J. Mol. Biol. (1998) **275**, 895–916

JMB

An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction

Ram Samudrala[1,2] and John Moult[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600, Gudelsky Drive, Rockville, MD 20850 USA*

[2]*Molecular and Cell Biology Program, University of Maryland at College Park College Park, MD 20742, USA*

**Corresponding author*

We present a formalism to compute the probability of an amino acid sequence conformation being native-like, given a set of pairwise atom–atom distances. The formalism is used to derive three discriminatory functions with different types of representations for the atom–atom contacts observed in a database of protein structures. These functions include two virtual atom representations and one all-heavy atom representation. When applied to six different decoy sets containing a range of correct and incorrect conformations of amino acid sequences, the all-atom distance-dependent discriminatory function is able to identify correct from incorrect more often than the discriminatory functions using approximate representations. We illustrate the importance of using a detailed atomic description for obtaining the most accurate discrimination, and the necessity for testing discriminatory functions against a wide variety of decoys. The discriminatory function is also shown to be capable of capturing the fine details of atom–atom preferences. These results suggest that the all-atom distance-dependent discriminatory function will be useful for protein structure prediction and model refinement.

Keywords: knowledge-based; conditional probability; discriminatory function; decoy sets

Introduction

Any algorithm that attempts to predict protein structure requires a discriminatory function that can distinguish between correct and incorrect conformations. These discriminatory functions can be extremely simple: for example, counting atomic contacts in a given conformation, or can involve elaborate calculations based on the physics of the system to determine the energy of a conformation (Brooks *et al*., 1983; Weiner *et al*., 1986; Jorgensen & Tirado-Rives, 1988).

A class of discriminatory functions is knowledge-based. These functions compile parameters from tendencies observed in a database of experimentally determined protein structures (Wodak & Rooman, 1993; Sippl, 1995). Historically, knowledge-based discriminatory functions have gained popularity in their application to the ''fold recognition'' problem, i.e. recognising the fold an amino acid belongs to in the absence of detectable sequence homology (Sippl, 1990; Bowie *et al*., 1991; Jones *et al*., 1992; Bryant & Lawrence, 1993). Since then, knowledge-based discriminatory functions have been used to validate experimentally determined protein structures (Lüthy *et al*., 1992; Sippl, 1993; MacArthur *et al*., 1994), for *ab initio* protein structure prediction (Sun, 1993; Simons *et al*., 1997), and have proven their worth in *bona fide* fold recognition experiments (Lemer *et al*., 1995; Madej *et al*., 1995; Flöckner *et al*., 1995; Jones *et al*., 1995; Levitt, 1997).

Abbreviations used: CASP, critical assessment of protein structure prediction methods; CDF, contact discriminatory function; crabpi, cellular retinoic acid binding protein I; edn, eosinophil derived neurotoxin; e5.2, immunoglobulin domain protein; GA, genetic algorithm; hpr, histidine-containing phosphocarrier protein; IFU, independent folding unit; IRAPDF, linearly interpolated residue-specific all-atom conditional probability discriminatory function; NMR, nuclear magnetic resonance; mm23, nucleoside diphosphate kinase protein; NVPDF, non-residue-specific virtual-atom conditional probability discriminatory function; PDB, Protein Data Bank; PDF, probability discriminatory function; p450, heme protein; RAPDF, residue-specific all-atom conditional probability discriminatory function; RVPDF, residue-specific virtual-atom conditional probability discriminatory function; rmsd, root mean square deviation.

0022–2836/98/050895–22 $25.00/0/mb971479

Generally, knowledge-based discriminatory functions have used a simple one- or two-point-per-residue representation. That is, they usually represent each residue in a protein sequence with one or two positions in three-dimensional space. Discrimination is based on each residue's preference to be buried or exposed, its preference for a particular secondary structure conformation, and/or its preference to be in contact at a particular distance and sequence separation from other residues (Sippl, 1990; Bowie *et al*., 1991; Jones *et al*., 1992; Bryant & Lawrence, 1993). However, to capture the finer details of atom–atom interactions in proteins, a more detailed representation is necessary, and two such functions (DeBolt & Skolnick, 1996; Subramaniam *et al*., 1996) have been developed so far. For example, in a comparative modeling scenario where two possible models can be quite similar (within 1.0 to 3.0 Å in terms of root mean square deviation (rmsd) of the C_α atoms) to the experimentally determined structures (Mosimann *et al*., 1995), we need all the information we can possibly obtain from the two models to determine which one is more accurate. A one-point-per-residue discriminatory function may not be able to discriminate as well as an all-atom discriminatory function, which takes into account the environment of all the atoms on the main-chain and the side-chain of each residue. Also, a detailed all-atom model cannot be built using a simple representation.

A major issue in developing any discriminatory function for work with protein molecules is how to test performance. There are three principal strategies. Most popular for testing physics-based functions has been detailed comparison with experimental data from small molecule systems (Halgren, 1995). The assumption is that good results on such data must imply adequate performance on the large molecule systems. The second method is the use of ''decoy'' sets (Park & Levitt, 1992). That is, devising many incorrect structures, and testing whether a function can discriminate between these and the experimental conformation. Decoys have been based on lattice models (Park & Levitt, 1992), molecular dynamics trajectories (Wang *et al*., 1995), crystal structures of different resolutions (Subramaniam *et al*., 1996) and amino acid sequences mounted on radically different folds (Holm & Sander, 1992a). World Wide Web sites have been established to provide decoy test sets for fold recognition functions (Fischer & Eisenberg, 1996; Fischer, 1997) and for general protein structure prediction functions (Braxenthaler *et al*., 1997). To date, each function has generally only been tested on one or two classes of decoy. A danger here is that discrimination may be achieved utilizing some specific aritfacts of the decoys. For example, non-compactness or systematic distortion of detailed features such as abnormal hydrogen bond length. The third approach is to use the function to drive a search for a native like conformation, starting from some approximate structure. Tests of this sort have so far only been reported for physics-based potentials, and very rarely have they been even partially successful (Storch & Daggett, 1995). For protein structure prediction to work, this is the most relevant test, but it is also the most difficult and time consuming. We have opted for testing against decoys, and have used as wide a range of types as possible, taking advantage of the test sets available in PROSTAR, the Protein Potential Test Site (Braxenthaler *et al*., 1997).

Our goal is to develop a discriminatory function that will work well at identifying the best conformation among a set of incorrect or approximate conformations. To accomplish this, we derive pairwise distance-dependent all-atom conditional probability functions that represent atom–atom preferences in a residue specific manner. We evaluate the performance of these functions by seeing how well they distinguish correct conformations of an amino acid sequence from incorrect or approximate (decoy) conformations. We perform this evaluation for a wide variety of decoy types. We compare this discriminatory function to three more approximate representations to determine the effect of decreasing detail in the representation. Two of the approximate representations treat combinations of atoms as single ''virtual atoms''. The third approximate representation, a simple contact-based discriminatory function, is used to illustrate how much of the discriminatory information is obtained from compactness alone (Bahar & Jernigan, 1997). We discuss the implications of the these results for protein structure prediction and model refinement.

Methods

We will describe two formalisms. The first computes the conditional probabilities, and the second computes the free energies, of pairwise atom–atom preferences in proteins using statistical observations on native structures. We make the observation that these two formalisms are equivalent for all practical purposes. However, it is more straightforward to think of pairwise preferences of atoms in proteins in terms of probabilities rather than in terms of free energies: the Boltzmann formalism assumes an equilibrium distribution of atom–atom preferences, the physical nature of the reference state is not clear, and the probability of observing a system in a given state must change with respect to the temperature (Moult, 1997).

The conditional probability formalism

We divide the possible conformations of a structure or piece of structure into two subsets, the set of conformations we will regard as correct (i.e. native conformations, having low rms deviation from the experimental structure), {C}; and the rest, the set of incorrect structures {I}. We consider a set of properties of the structure $\{y_k\}$. The properties may be any aspects of protein structure that differ

significantly between the set of incorrect conformations we are dealing with and the conformations we regard as correct. Examples might be the strength of electrostatic interactions, close packing, exposure of non-polar groups to solvent, or, as in the present application, simply a set of interatomic distances within a structure, $\{d_{ab}^{ij}\}$, where d_{ab}^{ij} is the distance between atoms i and j, of type a and b, respectively. We wish to evaluate $P(C \mid \{d_{ab}^{ij}\})$, the probability the structure is a member of the ''correct'' set, given it contains the distances $\{d_{ab}^{ij}\}$. To perform this evaluation, we express $P(C \mid \{d_{ab}^{ij}\})$ as an explicit function of probabilities that can be derived from experimental structures. These are $P(d_{ab}^{ij} \mid C)$, the probability of observing a distance d between two atoms i and j of types a and b in a correct structure, and $P(d_{ab}^{ij})$, the probability of observing such a distance in any structure, correct or incorrect. Also required is $P(C)$, the probability that any structure picked at random is a member of the ''correct'' set. We use the general and exact chain rule for conditional probabilities (Mosteller *et al*., 1970):

$$P(C) \cdot P(d_{ab}^{ij}|C) = P(d_{ab}^{ij}) \cdot P(C|d_{ab}^{ij}) \tag{1}$$

and express the probabilities of observing the set of distances as products of the probabilities of observing each individual distance, making the important approximation that all distances are independent of one another:

$$P(\{d_{ab}^{ij}\}|C) = \prod_{ij} P(d_{ab}^{ij}|C); \qquad P(\{d_{ab}^{ij}\}) = \prod_{ij} P(d_{ab}^{ij}) \tag{2}$$

It then follows that:

$$P(C|\{d_{ab}^{ij}\}) = P(C) \cdot \prod_{ij} \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \tag{3}$$

$P(C)$ is a constant independent of conformation for a given amino acid sequence, and is not considered further. Note however that its omission means that scores from different sequences cannot be compared. It is useful to use a log form of equation (3), both to scale the quantities involved to a small range, and to give a form similar to that of a potential of mean force. We use a scoring function S proportional to the negative log conditional probability that the structure is correct, given a set of distances:

$$S(\{d_{ab}^{ij}\}) = -\sum_{ij} \ln \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \propto -\ln P(C|\{d_{ab}^{ij}\}) \tag{4}$$

Given an amino acid sequence conformation, we calculate all the distances between all pairs of atom types and compute the score S on the left-hand side by summing up the probability ratios assigned to each distance between a pair of atom types.

To make use of equation (4), we need distributions of $P(d_{ab} \mid C)$ and $P(d_{ab})$ for all combinations of atom types at all observed distances. The $P(d_{ab} \mid C)$ are derived directly from experimental structures as follows:

Using a set of known structures from the Brookhaven Protein Data Bank (PDB) (Bernstein *et al*., 1977), we can make observations of atom–atom contacts in particular distance bins. We compute the probability of observing atom type a and atom type b in a particular distance bin (with midpoint value d) in a native conformation C,$P(d_{ab} \mid C)$, like so:

$$P(d_{ab}|C) = f(d_{ab}) = \frac{N(d_{ab})}{\Sigma_d N(d_{ab})} \tag{5}$$

Here $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d. The denominator is the number of a–b contacts observed for all distance bins. We assume that the frequency distributions obtained from the database, $f(d_{ab})$, here and elsewhere, represent the probabilities.

For example, if the number of lysine N_ζ and glutamate $O_{\delta 1}$ (KN_ζ-$EO_{\delta 1}$) contacts within a distance range of 4.0 to 5.0 Å was found to be equal to 10 in the data set, and the total number of KN_ζ-$EO_{\delta 1}$ contacts observed in all distance bins was 100, the frequency of KN_ζ–$EO_{\delta 1}$ contacts at distance bin 4.5 is $10/100 = 0.1$.

$P(d_{ab})$ may be thought of as describing a property of the reference state, specifically, the probability of seeing a separation d between atoms types a and b in any possible structure, correct or incorrect. In terms of Bayesian statistics (Mosteller *et al*., 1970), $P(d_{ab})$ is a prior distribution, that is, our knowledge of the interatomic distances in proteins for a particular pair of atom types, before we have observed any experimental structures. An advantage of the Bayesian viewpoint compared with the requirements of the potential of mean force is that it more clearly allows a choice of any prior distribution convenient to our purpose. Many different prior distribution choices are possible. For example, we may assume that all distances are equally probable (Subramaniam *et al*., 1996), or that the set of distances observed in some random coil model is appropriate, a choice often made in potential of mean force work (Avbelj & Moult, 1995b) because of the relevance to the physical folding process. Such distributions are perfectly valid, but do not necessarily make the best use of the available information. For example, it has been shown that much of the apparent signal in a potential of mean force using a random coil reference state originates in the non-specific compactness of correct structures (Jernigan & Bahar, 1996; Bahar & Jernigan, 1997). To obtain the most information supporting or refuting the hypothesis that we are considering a native structure, we would like to use a distribution that includes all the possible incorrect structures that might be encountered, but one that is not unnecessarily broad. In a particular application where a large number of representative

conformations have been generated, for example, when examining all possible conformations of a loop in a protein molecule (Fidelis *et al*., 1994), it is possible and presumably advantageous to compile the reference distributions from the available correct and incorrect conformations. More typically, the sample of incorrect conformations available is too sparse to allow this direct approach. In the present work, we deal only with a prior distribution that is generally useful for structure prediction. Methods used to predict protein structure usually produce reasonably compact models, even when the result is incorrect. Thus, a good choice of prior distribution is that found in the set of possible compact conformations.

We assume that averaging over different atom types in experimental conformations is an adequate representation of the random arrangements of these atom types in any compact conformation. Then we can approximate $P(d_{ab})$, the probability of finding atom types a and b in a distance bin d in *any* compact conformation, native or otherwise, as equal to $P(d)$, the probability of seeing *any* two atom types in a distance bin d:

$$P(d_{ab}) = P(d) = f(d) = \frac{\Sigma_{ab} N(d_{ab})}{\Sigma_d \Sigma_{ab} N(d_{ab})} \qquad (6)$$

Here, $\Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types in a particular distance bin d, and the denominator is the total number of contacts between all pairs of atom types summed over the distance bins d.

There are three principal approximations in this formalism:

(1) All the interatomic distances are independent of one another. This is clearly not correct: if atom A is close to atom B and to atom C, then we know that atoms B and C must also be fairly close. For example, if the $C_{\delta 2}$ of leucine is close to the $C_{\delta 1}$ atom of another leucine, it is likely that the $C_{\delta 1}$ of the first residue will also be close to this atom. In this sense, there is some degree of double counting of interactions. Although it is possible to write the probabilities in a way that takes these correlations into account, much more data would be required to obtain adequate statistics, so that it is difficult to assess the impact of this approximation.

(2) The distribution of interatomic distances in proteins is not significantly different in different environments. This is not universally true. Two examples: the distribution of distances between ionic groups on the surface of proteins is very different from that of buried pairs (Wodak & Rooman, 1993), and the probability of a hydrogen bond being formed between main-chain groups varies sharply as a function of the sequence separation between them (Sippl *et al*., 1996). In principle, given a large enough set of experimental structures, it is possible to compile separate probability distributions for each such environmental variable, subject to availability of sufficient observations. We have not attempted to do that in the present study.

(3) The interatomic distance probability distributions derived from experimental structures will be sufficiently sharp, and different for different atom types, that a useful discriminatory signal will be obtained. We examine the extent to which this is true later in this work.

The potential of mean force

The potential of mean force method for discriminating between correct and incorrect protein structures rests on three assumptions:

(1) The total free energy of a protein molecule relative to some reference state, ΔG_{tot}, can be expressed as a sum of the relative free energy $\Delta G(R)$ of a number of individual contributions, where R represents the value of a ''reaction co-ordinate''. As with the conditional probability analysis, the reaction co-ordinate may be any convenient measure of property of the molecule (Beveridge & DiCapua, 1989), such as the separation of specific types of atoms (Bahar & Jernigan, 1997) or residues (Sippl, 1990; Jones *et al*., 1992), the distance between hydrogen bonding groups (Sippl *et al*., 1996), or the value of the local main chain electrostatic energy (Avbelj & Moult, 1995b). We again consider the particular case of a reaction co-ordinate representing the distance of d between atoms i and j of type a and b, respectively, so that:

$$\Delta G_{\text{tot}} = \sum_{ij} \Delta G(d_{ab}^{ij}) \qquad (7)$$

Although enthalpy may be reasonably approximated by such a pairwise sum of interactions, and typically is in empirical force field application (McCammon & Harvey, 1987), it is not generally valid to do this for free energy, because entropy contributions can not be regarded as additive (Mark & van Gunsteren, 1994). This issue is related to the assumption of independence of contributions in the conditional probability formalism discussed earlier, but the nature of the approximation is harder to express quantitatively.

(2) The relative free energy of a particular interaction between any pair of components can be deduced from the inverse of Boltzmann's law:

$$G(d_{ab}) = -kT \ln \frac{\rho(d_{ab}|C)}{\rho(d_{ab})} \qquad (8)$$

where k is the Boltzmann constant, T is the absolute temperature, $\rho(d_{ab}|C)$ is the density of groups a and b at a separation d in experimental protein structures, and $\rho(d_{ab})$ is the same density in the reference state (McQuarrie, 1976). Two assertions underlie this expression: each occurrence of d_{ab} in experimental structures is independent of any thing else in the environment, and the distribution over a set of observations of d_{ab} in a number of different proteins should obey Boltzmann's

distribution. Sippl *et al*. (1996) have argued cogently that the latter point is true.

(3) The lowest free energy conformation represents the native state (often called the thermodynamic hypothesis) (Anfinsen, 1973).

Substituting equation (8) into equation (7), we have:

$$G_{\text{tot}} = -kT \sum_{ij} \ln \frac{\rho(d_{ab}^{ij}|C)}{\rho(d_{ab}^{ij})} \qquad (9)$$

This expression is similar to our scoring function for the conditional probability formalism (equation (4)), except that we are dealing with the density of states with particular distance values, rather than the probability. However, such densities are also equivalent to the frequencies defined in equation (4), and are evaluated inexactly the same way from experimental structures. Thus, seeking the conformation with the lowest free energy based on a potential of mean force analysis is formally equivalent to seeking the one with the largest probability of presenting a native structure in terms of a Bayesian statistics. Because of the less clear assumptions involved in the potential of mean force analysis, we prefer to use the simpler Bayesian probability formalism in this work.

The residue-specific all-atom probability discriminatory function

The conditional probabilities for the residue-specific all-atom probability discriminatory function (RAPDF) are compiled from frequencies of contacts between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered, and the description of the atoms is residue specific, i.e. the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. Contacts between atoms within a single residue are excluded from the counts. We divide the distances observed into 1.0 Å bins ranging from 3.0 to 20.0. Contacts between pairs of atom types in the 0.0 to 3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins. Table 1 lists the atom types used for this discriminatory function.

A table containing the negative log conditional probability scores, *S*, for all pairs of atom types for all distances is compiled from a database of known structures using equation (4). Given an amino acid sequence in a particular conformation, the scores of all contacts between pairs of atom types with distances that fall within the distance cutoff above (ignoring contacts between atoms within a single residue) are summed up to yield the total score supporting the hypothesis that the conformation is correct. All counts were initialized to one to avoid taking the log of zero. This procedure is used for all probability discriminatory functions described in this paper.

The residue-specific virtual-atom probability discriminatory function

The conditional probabilities for the residue-specific virtual-atom probability discriminatory function (RVPDF) are compiled from frequencies of contacts between pairs of virtual atom types in a database of protein structures. A virtual atom approximation similar to the one developed by Head-Gordon & Brooks (1991) is used. This representation combines a group of atoms into a single virtual atom type by averaging over the corresponding *x*, *y*, and *z* Cartesian coordinates of the individual atoms. Aside from labeling conventions, the present representation differs from the original one in the determination of the virtual centers for virtual atoms vNH and vOH, taken here to be represented by the positions of the N and O atoms, respectively, rather than the geometric centres. The distance bins are the same as in the RAPDF. Each of the virtual atom types is prefixed by the type of the residue, resulting in 105 different virtual atom types. Table 2 lists the virtual atoms used for this discriminatory function and the combinations of atom types they represent.

The non-residue-specific virtual-atom probability discriminatory function

The non-residue-specific virtual-atom probability discriminatory function (NVPDF) differs from RVPDF only in that the virtual atom types are not residue-specific. For example, all vC_α atom types are considered the same, and all vC_β atom types are considered the same, and so on. The total number of virtual atom types considered under this approximation is 21.

The contact discriminatory function

To examine how well non-specific compactness alone can discriminate between correct and incorrect conformations relative to the three PDFs (RAPDF, RVPDF and NVPDF) described above, we use a simple contact discriminatory function which assigns a score of −1.00 for every atom–atom contact within 6.0 Å in an amino acid

Table 1. List of atom types used in the all-atom residue specific probability discriminatory function (RAPDF). Each of these atom types is prefixed by the type of the residue, resulting in 167 different atom types

C	C_α	C_β	C_δ	$C_{\delta1}$	$C_{\delta2}$	C_ε	$C_{\varepsilon1}$	$C_{\varepsilon2}$	$C_{\varepsilon3}$	C_γ	$C_{\gamma1}$
$C_{\gamma2}$	CH_2	C_ζ	$C_{\zeta1}$	$C_{\zeta2}$	N	$N_{\delta1}$	$N_{\delta2}$	N_ε	$N_{\varepsilon1}$	$N_{\varepsilon2}$	NH_1
NH_2	N_ζ	O	$O_{\delta1}$	$O_{\delta2}$	$O_{\varepsilon1}$	$O_{\varepsilon2}$	O_γ	$O_{\gamma1}$	OH	S_δ	S_γ

Table 2. List of virtual atom types used in the residue-specific and non-residue-specific virtual-atom probability discriminatory functions (RVPDF and NVPDF, respectively). The table lists the virtual atom type, the atom types of the components, and the residues it is present in (in one-letter code). For the RVPDF, each of the virtual atoms is prefixed by the type of residue (in one-letter code), resulting in 105 different virtual atom types.

Virtual atom	Components	Present in residue
vCO	$C + O$	all
vNH1	N	all
vNH2	$N_{\delta 2}$	N
vNH2	$N_{\varepsilon 2}$	Q
vNH2	NH_1	R
vNH2	NH_2	R
vNH3	N_{ζ}	K
vNHE	$N_{\delta 1}$	H
vNHE	$N_{\varepsilon 2}$	H
vNHE	N_{ε}	R
vNHE	$N_{\varepsilon 1}$	W
vCOS	$C_{\gamma} + O_{\delta 1}$	N
vCOS	$C_{\delta} + O_{\varepsilon 1}$	Q
vCSC	$C_{\gamma} + S_{\delta} + C_{\varepsilon}$	M
vCCC	$C_{\beta} + C_{\gamma 1} + C_{\gamma 2}$	V
vCCC	$C_{\gamma} + C_{\delta 1} + C_{\delta 2}$	L
vC3R	$C_{\beta} + C_{\gamma 1} + C_{\varepsilon 1}$	F
vC3R	$C_{\delta 2} + C_{\varepsilon 2} + C_{\zeta}$	F
vC3R	$C_{\beta} + C_{\gamma 1} + C_{\varepsilon 1}$	Y
vC3R	$C_{\delta 2} + C_{\varepsilon 2} + C_{\zeta}$	Y
vC3R	$C_{\delta 2} + C_{\varepsilon 3} + C_{\zeta 3}$	W
vC3R	$C_{\varepsilon 2} + C_{\zeta 2} + CH_2$	W
vCOO	$C_{\gamma} + O_{\delta 1} + O_{\delta 2}$	D
vCOO	$C_{\delta} + O_{\varepsilon 1} + O_{\varepsilon 2}$	E
vOH	O_{γ}	S
vOH	$O_{\gamma 1}$	T
vOH	OH	Y
vCC	$C_{\beta} + C_{\gamma 2}$	I
vCC	$C_{\gamma 1} + C_{\delta 1}$	I
vCC	$C_{\beta} + C_{\gamma}$	K
vCC	$C_{\delta} + C_{\varepsilon}$	K
vCC	$C_{\gamma} + C_{\delta}$	R
vCC	$C_{\beta} + C_{\gamma 2}$	T
vCCP	$C_{\beta} + C_{\gamma}$	P
vCCR	$C_{\gamma} + C_{\delta 1}$	W
vCCH	$C_{\gamma} + C_{\delta 2}$	H
vSH	S_{γ}	C
vCE1	$C_{\varepsilon 1}$	H
vC	C_{ζ}	R
vCH3	C_{β}	A
vCH2	C_{α}	G
vCH2	C_{β}	C,D,E,F,H,K,M,N,Q,R,S,W,Y
vCH2	C_{γ}	E,Q
vCH2	C_{δ}	P
vCH	C_{α}	A,C,D,E,F,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y

sequence conformation, excluding contacts within a single residue.

The linearly interpolated residue-specific all-atom probability discriminatory function

The three PDFs (RAPDF, RVPDF, NVPDF) use discrete bins to compile the probability scores. This leads to a situation where the score for observing any distance within a bin width is the same for a given pair of atom types. In reality, the preferences between atom types vary in a continuous manner as the distances between the contacts vary. We thus also evaluate the total score of a conformation by linearly interpolating between the scores for the discrete bins to uniquely define the score for a given distance. The linear interpolation is used to modify the RAPDF.

Interpolation is based on two assumptions: (1) The value at the mid-point of the distance bins is given by the score for that distance bin. (2) There is a linear relationship between probability values observed for neighbouring bins.

Thus, if d_a represents the actual distance encountered, d_l represents the mid-point of the closest distance bin value on the left-hand side, d_r represents the mid-point of the closest distance bin on the right-hand side, and S_l represents the score for d_l, and S_r the score for d_r. S_a, the score for d_a is given by:

$$S_a = S_l + \left((S_r - S_l) \cdot \frac{d_a - d_l}{d_r - d_l} \right) \quad (10)$$

Low counts analysis

We have investigated the effect of finite counts used to define the frequencies required for a PDF (equation (4)), by introducing a simple model of count variation. For each set of counts obtained from observations of particular pairs of atom types in the database in equations (5) and (6), we assume a Gaussian distribution for variation in counts in repeated experiments and modify the counts accordingly. A set of possible alternative counts is then obtained using the equation:

$$N' = \left(\sum_m R - \frac{m}{2} \right) \cdot \sigma + N \quad (11)$$

where N is the observed count value, N' is the modified count value, $\Sigma_m R$ represents the sum of m random numbers R in the interval [0,1] ($m = 12$), and σ represents the standard deviation in the observed counts, assumed to be $N^{1/2}$.

We generate 100 different sets of modified counts each of the terms in equation (4) for the 18 distance bins using the above procedure, thus providing 100 different sets of conditional probability curves that might have been obtained from a protein database of the same type and size used in this work. These modified curves are compared to the observed one for particular pairs of atom types.

Selection of the structure library for obtaining conditional probabilities

Table 3 lists the PDB codes of the 265 structures that were used for compiling the conditional probabilities for the discriminatory functions described above. The PDB codes were initially obtained from the CATH database and are a set of non-homologous (less than 30% sequence identity between any proteins in the set) X-ray structures better than 3.0 Å resolution (Orengo *et al.*, 1993). Structures with multiple side-chain conformations have been modified such that only the side-chain confor-

Table 3. List of PDB (Bernstein *et al.*, 1977) codes of the 265 protein chains used for compilation of conditional probabilities. In cases where a single chain of the protein is used for the compilation, the chain identifier is shown

1351	1c5a	1gox	1oma	1tabI	2hpdA	3pgk
1aaf	1cauA	1gpb	1omf	1ten	2ltnA	3pgm
1aak	1cdb	1gpr	1ovb	1tfi	2ltnB	3rubS
1aba	1cde	ahcc	1pba	1tgl	2mev4	3sc2A
1aco	1cdg	1hgeA	1pfa	1thg	2mnr	3sc2B
1acp	1cewI	1hgeB	1pdc	1tml	2msbA	4enl
1add	1cmbA	1hleA	1pfkA	1tnfA	2nckL	4fgf
1adn	1cobA	1hmy	1pgd	1tplA	2ohxA	4gcr
1ads	1colA	1hoe	1pgx	1tpm	2ovo	4htcI
1ak3A	1coy	1hsbA	1pha	1ttaA	2pia	4mt2
1ala	1cpcA	1hstA	1phh	1ttf	2plv4	4sbvA
1alkA	1cpt	1huw	1pii	1ula	2pmgA	4sgbI
1aozA	1csc	1hyp	1pkp	1utg	2polA	5fd1
1apa	1cseE	1ifc	1plc	1vil	2reb	5p21
1apmE	1cseI	1ipd	1poa	1vsgA	2rhe	5pti
1aps	1ctf	1isuA	1poxA	1wsyA	2rn2	5rubA
1arb	1d66A	1kst	1ppn	1wsyB	2sicI	5timA
1arqA	1dhr	1lab	1prcC	1xis	2sn3	6insE
1atnA	1dmb	1lct	1prcH	1ycc	2sns	7aatA
1atr	1eca	1lfi	1prcL	1ysaC	2stv	7catA
1atx	1ede	1lis	1ptf	1zaaC	2tgi	7rsa
1ayh	legf	1lla	1pyaA	256bA	2tmdA	8abp
1bal	1etrL	1lmb3	1pyaB	2aaiB	2tmvP	8fabB
1bbpA	1ezm	1ltsA	1pyp	2bbkH	2tsl	8rxnA
1bbt1	1fbaA	1ltsC	1raiA	2bbkL	2tscA	9wgaA
1bbt2	1fc2D	1lyaA	1raiB	2bopA	2yhx	
1bbt3	1fiaA	1mat	1rcb	2bpal	3b5c	
1bbt4	1fkb	1mfaH	1rec	2bpa2	3bcl	
1bds	1fnr	1mfa:	1rfbA	2cas	3blm	
1bgc1	1fus	1minA	1rhd	2cba	3cla	
1bgc2	1fxd	1minB	1ribA	2cdv	3cts	
1bgh	1gal	1mypA	1rip	2cmd	3dfr	
1bha	1gatA	1mypC	1rro	2cpl	3ebx	
1bia	1gd1O	1nar	1rveA	2ctc	3ecaA	
1bllE	1gdhA	1nipA	1sbp	2ctvA	3gapA	
1bmv1	1gky	1noa	1shaA	2cyp	3grs	
1bmv2	1glaG	1nrcA	1shg	2dnjA	3il8A	
1brnL	1glt	1nrd	1sim	2er7E	3mdsA	
1btc	1gluA	1nscA	1sryA	2gstA	3monA	
1bw3	1gof	1ofv	1stp	2hhmA	3monB	

Table 4. Class I decoys. The name used to identify the specific decoy set, the number of decoys in the set, and C_α rmsd range of the decoys to the experimental structure, and the appropriate references are given

Decoy set name	Number of decoys	C_α RMSD range (Å)	Reference
MISFOLD	25	8.66–22.43	(Braxenthaler *et al.*, 1997; Holm & Sander, 1992b)
CASPI	42	0.53–7.40	(Mosimann *et al.*, 1995; Braxenthaler *et al.*, 1997)
IFU	44	0.21–10.02	(Braxenthaler *et al.*, 1997; Pedersen & Moult, 1997)
PDBERR	3	0.81–13.21	(Braxenthaler *et al.*, 1997)
SGPA	2	1.91–2.06	(Braxenthaler *et al.*, 1997; Avbelj *et al.*, 1990)

mations with atoms having the highest occupancy and/or lowest temperature factors are used.

Decoy set generation

The decoy sets used were obtained from the Protein Potential Site (PROSTAR) (Braxenthaler *et al.*, 1997) and can be divided into two classes. Decoy sets in class I discriminate between one correct and one or more incorrect or approximate conformations. Decoys sets in class II are a set of approximate conformations that vary in rmsd to the experimental conformation, excluding the experimental conformation itself.

Table 4 lists the decoy sets in class I. The MISFOLD decoy set, generated by Holm & Sander (1992a), consists of 25 examples of pairs of proteins with the same number of residues in the chain, but different sequences and conformations. Sequences were swapped between members of a pair, and side-chain packing annealed using a Monte Carlo process (Holm & Sander, 1992a). These provide inappropriate environments for most of the side-chains in the structures.

The first experiment on the Critical Assessment of protein Structure Prediction methods (CASP2) produced a set of 42 comparative models of six different proteins (Mosimann *et al.*, 1995). These form the CASP1 decoy set. The models vary in C_α rmsd to the corresponding experimental conformation, ranging from 0.53 to 7.40 Å, depending on the difficulty of the model building process.

The IFU decoy set is based on a set of 44 peptides which are proposed to be independent folding units as determined by local hydrophobic burial and experimental evidence (Unger & Moult, 1991). The set consists of the structure of the peptides as observed in the complete experimental protein structure, and a conformation of the fragment generated with a Genetic Algorithm (Pedersen & Moult, 1997) and a physics-based potential of mean force (Avbelj & Moult, 1995a).

The PDBERR decoy set consists of structures determined using X-ray crystallography which were later found to contain errors, and the corresponding corrected experimental conformations (Braxenthaler *et al.*, 1997). The SGPA decoy set consists of two conformations generated by molecular dynamics simulations starting with the *Streptomyces griseus* Protease A experimental structure (PDB code 2sga) (Avbelj *et al.*, 1990), and the 2sga experimental structure.

Among all the decoy sets referenced in this paper, only the LOOP decoy set belongs to class II. This decoy set consists of sets of conformations for short loops (four or five residues) that were systematically generated using the methods of Moult & James (1986) and Fidelis *et al.* (1994). The loop conformations are evaluated in the context of the rest of the experimental structure. Table 5 gives details about each of the loops in the LOOP decoy set.

Table 5. Class II decoys. The LOOP decoy set is a set of loop conformations that were systematically generated using the methods of Moult & James (1986) and Fidelis *et al.* (1994). For each loop, the number of different loop conformations and the all-atom rmsd range is given

	PDB code	Residue range	Sequence	Number of conformations	All-atom rmsd range (Å)
1	3dfr	20–23	PWHL	394	0.75–4.58
2	3dfr	27–30	LHYF	1390	0.81–3.47
3	3dfr	64–68	HQED	71439	0.89–4.19
4	3dfr	120–124	GSFEG	474	0.57–2.91
5	3dfr	136–139	FTKV	10782	1.39–2.15
6	2sga	35–39	TNISA	15453	1.20–3.17
7	2sga	97–101	GSTTG	2079	0.60–3.34
8	2sga	116–119	YGSS	26572	0.47–4.91
9	2sga	132–136	AQPGD	206	0.97–2.58
10	2fbj	265–269	HPDSG	393	0.96–3.90
11	2hfl	264–268	LPGSG	339	1.11–2.81

Decoy set evaluation

For Class I decoy sets, the ratio of the score of the incorrect conformation to that of the correct conformation is determined. A discrimination ratio less than 1.0 (or log discrimination ratio less than 0.0) indicates that the discriminatory function is able to distinguish between the correct conformation and the incorrect one. The lower the log discrimination ratio, the more reliable the discrimination.

For class II decoy sets, two evaluation measures are used. One measure is the probability of choosing by chance a conformation with an equal or lower rmsd than the one selected by the discriminatory function. That is, the number of conformations that have the same or lower all-atom rmsd, as the rmsd of the conformation with the lowest score is divided by the total number of conformations in the set. The second measure is whether or not a conformation with an all-atom rmsd within 1.0 of the lowest rmsd conformation present in the decoy set is selected by the discriminatory function.

Results of a complete decoy set are summarized by calculating the average discrimination ratio of all the decoys and/or as a percentage of decoys correctly discriminated. Further details on the evaluation protocols and the decoy set generation are given by Braxenthaler *et al*. (1997). Proteins in the structure library that are more than 30% identical in sequence to a protein in a particular decoy set were not included in the compilation of the discriminatory function, i.e. the procedure was jack-knifed.

Results

The residue-specific all-atom discriminatory function performs the best across a wide variety of decoys

An ideal discriminatory function is one that correctly discriminates 100% of class I decoys and selects conformations with low all-atom rmsds (within 1.0 Å of the conformation with the lowest rmsd) in the LOOP decoy set. In addition, the average discrimination ratios should be as low as possible.

The RAPDF comes close to achieving the goal of 100% discrimination, and performs significantly better than the RVPDF and NVPDF. Figure 1a and b shows that the RAPDF has the best average discrimination ratio and the largest percentage of decoys correctly discriminated across a range of decoy sets. In the case of the MISFOLD, PDBERR, and SGA decoy sets, it correctly discriminates 100% of the decoys in the set. Further, the average discrimination ratios show that the scores for the correct conformations in the MISFOLD, PDBERR, and SPGA decoy set are on average lower (better) by 60%, 50% and 25%, respectively, compared to the scores for the incorrect conformations.

For the IFU decoy set, the percentage of conformations correctly discriminated by the RAPDF is 73%. The average difference between the scores for the correct conformations and the incorrect conformations for the RAPDF is 10%.

Figure 2 shows the results in more detail for some of decoys in the CASP1 set, which is a set of comparative models and their corresponding experimental structures. Here, the overall percentage of decoys correctly discriminated by the RAPDF is 93%, for the 42 decoys in this set. The average difference between the scores for the correct conformations and the incorrect conformations for the RAPDF is 15%. The Figure shows the results for only the models with the lowest C_α rmsd to the corresponding experimental structures. The RAPDF does not perform as well as the two other discriminatory functions in terms of the discrimination ratio in two instances (see the bars representing nm23 and hpr in Figure 2).

It is less obvious which discriminatory functions perform best from the log probability and all-atom rmsd data for the LOOP decoy set (Figure 3). However, if we examine the actual rmsd values, we note that for 10 out of 11 loops, the RAPDF picks a conformation that is within 1.0 Å of the lowest rmsd conformation in the sample space (where the experimental structure is not included). RVPDF, NVPDF, and CDF have ratios of 5/11, 9/11, and 8/11, respectively.

Relationship between probability score and rmsd

Plotting the rmsd *versus* the score for a set of conformations from the LOOP decoy set (Figure 4) shows that the RAPDF correlates reasonably well with the all-atom rmsd up to quite large rmsds. This suggests that this discriminatory function can be useful in simulations that attempt to get closer to the native conformation starting from a distant conformation.

Discriminatory power decreases upon successive approximations

There is an overall progressive deterioration in the signal going from the all-atom residue-specific discriminatory function to the CDF, as the description gets more and more approximate (Figure 1). One major exception is the LOOP decoy set (Figure 3) where the NVPDF performs significantly better than the RVPDF. The other exception is apparent in Figure 1b, for the MISFOLD decoy set.

The compactness term alone is useful for discriminating between correct and incorrect conformations

The contribution of the compactness term, which is measured by the CDF, is better than some of the other PDFs for certain decoy sets (Figure 1b under MISFOLD, and Figure 3). Further, in a majority of the decoy sets, it is adequate to distinguish

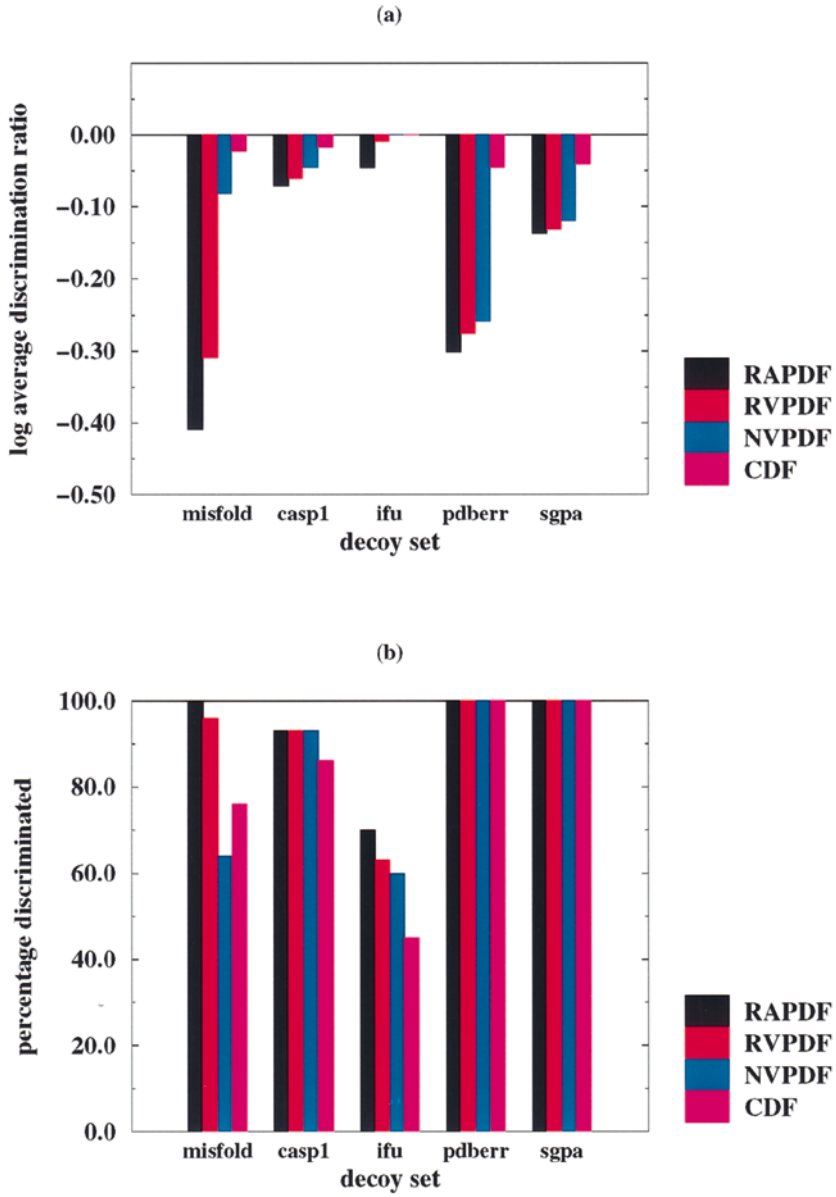

Figure 1. Comparison of the performances of the residue-specific all-atom probability discriminatory function (RAPDF), the residue-specific virtual-atom probability discriminatory function (RVPDF), the non-residue-specific virtual-atom probability discriminatory function (NVPDF), and the contact discriminatory function (CDF) for class I decoy sets. The log of the average discrimination ratios between incorrect and correct conformations for the five decoy sets in class I are shown in (a). The lower the log average discrimination ratio, the better the discrimination. The percentage of decoys that were correctly discriminated within a decoy set are shown in (b). In all cases, the most detailed function, RAPDF, performs best, and achieves 100% discrimination for three out of five decoy sets.

between correct and incorrect conformations most of the time (Figure 1b under PDBERR and SGA, and Figure 3).

Using a large distance cutoff helps in discrimination

Figure 5 shows the percentage of decoys correctly discriminated at different distance cutoffs (5.0, 10.0, 15.0, and 20.0 Å). There is a significant advantage overall to using a larger distance cutoff. A distance cutoff of at least 15.0 Å is necessary to accurately discriminate all the 25 decoys in the MISFOLD decoy set. In the cases of the PDBERR and SGPA decoy sets, it does not appear to make a difference which cutoff is chosen. The average discrimination ratios for the four different cutoffs show a similar trend for all decoy sets.

Comparison of the contribution of electrostatics and non-electrostatics terms

To gain some insight into the nature of the signal in the RAPDF, we partition the discriminatory function according to contributions from electrostatic and non-electrostatic contacts. Interactions between any combination of nitrogen and oxygen atoms are defined to be electrostatic in nature. All other contacts are considered non-electrostatic. Figure 6 compares the effectiveness (by measuring the percentage of conformations correctly discriminated) of using only the electrostatic or non-elec-

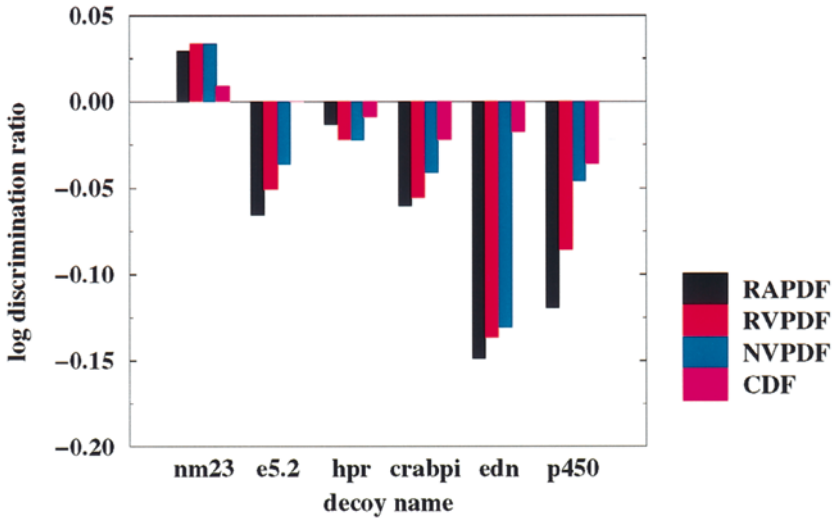

Figure 2. Comparison of the performances of the four discriminatory functions (RAPDF, RVPDF, NVPDF, and CDF) for selected decoys in the CASP1 set. The log discrimination ratios between the experimental conformation and the model is shown. The model with the lowest C_α rmsd to the corresponding experimental conformation is chosen from a given set of models for this evaluation. The identifiers used to label the decoys are those used by Mosimann *et al.* (1995). Best discrimination is achieved with the RAPDF, except in two cases where all the decoys have a low C_α rmsd to the experimental structure.

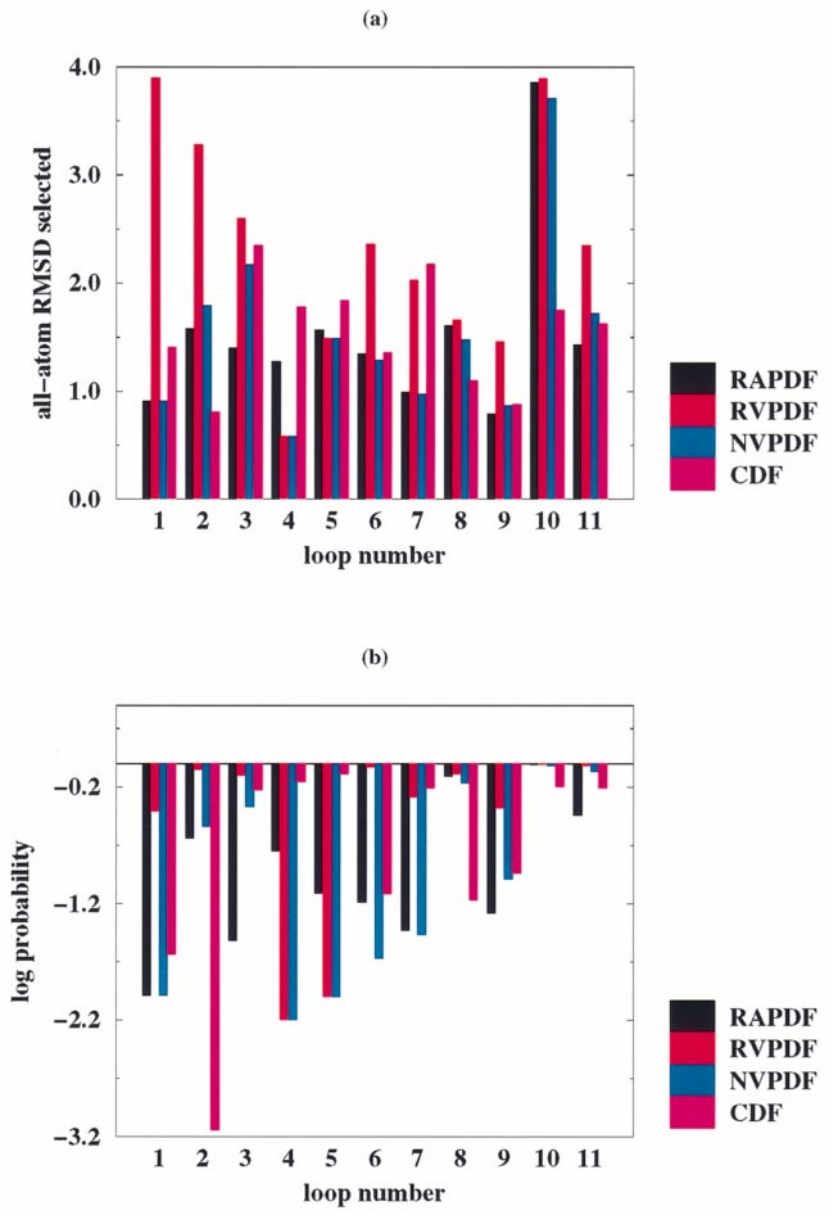

Figure 3. Comparison of the performances of the residue-specific all-atom probability discriminatory function (RAPDF), the residue-specific virtual-atom probability discriminatory function (RVPDF), the non-residue-specific virtual-atom probability discriminatory function (NVPDF), and the contact discriminatory function (CDF) for the LOOP decoy set. The all-atom rmsd of the conformation selected by each discriminatory function (a) and the log probabilities of observing an equal or lower rmsd by chance (b) are shown. The loop numbers in the horizontal axis correspond to the numbers in Table 5, column 1. Performance of the different PDFs is more varied with this decoy set, although the RAPDF picks out a conformation with an rmsd that is within 1.0 Å of the lowest available in ten out of 11 cases.

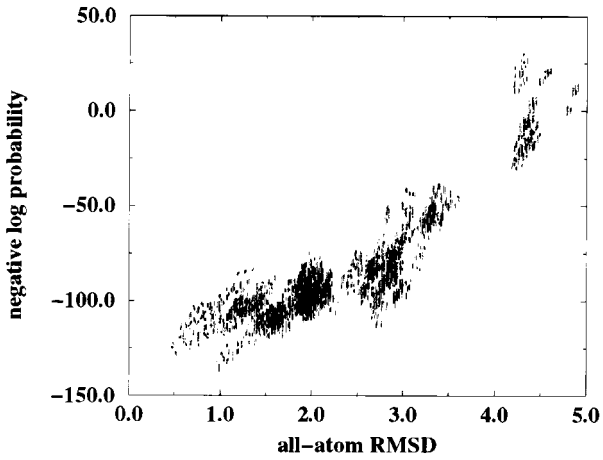

Figure 4. Performance of the residue-specific all-atom probability discriminatory function (RAPDF) for a selected loop in the LOOP decoy set. The all-atom rmsd Å *versus* the negative log conditional probability of the 26,572 conformations for the 2sga 116 to 119 LOOP set is shown. The negative log conditional probability of a conformation increases as the rmsd progressively gets worse, a useful property for long-range model refinement.

trostatic terms, relative to the full PDF. Although the non-electrostatic terms alone are adequate for correct discrimination in most cases, the electrostatic term does not play a significant role in enhancing the signal. This is particularly noticeable in the CASP1, IFU, and LOOP decoy sets. The average discrimination ratios for the decoy sets using the electrostatics, non-electrostatics, and combined terms (data not shown) indicate a similar trend except in the case of the CASP1 and IFU decoy sets, where the ratios are similar regardless of the terms used.

Linear interpolation improves discrimination

Comparison between the IRAPDF and the RAPDF (Figure 7) shows that linear interpolation helps discriminate between correct and incorrect conformations. This is most obvious in the LOOP decoy sets where the improvement is quite dramatic, but for each decoy set there is some improvement. For all decoys, the percentage of structures correctly discriminated is identical whether or not the conditional probabilities are interpolated.

The problem of sparse data for compilation of probabilities is negligible

We examine the uncertainty due to the finite counts in the individual contributions to $P(C \mid d_{ab})$, the probability of observing a correct conformation given a contact between atom types a and b at a distance d:

$$P(C|d_{ab}) = \frac{P(d_{ab}|C)}{P(d_{ab})} = \frac{N(d_{ab})/\Sigma_d N(d_{ab})}{\Sigma_{ab} N(d_{ab})/\Sigma_d \Sigma_{ab} N(d_{ab})} \quad (12)$$

Equation (12) is formed by expanding equation (4) in terms of equations (5) and (6). Detailed explanations for these terms are given in the Methods section. To begin our analysis, let us examine Table 6 for the nature of the raw counts that we encounter in our observations for each of the four terms in the above expression. Since both terms in the denominator in equation (12) are sums over all atom types, they are well determined and do not lead to significant uncertainty in the probabilities. In the denominator, $\Sigma_d N(d_{ab})$ is also well determined, even for the rarest atom types. Thus, all the uncertainty arises for the individual $N(d_{ab})$ counts, which may indeed be low: in a rare atom pair, we might count zero or one observation, for example, and have an uncertainty of a factor of 2 in the resulting probability.

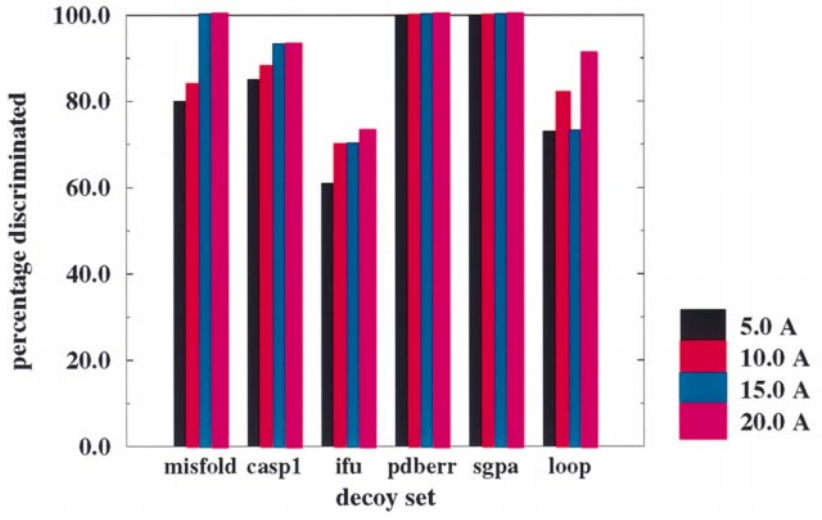

Figure 5. Performance of the residue-specific all-atom probability discriminatory function (RAPDF) at different cutoffs. The percentages of structures correctly discriminated for six decoy sets at four different cuttoffs is shown. In the case of the LOOP decoy set, ''correct'' discrimination is defined to be the selection of conformation that is within 1.0 Å of the lowest all-atom rmsd conformation for each loop. Generally, discrimination progressively improves up to a cutoff of 20.0 Å.

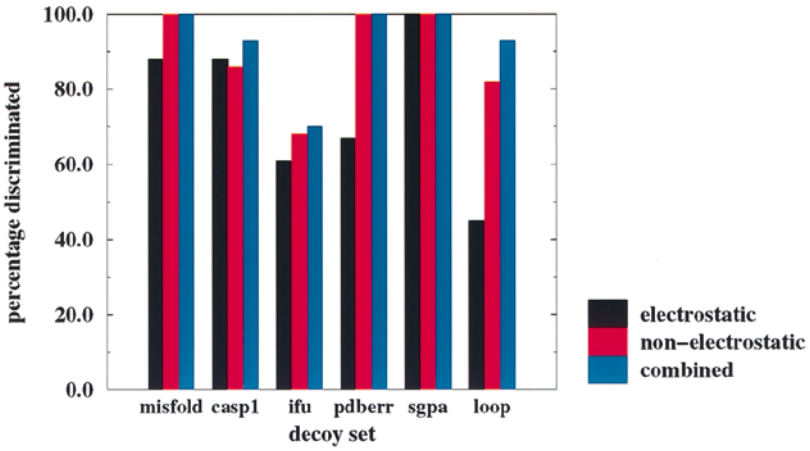

Figure 6. Comparison of electrostatic, non-electrostatic, and combined terms in the residue-specific all-atom probability discriminatory function (RAPDF). The percentages of structures correctly discriminated for various decoy sets is shown. In general, the non-electrostatic terms provide the most signal.

To analyse the uncertainty in the conditional probabilities due to the dataset dependence of the counts, we examine two atom–atom preferences, cysteine N–tryptophan O (CN-WO), a minimum counts situation, and isoleucine C_α–leucine $C_{\delta 2}$ (IC_α-$LC_{\delta 2}$), an average counts situation.

Figure 8 compares the effect of uncertain count values (generated as described in the Methods section) on the conditional probabilities for the preferences between these two pairs of atom types. There is significantly more error in the conditional probabilities for the worst case (CN-WO) than for the average case (IC_α-$LC_{\delta 2}$). In the average case, the absolute error in the scores is generally less than 0.1 and has a maximum value of about 1.0 (in the 0 to 3.0 Å distance bin). In the worst case, the absolute error in the scores is generally less than 0.5 and has a maximum value of about 1.0.

Relationship between the conditional probabilities and the nature of physical interactions in proteins

One would expect that the conditional probability distributions should reflect the known physical and entropic effects that are believed to be important in determining protein structure (Dill, 1990). With so much data, it is not possible to systematically assess this relationship. We have examined a few of the distributions to try to understand the features in these terms.

We first examine the set of C_α–C_α and C_β–C_β interaction probabilities (Figure 9). Although there is a significant spread among the residue types, the curves show several consistent features. In the C_α–C_α distributions, there are two minima, between 3.0 and 4.0 and between 5.0 and 6.0. The C_β–C_β curves show only the second, in the 5.0 to 6.0 range. The first C_α–C_α minimum is a simple consequence of the covalently determined distance between adjacent C_α atoms in the polypeptide chain, around 3.8 Å. The origin of the second one is less obvious. An analysis of the interactions in the proteins in Table 3 reveals that this feature arises primarily from C_α–C_α distances separated by two or three residues (i, $i+2$ and i, $i+3$) in α-helices. Similarly, the single C_β–C_β minimum has its origin in i, $i+1$ and i, $i+2$ separations in both α-helices and β-strands.

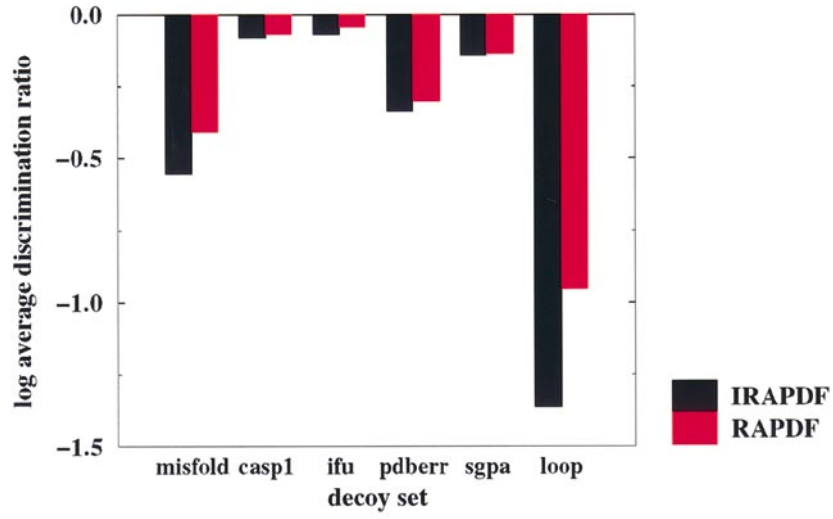

Figure 7. Comparison of the residue-specific all-atom probability discriminatory function (RAPDF) to the linearly-interpolated version (IRAPDF). For five of the decoy sets, the bars represent the log of the average discrimination ratio of the probabilities between the correct and incorrect structures. For the LOOP decoy set, the bars represent the log average of the probabilities of finding at least one structure with a lower all-atom rmsd than the one with the best discrimination by chance (i.e. the sum of log probabilities divided by the number of loops). In all cases, interpolation improves discrimination.

Table 6. Details of the counts obtained when compiling the conditional probabilities. For each term in equation (12), the minimum counts and the average counts is given for all atom types and all distance bins. Only the first term, $N(d_{ab})$, contributes significantly to the uncertainty in the probabilities

Term	Minimum counts	Average counts
$N(d_{ab})$	1	648
$\Sigma_d N(d_{ab})$	1525	11,708
$\Sigma_{ab} N(d_{ab})$	1,011,903	9,143,295
$\Sigma_d \Sigma_{ad} N(d_{ab})$	164,980,971	164,980,971

More dramatic than the consistent features is the large spread in values in both plots in the 0.0 to 3.0 Å bin and in the 3.0 to 4.0 Å bin in the C_β–C_β plot. On closer inspection, these turn out to be mostly irrelevant low count artifacts: there are almost no such distances in experimental structures. However, since all counts are initialized to one, zero probabilities are not possible, and apparent probabilities are dominated by the variation in the total counts over all bins for a pair of atom types, $\Sigma_{ab} N(d_{ab})$. In practice these probabilities will not be used to evaluate conformations. A real component of the signal in the lowest bin for the C_β–C_α plot is provided by *cis*-peptides: the lowest scores are for distances between C_α atoms of proline residues to C_α atoms of other residues, reflecting the higher frequency of such conformations before prolines. In the C_β–C_β plot, the outlier with the score of almost −3.0 is for disulphide linked cysteine-cysteine pairs.

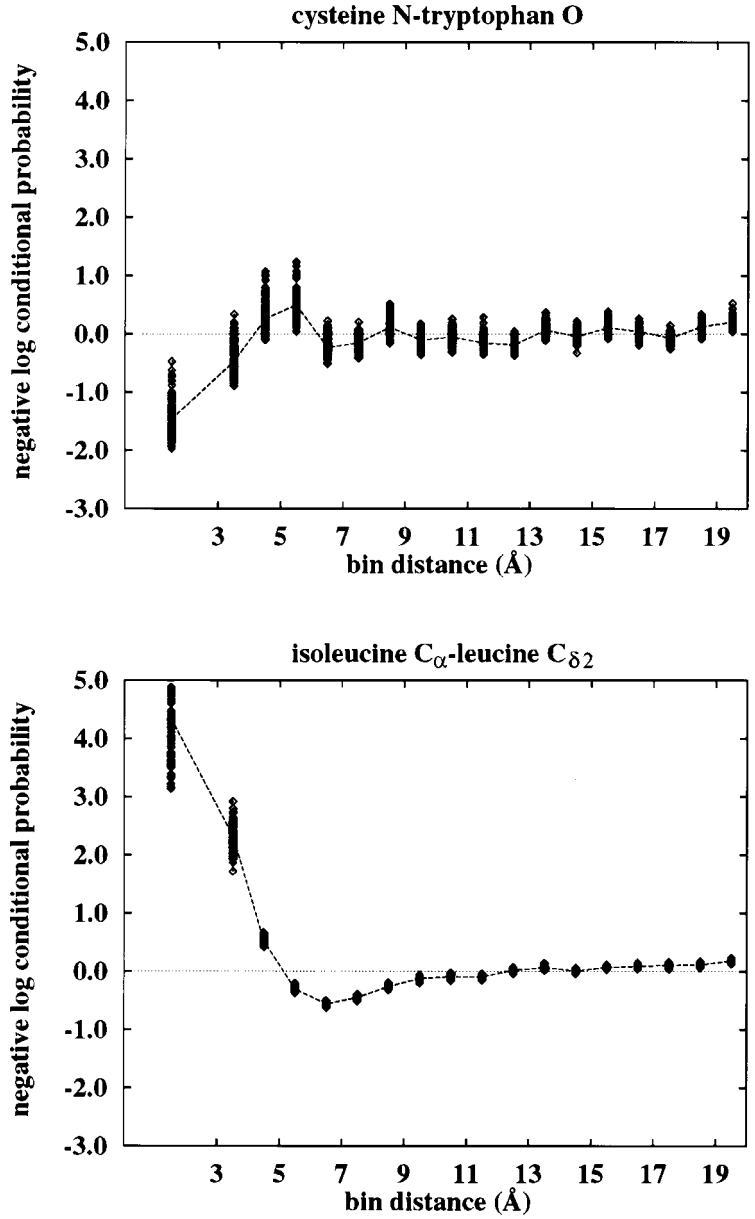

Figure 8. Comparison of the effect of counting uncertainties on the conditional probabilities for two pairs of atom types, cysteine N–tryptophan W (CN-WO), where the counts are among the lowest in the database, and isoleucine C_α–leucine $C_{\delta 2}$ (IC_α-$LC_{\delta 2}$), where the counts in the 0.0 to 3.0 Å bin distance are similar to the average counts in Table 6. The broken line connects observed conditional probabilities and the points around the broken line represent the variation due to the uncertainty in the counts (see the Methods section).

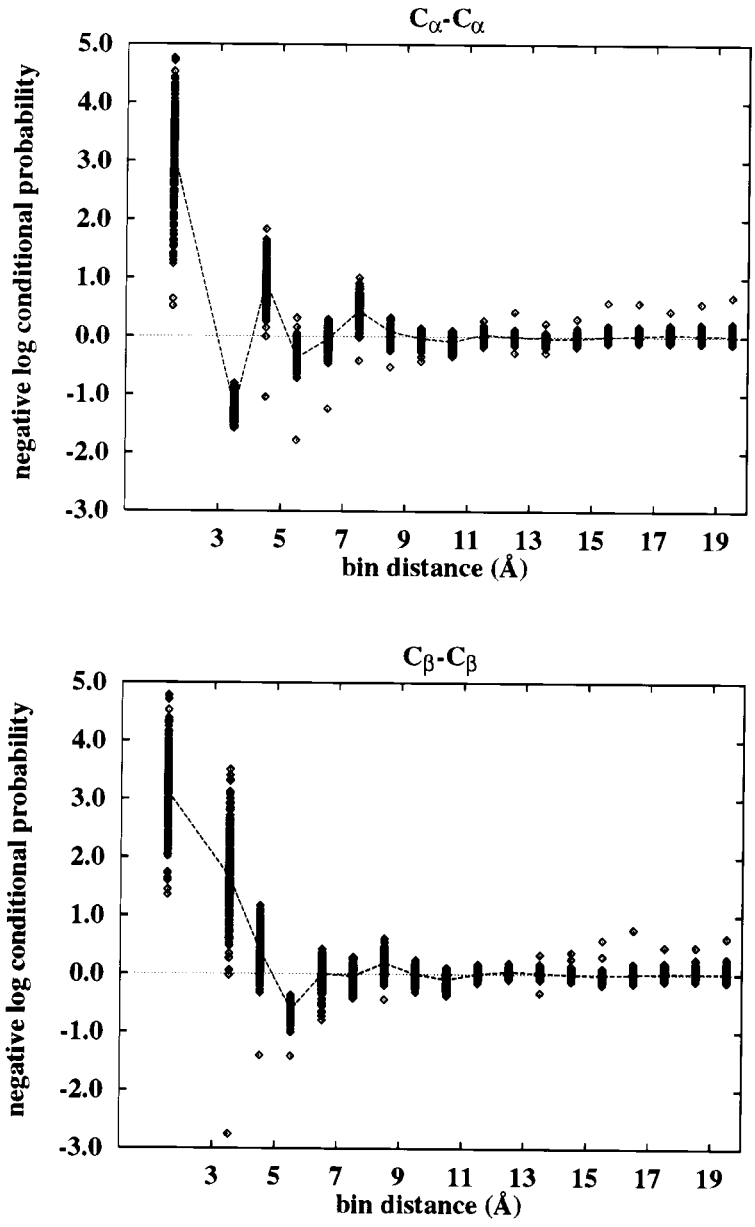

Figure 9. Plot illustrating the conditional probabilities encountered in the 18 distance bins for all C_α–C_α contacts and all C_β–C_β contacts. The spread at a given distance bin illustrates the differences in probabilities for the atom types in different residues. The average of the scores for each bin is connected by the broken line. Consistent, though weak, minima are seen.

C_α–C_α distance plots for particular residue pairs show unique features. For example, differences between valine–valine and alanine–alanine reflect the different frequency of occurrence of these residues in α and β secondary structures (Figure 10). α-Helices have C_α–C_α distances between 5.0 and 6.0 Å for i, $i+1$ and i, $i+3$ pairs, and this is reflected in the slightly deeper minimum in the alanine curve compared to that of valine. Conversely, β strands have distances in the 6.0 to 7.0 Å range for i, $i+2$ residues, producing the deeper minimum for the corresponding bin in the valine curve. This is an example of the type of signal the conditional probabilities incorporate automatically, but which is not easily captured in a physics-based potential.

Similar significant residue specific features can be found in many of the curves. For example, the spread in values in the lowest distance bins for the N-O curves (Figure 11) reflects the different frequencies of backbone hydrogen bonding for different pairs of residue types and is also directly linked to the variation in frequency of occurrence of different residues in alpha and beta secondary structure. Figure 12 shows two extremes of hydrogen bonding preferences between residue pairs. Not surprisingly, curves for proline nitrogen have very low probability of close contact with a main-chain oxygen, reflecting the inability of this nitrogen atom to participate in hydrogen bonds. A little more subtly, the highest probability of hydrogen bonding is between aspartate N and lysine O,

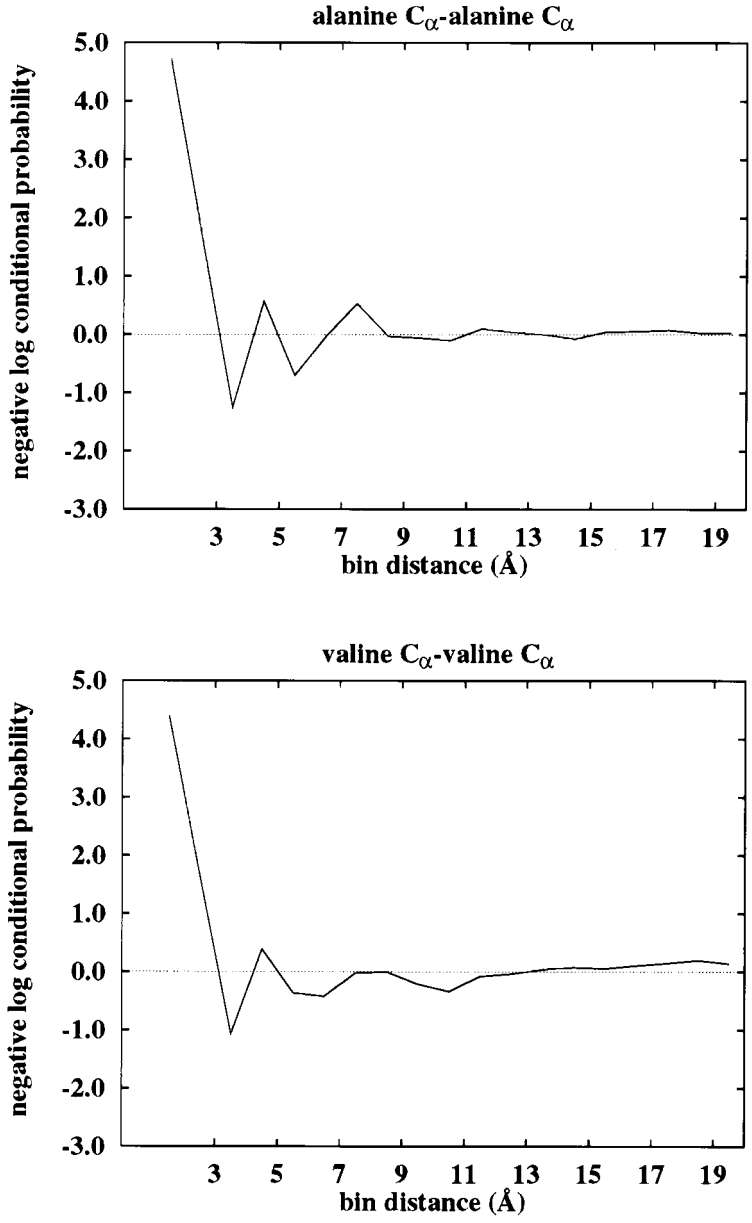

Figure 10. Plots illustrating the conditional probabilities for alanine C_α–alanine C_α (AC_α-AC_α) and valine C_α–valine C_α (VC_α-VC_α). The negative log conditional probabilities are plotted against the 18 distance bins.

reflecting the tendency for i, $i+4$ salt bridges in α-helices, and for salt bridges between residues on adjacent β-strands. Figure 11 also shows the curves for a side-chain–main-chain interaction, N–aspartate $O\delta_1$. The variation in the curve at low distances reflects the different propensity for aspartic acid side-chains to interact with the NH groups of different residue types.

Discussion

Performance of the all-atom residue-specific probability discriminatory function

The most detailed discriminatory function we have tested is successful against a wide range of decoys. Although there are some failures (discussed below), this level of performance suggests the function will at least be useful for protein structure prediction. Other work (Samudrala & Moult, 1997), which assess the predictive power of this discriminatory function in blind tests, indicates that it can distinguish correct side-chain and main-chain conformations from incorrect ones in a real-life modeling scenario.

However, these tests do not rule out that there are sparse false minima in the function's surface, not sampled by the decoys. Further decoy testing is planned, as well as tests searching for experimental structures starting from approximate ones. In this connection, it is encouraging to note that there is a good correlation between the score from the function and the all atom rmsd of a confor-

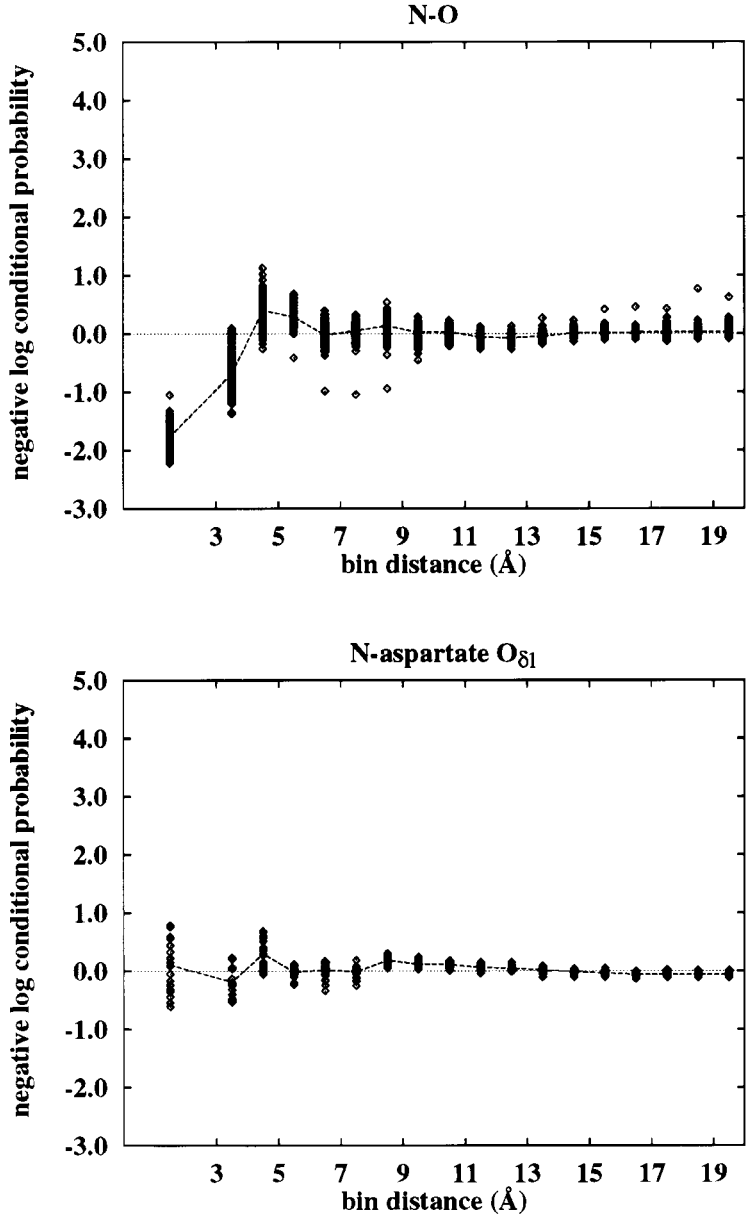

Figure 11. Plot illustrating the conditional probabilities encountered in the 18 distance bins for N-O preferences and preferences between main chain nitrogens and aspartic acid $O_{\delta 1}$ for all residue pairs. The average of the negative log conditional probabilities for each bin is connected by the broken line. The variation in the low bin distances reflects the varying propensities for different residues to form these interactions.

mation (Figure 4); so that a reasonable range of convergence can be expected in such searches.

Effect of approximating the detail in the discriminatory function representation

The more approximate discriminatory functions (RVPDF, NVPDF, CDF) are all able to discriminate between the correct and incorrect structures to some degree, but for the most accurate discrimination across a range of different decoys, an all-atom residue-specific representation is necessary (Figure 1a and b).

Effect of the compactness term on predictive power

As one would expect, the signal from the CDF is consistently less strong than from the other discriminatory functions tested (Figure 1a), except with the loop decoy set, where it performs competitively (Figure 3). In the case of the PDBERR, SGPA and the LOOP decoy sets, the signal is sufficiently strong to provide overall discrimination in most instances. For the MISFOLD, CASP2, and IFU decoy sets, it is less effective. Other workers have noted that compactness alone provides a strong signal when the competing conformations are random coil like (Jernigan & Bahar, 1996; Bahar & Jernigan, 1997). Here we see that the signal may still be very significant in much less obvious situations. Before applying this test, we were unaware that our high rmsd loop conformations tended to be the least compact. These results provide a good illustration of the need for multiple types of decoy set in testing a discriminatory function. A particular set may have a specific property, such as being

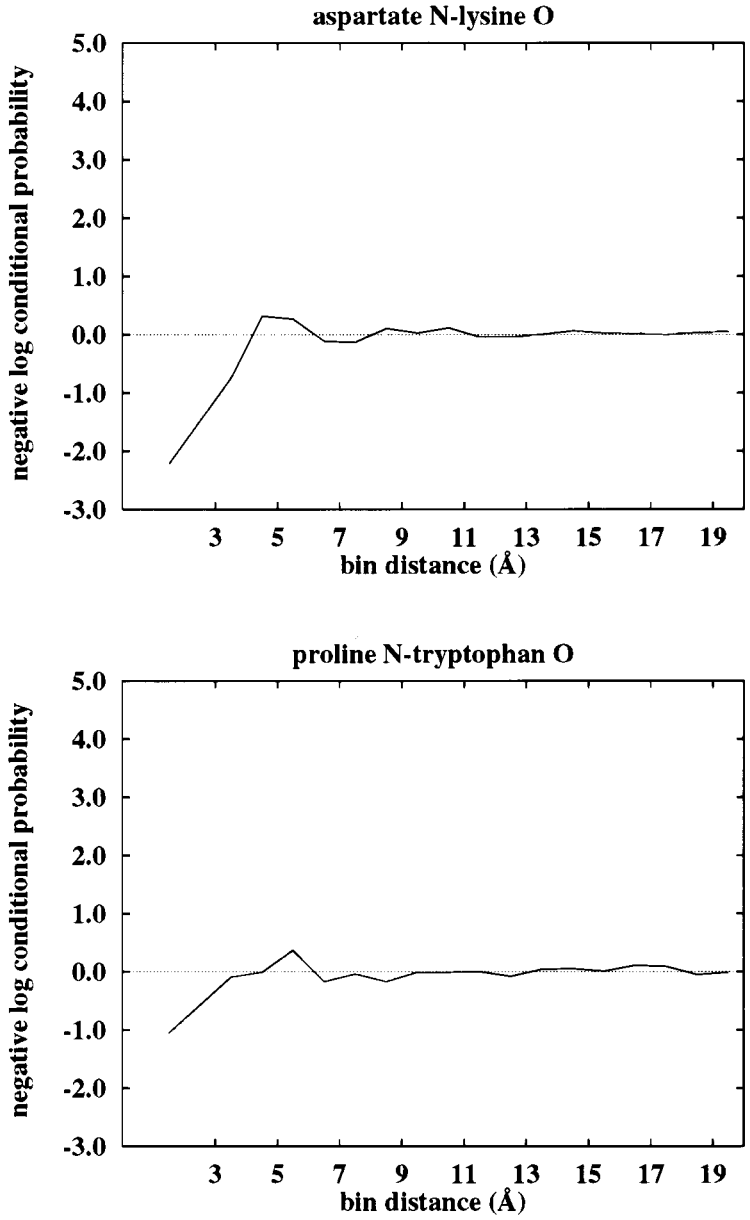

Figure 12. Plots illustrating the conditional probabilities for aspartate N–lysine O and proline N–tryptophan O. The shorter minimum in the 0.0 to 3.0 Å bin for proline N–tryptophan O reflects the inability of proline N to hydrogen bond. The deep minima for aspartate N–lysine O reflects the tendency of these residues to salt bridge in α-helices and β-sheets, resulting in main-chain hydrogen bonds.

less compact than the native structures, and excellent performance may be obtained from a function utilizing that signal. But in a real life application, (for example, assessing comparative models) that signal may not be present.

Effect of using a large distance cuttoff

Using a 20.0 Å distance cutoff results in the most accurate discrimination for the RAPDF (Figure 5). The signal from each atom–atom interaction is extremely weak at such large distances (see Figures 9, 11, 10, and 12), but each atom has a very large number of contacts, so that the combined signal still has an impact. It is unlikely that the energy interaction is significant, except perhaps between charged groups. However, the overall tendency of proteins to be organized ''hydrophobic inside, hydrophilic outside'' may result in a signal in the probability curves (Huang *et al.*, 1995).

Contributions of electrostatic and non-electrostatic terms

Both electrostatic and non-electrostatic terms (Figure 6) contribute to discrimination, but neither one alone performs as well as the combination. With these decoys, the non-electrostatic contribution is usually the most significant, an effect that is most pronounced for the PDBERR and the LOOP decoy sets. The electrostatic contribution is probably dominated by the main-chain hydrogen binding, since these probabilities tend to have pronounced features (Figure 10), and there are many such terms contributing. In the loops set, the backbone hydrogen bonding groups tend to be more

solvated than in most of the structure contributing to the probabilities.

Effect of linear interpolation and the problem of sparse data

The discriminatory functions described here were constructed using the simplest possible models. A more sophisticated model would take into account the effects of low counts in the computation of the frequencies and would also perform some sort of ''smoothing'', or interpolation, between the discrete conditional probabilities.

The problem of low counts is less severe with this formalism than with some others, where counts are further subdivided by the sequence separation between the atoms involved (Sippl, 1990; Subramaniam *et al.*, 1996). The total conditional probability score for a given protein conformation represents a sum over a very large number of individual terms (in the order of 10^6 contributions) so that the effect of the uncertainties is reduced.

Linear interpolation of the conditional probabilities in the RAPDF does result in better discrimination for these decoy sets (Figure 7). This suggests that the discrete points for each distance bin can be represented by a continuous function and so used with a gradient-dependent protein folding simulation technique.

Effect of aritfacts in the decoy sets

A discriminatory function may be able to select the correct conformation by utilizing subtle differences in the incorrect conformations which distinguish them from the experimentally determined structures. For example, refinement of structures determined using X-ray crystallography is usually done with programs (like X-PLOR (Brünger, 1992)) which effectively restrain particular distances for atom–atom interactions, such as hydrogen bonds. Since the PDFs are parameterized on high resolution X-ray crystallography structures, it is important to demonstrate that discrimination of correct conformations from incorrect ones is not based on this kind of fine detail.

For the LOOP decoy set (Figures 3 and 4) this is not the case, as the criteria for correctness depend on selecting a low rmsd conformation to the experimental structure, not the experimental structure itself. For the crystallographic error (PDBERR) decoy set, discrimination based on fine detail is unlikely, as both the correct and incorrect conformations were refined using X-PLOR (Bruñger, 1992) or PROLSQ (Hendrickson *et al.*, 1985).

To test whether such an artifact is responsible for accurate discrimination in the CASP1 decoy set, consisting of homology models and their corresponding experimental structures, we energy minimized two models together with the experimental structures using the DISCOVER CVFF forcefield (Hagler *et al.*, 1974; Dauber-Osguthorpe *et al.*, 1988). The discrimination ratios of the negative log conditional probabilities between the unminimized model and the unminimized experimental structure are 0.82 and 0.87 for the two cases The corresponding ratios of the minimized model to the minimized experimental structure are 0.78 and 0.83. Thus, surprisingly, discrimination is slightly improved by this procedure, rather than deteriorating as would be the case if the signal were arising from fine scale differences. While this is not an exhaustive test, it suggests that the signal that separates correct and incorrect conformations is not due to fine details in the experimental structures.

A similar test was performed for the SGPA decoy set, in this case, minimizing all structure using 1000 steps of steepest descent using the CHARMM forcefield (Brooks *et al.*, 1983). The discrimination ratios between the simulation structures and the unminimized experimental structure are 0.72 and 0.74 for the two cases. The corresponding ratios of the minimized simulation structures to the minimized experimental structure are 0.79 and 0.80.

In the MISFOLD decoy set, the main chains of both the correct and incorrect conformations are from structures determined using X-ray crystallography. However, the percentages of structures correctly discriminated using non-specific compactness, measured by the CDF (Figure 1a) indicates that the side-chain packing in the incorrect conformations does not generate as many atom–atom contacts within 6.0 Å as would normally be observed in an experimental conformation for that sequence. So for these decoys the signal may be partially due to fine differences between correct and incorrect conformations.

Limits on the discriminatory power

There are definite limits as to what the discriminatory functions can achieve. The weakest performance is for the IFU decoy set, where the percentage discrimination by the RAPDF is 70%. The independent folding units in this set are short fragments (between 10 and 20 residues) that are removed from the context of the rest of the structure. In some cases, the failure may be because these fragments do not represent *bona fide* independent folding units. In others, the information available in the fragments may not be adequate to uniquely identify the correct conformation from the incorrect conformation, as the scores are evaluated on relatively small numbers of atom pairs.

Discrimination is not attained for one member of the LOOP decoy set (residues 265 to 269 in the immunoglobulin A FAB fragment; PDB code 2fbj). This region is involved in several (>10) intermolecular crystallographic contacts of less than 4.0 Å in the experimental structure. However, the CDF (Figure 3) and other discriminatory functions (Braxenthaler *et al.*, 1997) are able to discriminate well.

In the CASP1 decoy set (Figure 2), the RAPDF with one protein, nucleoside diphosphate kinase

(nm23), fails to discriminate between the correct and approximate conformations, and with another protein (hpr) performs worse than RVPDF and NVPDF. In both these cases, the approximate conformations are very similar (within 0.53 and 1.05 Å C_α rmsd, respectively) to the experimental structure. This suggests that the discriminatory function may be unable to consistently discriminate correct from incorrect when the conformations are close (around 1.0 Å C_α rmsd) to the experimental conformation. Present comparative modelling techniques do not approach this level of accuracy except at very high sequence identity (Martin *et al*., 1997), so meanwhile this is not a concern.

Effect of experimental accuracy

An alternative explanation for poor discrimination with nm23 is the relatively low quality of the experimental structure (2.8 Å resolution with an *R*-factor of 0.25 (Webb *et al*., 1995)). The parent structure used for the comparative modelling, PDB code 1ndl, has a high sequence identity (77%), and has been solved to a 2.4 Å resolution with an *R*-factor of 0.16 (Chiadmi *et al*., 1993). Thus the poor discrimination of the RAPDF may simply reflect the moderate resolution and incomplete refinement of the target experimental structure relative to the parent structure.

Relevance of conditional probabilities to the nature of physical interactions observed in proteins

The discriminatory function was compiled using statistical observations, averaging over different environments in protein structures. As a result, it displays some features not observed in a direct way in a physics-based energy function, as shown in Figures 9, 10, 11, and 12.

Caution should be used when interpreting the conditional probability or potential of mean force data in physical terms because of the averaging of environments that occurs during the compilation of the probabilities. To illustrate with an extreme example, the minimum in the N–O plot (Figure 11) in the 0.0 to 3.0 Å bin averages over hydrogen bonds between *N* and O atoms in i, $i+4$ residues in α-helices and between N–O distances in i, $i+1$ (neighbouring) residues. It is thus difficult to ascertain exactly where the signal is coming from given the two different environments.

Nevertheless, many of these features seen are physically significant. Some are obvious: for example, the largest minima in the plot of C_α–C_α contacts (Figure 9) reflects the geometrical constraints imposed by the covalent structure of the polypeptide chain, and the low score for contacts between proline nitrogen and other main-chain oxygen reflects the absence of the hydrogen atom in proline nitrogen (Figure 11).

Some of the features are less obvious: the smallest negative log conditional probabilities for nitrogen–oxygen contacts are in the 0.0 to 3.0 Å bin for aspartate and lysine, reflecting the tendency of these residues to form salt bridges in α-helices and β-sheets, resulting in main-chain hydrogen bonds between them (Figures 11 and 12).

Availability of the PDFs

The Tables containing the parameters for the PDFs, the decoy sets, and software to handle the data are available from the PROSTAR Web site (Braxenthaler *et al*., 1997).

Acknowledgements

Thanks to Jan Pedersen, Brett Milash, Michael Braxenthaler, and Rui Luo for valuable discussions. We also thank Lisa Holm and Chris Sander for making available the co-ordinates for the MISFOLD decoys. This work was supported in part by a Life Technologies Fellowship to Ram Samudrala and NIH grant GM41034 to John Moult.

References

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science,* **181**, 223–230.

Avbelj, F. & Moult, J. (1995a). Determination of the conformation of folding initiation sites in proteins by computer simulations. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Avbelj, F. & Moult, J. (1995b). Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry,* **34**, 755–764.

Avbelj, F., Moult, J., Kitson, H., James, M. & Hagler, A. (1990). Molecular dynamics study of the structure and dynamics of a protein molecule in crystalline ionic environment, *Streptomyces griseus* Protease A. *Biochemistry,* **29**, 8658–8676.

Bahar, I. & Jernigan, R. (1997). Inter-residue potentials in globular proteins: dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266**, 185–214.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tsumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Beveridge, D. & DiCapua, F. (1989). *Computer Simulations of Biomolecular Systems. Theoretical and Experimental Applications* (Gunsteren, W. & Weiner, P., eds), Escom, Leiden, The Netherlands.

Bowie, J., Lüthy, R. & Eisenberg, D. (1991). Method to identify protein sequences that fold into known three-dimensional structure. *Science,* **253**, 164–170.

Braxenthaler, M., Samudrala, R., Pedersen, J., Luo, R., Milash, B. & Moult, J. (1997). PROSTAR: The protein potential test site. <http://prostar.carb.nist.gov>.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.

Brünger, A. (1992). *X-PLOR Version 3.1: A System for X-ray Crystallography and NMR,* Yale University Press.

Bryant, S. & Lawrence, C. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Chiadmi, M., Morera, S., Lascu, I., Dumas, C., Le Bras, G., Veron, M. & Janin, J. (1993). Crystal structure of the Awd nucleotide diphosphate kinase from *Drosophila*. *Structure,* **1**, 283–293.

Dauber-Osguthorpe, P., Roberts, V., Osguthorpe, D., Wolff, J., Genest, M. & Hagler, A. (1988). Structure and energetics of ligand binding to proteins: *E.coli* dihydrofolate reductase-trimethoprime, a drug receptor system. *Proteins: Struct. Funct. Genet.* **4**, 31–47.

DeBolt, S. & Skolnick, J. (1996). Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of proteins structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng.* **9**, 637–655.

Dill, K. (1990). Dominant forces in protein folding. *Biochemistry,* **29**, 7133–7155.

Fidelis, K., Stern, P., Bacon, D. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7**, 953–960.

Fischer, D. (1997). UCLA-DOE Fold Recognition Server. <http://www.mbi.ucla.edu:-88/>.

Fischer, D. & Eisenberg, D. (1996). Fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.

Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. & Sippl, M. (1995). Progress in fold recognition. *Proteins: Struct. Funct. Genet.* **23**, 376–386.

Hagler, A., Huler, E. & Lifson, S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Amer. Chem. Soc.* **96**, 5319–5335.

Halgren, T. (1995). Potential energy functions. *Curr. Opin. Struct. Biol.* **5**, 205–210.

Head-Gordon, T. & Brooks, C. (1991). Virtual rigid body dynamics. *Biopolymers,* **31**, 77–100.

Hendrickson, W., Smith, J. & Sheriff, S. (1985). Direct phase determination based on anomalous scattering. *Methods Enzymol.* **115**, 41–55.

Holm, L. & Sander, C. (1992a). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.

Holm, L. & Sander, C. (1992b). Fast and simple monte carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins: Struct. Funct. Genet.* **14**, 213–223.

Huang, E., Subbiah, S. & Levitt, M. (1995). Recognising native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709–720.

Jernigan, R. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 192–209.

Jones, D., Miller, R. & Thornton, J. (1995). Successful protein folding recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387–397.

Jones, D., Taylor, W. & Thornton, J. (1992). A new approach to protein fold recognition. *Nature,* **258**, 86–89.

Jorgensen, W. & Tirado-Rives, J. (1988). The OPLS potential function for proteins. Energy minimisations for crystals of cyclic peptides and crambin. *J. Amer. Chem. Soc.* **110**, 1657–1666.

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

Levitt, M. (1998). Competitive assessment of protein folding recognition and threading accuracy. *Proteins: Struct. Funct. Genet.* In the press.

Lüthy, R., Bowie, J. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature (London),* **356**, 83–85.

MacArthur, M., Laskowski, R. & Thornton, J. (1994). Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. *Curr. Opin. Struct. Biol.* **4**, 731–737.

Madej, T., Gibrat, J. & Bryant, S. (1995). Threading a database of protein cores. *Proteins: Struct. Funct. Genet.* **23**, 356–369.

Mark, A. & van Gunsteren, W. (1994). Decomposition of the free energy of a system in terms of specific interactions. *J. Mol. Biol.* **240**, 167–176.

Martin, A., MacArthur, M. & Thornton, J. (1998). Assessment of comparative modelling in CASP2. *Proteins: Struct. Funct. Genet.* In the press.

McCammon, J. & Harvey, S. (1987). *Dynamics of Proteins and Nucleic Acids,* Cambridge University Press.

McQuarrie, D. (1976). *Statistical Mechanics,* Harper and Row, NY.

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

Mosteller, F., Rourke, R. & Thomas, G., Jr. (1970). *Probability with Statistical Applications,* Addison-Wesley Publishing Company, Reading, MA.

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7**, 194–199.

Moult, J. & James, M. N. G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Struct. Funct. Genet.* **2**, 146–163.

Orengo, C., Michie, A., Jones, S., Swindells, M., Jones, D. & Thornton, J. (1993). Protein structure classification. <http://www.biochem.ucl.ac.uk/bsm/cath/>.

Park, B. & Levitt, M. (1992). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392.

Pedersen, J. T. & Moult, J. (1997). Folding simulation with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **269**, 240–259.

Samudrala, R. & Moult, J. (1998). Handling context-sensitivity in protein structures using graph theory: bona fide prediction. *Proteins: Struct. Funct. Genet.* In the press.

Simons, K., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

Sippl, M. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct. Funct. Genet.* **17**, 355–362.

Sippl, M. (1995). Knowledge based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.

Sippl, M., Ortner, M., Jaritz, M., Lackner, P. & Flöckner, H. (1996). Helmholtz free energies at atom pair interactions in proteins. *Folding and Design,* **1**, 289–298.

Storch, E. & Daggett, V. (1995). Molecular dynamics of cytochrome b5: implications for protein–protein recognition. *Biochemistry,* **34**, 9682–9693.

Subramaniam, S., Tcheng, D. K. & Fenton, J. (1996). A knowledge-based method for protein structure refinement and prediction. In *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology* (States, D., Agarwal, P., Gaasterland, T., Hunter, L. & Smith, R., eds), pp. 218–229, AAAI Press, Boston, MA.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.*, 762–785.

Unger, R. & Moult, J. (1991). An analysis of protein folding pathways. *Biochemistry,* **30**, 3816–3823.

Wang, Y., Zhang, H. & Scott, R. (1995). Discriminating compact non-native structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA,* **92**, 709–713.

Webb, P., Perisic, O., Mendola, C., Backer, J. & Williams, R. (1995). The crystal structure of a human nucleoside diphosphate kinase, NM23-H2. *J. Mol. Biol.* **251**, 574–587.

Weiner, S., Kollman, P., Nguyen, D. & Case, D. (1986). An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.* **7**, 230–252.

Wodak, S. & Rooman, M. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247–259.

Edited by F. Cohen

(Received 10 July 1997; received in revised form 21 October 1997; accepted 21 October 1997)