensitive to side-chain rearrangements. The key advantage of the CAD measure over the coordinate RMSD is its ability to adequately rank partially correct models because the CAD number is not sensitive to changes in the wrong or unimportant parts of the model. Table 2 summarizes differences between the cRMSD and the CAD number. Obviously, the CAD is a superior quantitative measure for comparison of different conformations of the same polypeptide chain in a wide range of prediction accuracies. However, if two polypeptide chains are different, as in two homologous proteins with different amino acid sequences, the CAD measure in the current form is not applicable, whereas the main-chain cRMSD as well as the C^{α}–C^{α} distance map still is.

Table 2. Comparison of the CAD and cRMSD measures of difference between a reference structure and a model

	Type of conformational change	CAD number	cRMSD
A	Insignificant rearrangements in long exposed side-chains (Lys, Met, Arg, etc.)	Mostly insensitive if they do not interact with other residues	Highly sensitive, e.g. in lysozyme a simple randomization of only side-chain torsions of the surface lysine and arginine residues (13% or the residues) leads to the all-atom cRMSD of 1.5 Å
B	Rearrangements of two correctly predicted parts (domain rearrangements)	Adequately reflects the change; only the domain interface residues contribute to the number. Roughly speaking CAD will be the fraction of residues with changed contacts	Can give any meaningless number, i.e. two immunoglobulin molecules (3bjl.a and 3bjl.b) superimpose with cRMSD (C^{α}) = 14.7 Å (CAD for the same pair is only 10%)
C	Unfolding of a protein with the secondary structure preserved	Shows loss of only about 40 to 60% of information, since correct secondary structure does contribute to the CAD number	A meaningless large value; no obvious correlation with the correctness of the secondary structure; wrongly folded model gives better cRMSD than the unfolded model with correct secondary structure
D	Partially correct structure (i.e. loops and/or termini are incorrect)	The contributions to the CAD difference grow only until the correct contacts are completely broken, after which the measure does not depend on the conformation of the incorrect fragment	Contributions from the incorrect parts dominate and fluctuate strongly; model with larger correct part may easily get larger cRMSD than the model with smaller correct part
E	Rearrangements in the residues involved in crystal contacts	Contributions much smaller than the buried residues	Contributions similar to those from the buried residues
F	Crystallographically indistinguishable side-chain rotamers	Identical CAD number	Needs special treatment
G	Unreliable side-chain rotamers (e.g. χ^1 of Val)	Minimal influence	Needs special treatment

Some types were taken from John Moult's evaluation criteria (Moult, 1996).

Why do we want to take a sum of absolute contact area deviations rather that a root-mean-square measure:

$$\mathrm{RMS\_CAD} = \sqrt{\frac{1}{N_{\mathrm{res}}^2}\sum_{i,j}(A_{ij}^R - A_{ij}^M)^2} \qquad (3)$$

where N_{res} is the number of residues?

There are two reasons. First, an elementary residue-residue contribution to stabilization energy is rather a linear function of the contact area than a quadratic function, therefore a linear measure will better represent the ''correctness'' of the residue-residue interactions. Second, a normalization to N_{res}^2 of residue pairs does not make much sense, given the fact that most of them cannot be formed.

Since the CAD contributions $W_iW_j|(A_{ij}^R - A_{ij}^M)|$ are not easily comparable and their average value has no clear meaning, we proposed a normalization which would approach 100% for totally incorrect models in which only the covalent structure is preserved. Simplistically, this measure gives the percentage of the incorrect structure.

To investigate properties of the CAD measure and develop a feeling of what these numbers mean, we applied it to several classes of models. The first class consists of protein domains related by non-crystallographic symmetry. Comparison within this subset allows us to estimate the ''CAD accuracy'' of the X-ray structure determination and conformational differences originating from different crystal packing. We demonstrate that the CAD difference is only about 4% and does not exceed 10%. Interestingly, while most of the backbone cRMSDs are less than 1 Å, there are several pairs of domains related by non-crystallographic symmetry which are characterized by high backbone cRMSD up to 2.2 Å (e.g. 1set.a and 1set.b, or 1sas.a and 1sas.b), while the CAD differences for the same pairs are only about 5%. Visual inspection shows that these structures contain long helical elements bent slightly differently in two subunits. This difference leads to the cRMSD which is larger than for some models by homology compared with their reference structures. These deformations, however, do not confuse the CAD measure which clearly discriminates crystallographic deformations and much larger distortions in models by homology (Figure 2(b), Table 2).

The second set includes NMR structures for which we evaluated structural differences between models deposited in the Protein Data Bank (Bernstein *et al.*, 1977) in the same entry. The mutual inconsistency of the models evaluated by the CAD measure spans the range up to 60% with the mean value of about 15% which is three times higher than a similar distribution of differences in a crystallographic set. This threefold increase may partly be attributed to solution dynamics, but appears to mainly reflect the accuracy of the experimental data and procedures used for the NMR structure determination. Inspection of the low ac-

curacy entries such as 1tiv (CAD errors from 46 to 62%) confirms that disorder of this scale can hardly be attributed to solution dynamics. The calculated distribution of the CAD values (Figure 2(b)) may suggest a quality threshold for submission of the NMR structures to the PDB.

The third set of comparisons includes two groups of models built by homology for the CASP1 (Moult *et al.*, 1995; Lemer *et al.*, 1995; Mosimann *et al.*, 1995). The most important conclusion from our results is that the models have clearly lower structural quality than the typical crystallographic quality even though the cRMSD values from the reference structure may be quite low. The HPR models are the most characteristic in this regard, since they are based on the alignment without any insertions or deletions and accurate backbone prediction was not a problem. Models and X-ray dimers are totally separated in the CAD scale but not in the cRMSD scale.

The sets of unfolded structures with and without the secondary structure and the set of deliberately misfolded structures show clear separation when evaluated with the CAD measure, while their cRMSD values fall within practically the same range. This separation of the CAD values properly reflects the presence of the correct secondary structure in the fourth set which is absent in the fifth set. On the other hand, cRMSD is obviously not capable of detecting this feature and evaluates the models with the correct secondary structure as being as bad as the completely wrong models. A number of advantages of the CAD number over cRMSD are listed in Table 2.

Based on the presented results we would propose to use the CAD number as a standard measure to evaluate the difference between a model and a reference structure and to rank the models. Alternatively, if the reference structure is not known, one can calculate the average pairwise CAD distance in a set of several models as a measure of shared structural information.

The program

The ICM_CAD program with its source code is available from the Web site (http://saturn.med.nyu.edu/beta/groups/Abagyan.html). It uses modules and algorithms of the ICM program (Abagyan *et al.*, 1994; Molsoft, 1996).

Materials and Methods

CAD calculations

The surface area A_{ij} between two residues was calculated by the Shrake and Rupley algorithm (Shrake & Rupley, 1973) modified to speed up the calculation (Abagyan *et al.*, 1994). The probe radius was set to 1.4 Å and the following van der Waals radii were used: C, 1.9 Å; S, 1.9 Å; N, 1.65 Å; O, 1.6 Å. Hydrogen atoms were ignored. The standard correction of 3% was applied to compensate for errors of surface calculations due to the relatively low number of points representing a van der Waals sphere. Two contact areas A_{ij} and A_{ji} were replaced by their average value.

A set of domain pairs solved by X-ray crystallography

To analyze the crystallographic accuracy we selected a set of PDB domain pairs related by non-crystallographic symmetry and solved at resolution better than 2.5 Å (the PDB code and the identifiers of two compared chains compared are given): 1apx(a,b), 1bre(a,b), 1buc(a,b), 1deh(a,b), 1dpg(a,b), 1ebg(a,b), 1ebh(a,b), 1gse(a,b), 1ids(a,b), 1les(a,b), 1ndp(a,b), 1pvd(a,b), 1pyd(a,b), 1sec(a,b), 1set(a,b), 1smn(a,b), 1tar(a,b), 1wgc(a,b), 2cst(a,b), 2nac(a,b), 2phi(a,b), 2wgc(a,b), 4mdh(a,b), 5p2p(a,b), 7wga(a,b), 8cat(a,b), 9wga(a,b).

A set of NMR models

The following NMR entries were selected for the comparison: 1hph, 1eci, 1erp, 1bnb, 1agt, 1psm, 1dec, 1mtx, 1erd, 1ica, 1apo, 1hev, 1ret, 1atx, 1gps, 1cfh, 1shi, 1ahl, 1tfi, 2ech, 4tgf, 1bal, 1afp, 1mbf, 3egf, 1tih, 1mbk, 1shp, 1aaf, 1dtk, 1bbo, 1drs, 2ccx, 1pce, 1tfs, 2cdx, 1ntx, 1den, 1tcp, 1cld, 1nea, 1nor, 1san, 2igg, 1prl, 4hir, 1cis, 3ci2, 1ocp, 1hrf, 1pog, 1igl, 1hom, 1kst, 1stu, 2hoa, 1rtn, 1pse, 1hun, 1ftz, 1bha, 2bbi, 2hsp, 1c5a, 1pih, 1hma, 1mgs, 1ner, 1bod, 1abt, 1gbr, 1ghc, 1tvt, 1adr, 1boc, 1tnt, 1hrz, 1mnt, 1cdr, 1cb1, 2ptl, 1aty, 1dvh, 1hra, 1pba, 1rip, 1ctl, 1grx, 1ego, 1tiv, 1aca, 1lpt, 1hue, 1and, 3hsf, 1ttf, 1aps, 1fli, 9pcy, 2sob, 2ple, 1mdj, 1pcp, 1put, 1hkt, 1fhb, 1ab2, 1exh, 1akp, 1svq, 1bip, 1cye, 1urk. The comparisons were performed between the first model (usually the best structure or the average one) and other deposited models.

Models by homology

Models by homology for CRABP (PDB code 1cbi) and HPR (1pch) submitted to the CASP1 (Moult *et al.*, 1995) were taken from John Moult's ftp-site (ftp://iris4.cad.nist.gov/pub/model_database). The so-called simple-minded models for CRABP and HPR were built from the 2hmb and 2hpr structures, respectively, with the ICM program (Molsoft, 1996) using a procedure including the following steps: (1) alignment of sequences of the query and the template proteins using automatic global sequence alignment procedure (Needleman & Wunch, 1970); (2) transferring the aligned backbone and side-chains identical in the alignment directly from the template to the model; (3) assignment of minimally deformed extended conformation to the inserted backbone and the most likely rotamer for the non-identical side-chains (without sampling).

Unfolded and misfolded models

Models without tertiary structure were generated by assigning random values to the phi and psi torsion angle of glycine residues for a set of 36 proteins (1aaj, 1aba, 1acf, 1acx, 1cdp, 1cot, 1crn, 1ctf, 1eca, 1fas, 1fdx, 1fkb, 1hfh, 1hoe, 1hpi, 1hvc, 1hyp, 1knt, 1lis, 1mjc, 1pmy, 1ppa, 1ppt, 1ptf, 1rds, 1ten, 1yea, 2bop, 2cdv, 2che, 2hmb, 2imm, 2mhr, 3b5c, 3fxn, 7pcy). The deliberately misfolded models (1bp2on2paz, 1cbhon1ppt, 1fdxon5rxn, 1hipon2b5c, 1lh1on2i1b, 1p2pon1rn3, 1ppton1cbh, 1reion5pad, 1rhdon2cyp, 1rn3on1p2p, 1sn3on2ci2, 1sn3on2cro, 2b5con1hip, 2cdvon2ssi,

2ci2on1sn3, 2ci2on2cro, 2coon1sn3, 2crooon2ci2, 2cypon1rhd, 2i1bon1lh1, 2pazon1bp2, 2ssion2cdv, 2tmnon2ts1, 2ts1on2tmn, 5padon1rei, 5rxnon1fdx) were generated and made publicly available by Chris Sander and Lisa Holm (Holm & Sander, 1992) (Web site http://www.embl-heidelberg.de/ ~ holm).

Acknowledgements

This work was inspired by the Cambridge meeting on Protein Structure Prediction (May, 1996) organized by Gerrit Vriend and Chris Sander. The authors are grateful to John Moult, Tim Hubbard, Scott Dixon and Steve Bryant for stimulating discussions, and to Vladimir Maiorov and Jonathan Weider for careful reading of the manuscript.

References

Abagyan, R. A. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225**, 519–532.

Abagyan, R. A. & Totrov, M. M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.

Abagyan, R. A., Totrov, M. M. & Kuznetsov, D. A. (1994). ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1122**, 535–542.

Braun, W. (1983). Representation of short and long-range handedness in protein structures by signed distance maps. *J. Mol. Biol.* **163**, 613–621.

Godzik, A. & Sander, C. (1989). Conservation of residue interactions in a family of Ca binding proteins. *Protein Eng.* **2**, 589–596.

Goebel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.

Guo, Z. & Thirumalai, D. (1995). Kinetics of protein folding: nucleation mechanism, time scales and pathways. *Biopolymers,* **36**, 83–102.

Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 92–105.

Kleywegt, G. J. (1996). Use of Non-crystallographic Symmetry in Protein Structure Refinement. *Acta Crystallog.* **52**, 842–857.

Kuntz, I. D. (1975). An approach to the tertiary structure of globular proteins. *J. Am. Chem. Soc.* **97**, 4362–4366.

Lemer, C. M.-R., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337–355.

Molsoft (1996). *ICM 2.5 manual*.

Mosimann, S., Meleshko, R. & James, M. N. G. (1995). A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins: Struct. Funct. Genet.* **23**, 301–317.

Moult, J. (1996). Comparative modeling evaluation criteria. *http://www.mrc-cpe.cam.ac.uk/casp2/cm-criteria.html*.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Struct. Funct. Genet.* **23**, 2–4.

Needleman, S. B. & Wunch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Nishikawa, K. & Ooi, T. (1974). Comparison of homologous tertiary structures of proteins. *J. Theoret. Biol.* **43**, 351–374.

Phillips, D. C. (1970). The development of crystallographic enzymology. *Biochem. Soc Symp.* **31**, 11–28.

Rossmann, M. G. & Liljas, A. (1974). Recognition of structural domains in globular proteins. *J. Mol. Biol.* **85**, 177–181.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chaincontacts secondary structure and accessibility. *J. Mol. Biol.* **244**, 332–350.

Shakhnovich, E. I., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice model simulation. *Phys. Rev. Letters,* **67**, 1665–1669.

Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and Insulin. *J. Mol. Biol.* **79**, 351–371.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science,* **250**, 1121–1125.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267–1289.

Edited by B. Honig

(Received 12 July 1996; received in revised form 17 February 1997; accepted 25 February 1997)