Protein Engineering, Design & Selection vol. 19 no. 9 pp. 431–437, 2006
Published online July 14, 2006 doi:10.1093/protein/gzl027

The effect of experimental resolution on the performance of knowledge-based discriminatory functions for protein structure selection

Tianyun Liu and Ram Samudrala[1]

Department of Microbiology, University of Washington, School of Medicine, Seattle, WA 98195, USA

[1]To whom correspondence should be addressed. E-mail: ram@compbio.washington.edu

The key to an accurate method of protein structure prediction is the development of an effective discriminatory function. Knowledge-based discriminatory functions extract parameters from statistical analysis of experimentally determined protein structures. We assess how the quality of the protein structures used for compiling statistics affects the performance of a residue-specific all-atom probability discriminatory function (RAPDF). We find that the discriminatory power correlates with the quality of the structural dataset on which the RAPDF is parameterized in a statistically significant manner. The overrepresentation of unfavorable contacts in the low-resolution and NMR structures contributes to the major errors in the compilation of the conditional probabilities. Such errors weaken the discriminatory power of the function, especially when decoy conformations also contain considerable numbers of unfavorable contacts. This indicates that using high-resolution structural datasets after filtering out unfavorable contacts can improve the performance of knowledge-based discriminatory functions.

Keywords: all-atom probability discriminatory function/ experimental resolution/decoy discrimination

Introduction

A large number of protein structure prediction methods, including those based on comparative modeling, fold recognition and *de novo* simulations, rely on an effective knowledge-based discriminatory function (Jernigan and Bahar, 1996; Moult, 1997; Lazaridis and Karplus, 2000). Knowledge-based discriminatory functions extract statistics from a database of experimentally determined protein structures. Often the distribution of pairwise distances is used to extract 'pseudo-potentials' between atomic interactions based on the inverse formulation of the Boltzmann equation. Alternatively, a knowledge-based discriminatory function may be viewed as a set of probability distributions that can be used to find the most native-like structure (Sippl, 1990, 1995). Basic physical principles, however, are often violated in both these formalisms (Godzik *et al*., 1995; Godzk, 1996; Thomas and Dill, 1996; Ben-Naim, 1997; Park *et al*., 1997). For example, the original derivation of the Boltzmann distribution is based on a thermodynamic equilibrium system, whereas a database of protein structures is an inhomogeneous collection of different systems, each with its own free energy minimum (Godzik *et al*., 1995; Jernigan and Bahar, 1996; Ben-Naim, 1997). This theoretical defect, together with other problems in the theoretical justification of the knowledge-based discriminatory function (Godzik *et al*., 1995; Godzk, 1996; Thomas and Dill, 1996; Ben-Naim, 1997; Park *et al*., 1997), causes uncertainties when selecting experimental structures for a database to compile the knowledge-based discriminatory function.

The properties of a structural database affect the statistical outcome derived from it. The database dependence of the knowledge-based discriminatory functions has been reported previously by Furuichi and Koehl (1998). Their study showed that knowledge-based discriminatory functions carry a memory of the quality of the database in terms of the amount and diversity of secondary structure it contains. For example, the distance-dependent discriminatory function extracted by the method of Sippl from an all-α protein structural database is quantitatively different from that extracted from an all-β protein structural database. In a recent study, Zhang *et al*. (2004) compared the database dependence on structure topology between three different knowledge-based approaches. They have suggested that a possible source for database dependence is the flawed reference state used in the knowledge-based approaches.

Databases of protein structures are growing in size. Knowledge-based discriminatory functions derived from them should, therefore, also increase in accuracy. However, because the theoretical basis of knowledge-based discriminatory function is not clear (Godzik *et al*., 1995; Godzk, 1996; Thomas and Dill, 1996; Ben-Naim, 1997; Park *et al*., 1997), a definitive 'rule of thumb' for selecting experimentally determined structures for a database has never been proposed. Corresponding to this issue, there is now an enormous difference in accuracy between the best and the worst experimentally determined protein structures owing to the limitations of methodologies and experimental errors (Cruickshank, 1999).

It has been considered that X-ray diffraction has a relatively high degree of inherent reliability. However, there are many minor inaccuracies or problems of interpretation that can affect reliability of the final coordinates (Laskowski *et al*., 1998). For example, if the data is poor and the quality of the electron density map is low, it can be difficult to trace a molecule using the electron density computed from the diffraction data. In Nuclear Magnetic Resonance (NMR) spectroscopy, insufficient experimentally derived restraints often result in the uncertainty of the atomic coordinates (Chalaoux *et al*., 1999). In addition, the problem of valid error estimation has not yet been solved, mainly because it is difficult to estimate the likelihood of occasional large mistakes in assigning starting coordinates that might not be correctable by refinement. Therefore, not all structures deposited in the Protein Data Bank (PDB) are of equally high quality, usually because of the quality of the experimental data from which they were determined. Software tools for validating protein structures

have been developed, which can detect some errors in the assessed structures. For example, the PROSA II program by Sippl can identify misfolded structures as well as faulty segments of structural models by calculating energy distributions of residues with statistical potentials of mean force (Sippl, 1993). Other tools include PROCHECK (EU 3-D Validation Network, 1998) and ERRAT (Colovos and Yeates, 1993).

We previously developed a residue-specific all-atom conditional probability discriminatory function (RAPDF) (Samudrala and Moult, 1998) that includes all protein heavy atoms and residue-specific atom types, rather than using a reduced set of atom types. The effectiveness of RAPDF has been demonstrated in a number of studies, including the selection of conformations in comparative modeling and evaluation of decoys in *de novo* simulations (Samudrala *et al*., 1999a, b). In this study, we investigate whether the quality of structures used for compilation will affect the performance of knowledge-based discriminatory functions with respect to discriminating near-native conformations from non-native ones. We accomplish this by comparing the effectiveness of RAPDFs derived from structural datasets with different experimental resolutions. For each RAPDF, we evaluate its ability to recognize and rank near-native conformations through a comprehensive 'decoy discrimination' test (Samudrala and Levitt, 2000; Tsai *et al*., 2003; Wang *et al*., 2004). We thus arrive at practical rules for selecting experimental structures for the compilation of conditional probabilities. In addition, we examine the conditional probabilities derived from the low-resolution and NMR datasets that lead to inaccurate discrimination. We further discuss the relationship between the quality of structures, the distance cutoff of interatomic contacts used for compilation and the efficacy of the knowledge-based discriminatory functions.

Methods

RAPDF

A complete description of RAPDF can be found in the original paper (Samudrala and Moult, 1998). In summary, we make observations of interatomic distances on a dataset of experimentally determined structures. The conditional probabilities are compiled by counting frequencies of distances between pairs of atom types in a dataset of protein structures. All non-hydrogen atoms are considered, and a residue-specific description of the atoms was used, that is, the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. The interatomic distances observed are divided into 1.0 Å bins ranging from 3.0 to 20.0 Å. Contacts between atom types in the 0–3 Å range are placed in a separate bin, resulting in a total of 18 distance bins. Distances within a single residue are not included in the counts.

The scores $S(d_{ab})$ proportional to the negative log conditional probability of observing a native conformation given an interatomic distance are compiled according to the formula:

$$S(d_{ab}) = -\ln\frac{P(d_{ab} \mid C)}{P(d_{ab})} \propto -\ln P(d_{ab} \mid C).$$

Here $P(d_{ab} \mid C)$ is the probability of observing a distance d between two atom types a and b in a correct conformation, and $P(d_{ab})$ is the probability of observing such a distance, d_{ab}, in any conformation, correct or incorrect. For a dataset of experimental structures, the counts of observations of d_{ab} in each structure are summed to generate an overall probability. We compiled tables of scores $S(d_{ab})$ for all possible pairs of the 167 atom types for the 18 distance ranges from a database of known structures.

Given an amino acid sequence in a particular conformation, the scores of all contacts between pairs of atom type that fall within the distance cutoff is summed to yield the total RAPDF score to evaluate the probability of a conformation being native-like.

Conformation files used for compilation of conditional probabilities

To build structural datasets for the compilation of conditional probabilities, a non-homologous subset was taken from the ASTRAL 1.69 database (Chandonia *et al*., 2004). A representative subset may be selected according to the similarity measure based on the *E*-value (Murzin *et al*., 1995; Chandonia *et al*., 2004). Specifically, a non-homologous subset containing 5439 structures was initially obtained from the ASTRAL 1.69 database according to the similarity measure with a threshold of 10^{-4} on the *E*-value. Conformations with incomplete side chains and theoretical models were excluded.

The non-homologous X-ray diffraction structures were divided into three datasets according to their resolution. The boundaries were chosen to ensure that the number of structures and the number of residues in each dataset were similar. The resolution ranges of the three X-ray diffraction datasets are: dataset 1, 0.54–1.79 Å (1486 structures); dataset 2, 1.80–2.10 Å (1532 structures); and dataset 3, 2.11–3.90 Å (1518 structures). A total of 616 structures solved by NMR spectroscopy were used as dataset 4. These four structural datasets were used for obtaining the conditional probabilities, resulting in four RAPDFs.

Decoy sets

Publicly available decoy sets provide a means to evaluate the performance of discriminatory functions. A total of eight multiple decoy sets generated by different simulation methods were used to test the performance of the RAPDFs. They include decoy conformations for 181 proteins: `rosetta` set containing decoy conformations for 41 proteins, 4 state_reduced for 7 proteins, `fisa_casp3` for 6, `hg_structal` for 29, `ig_structal` for 61, `ig_structal_hires` for 20, `lmds` for 11 and `semfold` for 6 proteins. The `rosetta` set were obtained from http://www.bakerlab.org (Samudrala and Levitt, 2000). All other decoy sets were obtained from the Decoys 'R' Us database <http://dd.compbio.washington.edu> (Tsai *et al*., 2003).

Evaluation of the discriminatory power of the RAPDFS

There are two ways to evaluate the discriminatory power of a discriminatory function on decoy sets (Samudrala and Levitt, 2000; Tsai *et al*., 2003; Wang *et al*., 2004). The first approach is to measure the likelihood of selecting the native structure from a set of decoys. Within any decoy set, an effective discriminatory function should be able to distinguish the native structure from non-native ones with a high degree of accuracy. However, the native structure can rarely be reproduced exactly.

The ability of picking the best-predicted or the near-native decoys is more important in protein structure prediction. Therefore, our preferred method is to assess how well a particular discriminatory function can distinguish near-native conformations from non-native ones in a particular decoy ensemble (Samudrala and Levitt, 2000; Tsai *et al*., 2003; Wang *et al*., 2004). An effective discriminatory function should show a consistent preference for the former.

Given a set of decoy conformations, we first plot the RAPDF score against the root mean square deviations of the C_α atoms (cRMSD) between the native conformation and each decoy conformation. The cRMSD is not a measure of the resolution of a decoy conformation but reflects its structural similarity to the native. Suppose the best-scoring conformation has the cRMSD rank of R in an ensemble of N decoy conformations, the log probability of selecting the best-scoring conformation (log P_{B1}) is calculated as log P_{B1} = log(R/N). This is the major criteria for evaluating the decoy discrimination of a function.

Other evaluation measures include: (i) The log probability of selecting the lowest cRMSD conformation among the top 10 best-scoring conformations (log P_{B10}), which is calculated as log P_{B0} = log (R_i/N), where R_i is the cRMSD rank of the decoy conformation, which has the lowest cRMSD among the 10 best-scoring decoy conformations. (ii) The fraction enrichment (FE) of the 10% lowest cRMSD conformations in the top 10% best-scoring conformations. (iii) Correlation coefficient (CC) between RAPDF scores and cRMSDs within a set of decoy conformations.

The discriminatory power of each of the four RAPDFs parameterized on different datasets was evaluated by log P_{B1} on the decoy conformations for 181 proteins from the eight multiple decoy sets. For comparison purposes, the value for each protein in the same decoy set was summed, resulting in a sum of log P_{B1} for the decoy set. To perform an evaluation on all the eight decoy sets, the overall sum of log P_{B1} on the eight decoy sets was calculated. The overall evaluation of the sum of, the average CC and the average FE were calculated in the same manner.

Determining the source of errors in the compilation of conditional probabilities

For each structural dataset, a table containing the scores $S(d_{ab})$ for all pairs between the 167 atom types in the 18 distance bins was compiled (Samudrala and Moult, 1998). We compared the four sets of scores compiled from structural datasets with different experimental resolutions. Variances between the four equivalent scores were calculated and plotted against the corresponding atom types and the distance bin indexes. Representative residue-specific interatomic contacts contributing to the significant differences between the four sets of scores were analyzed. The residue-specific atom type is named using the following convention: one letter abbreviation of the residue followed by one or more letters representing the atom type. For example, VCG1 represents the Cγ1 atom in valine.

To explain how errors in the conditional probabilities of RAPDF weaken decoy discriminations, we compared decoy conformations to the native one in the decoy set `4state_reduced/1ctf`. Images of corresponding conformations were prepared using Molscript (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997).

Results and discussion

The discriminatory power of RAPDF correlates with the quality of the structural dataset used for the compilation of conditional probabilities

Our goal is to assess the relationship between the experimental quality of the dataset and the discriminatory power of RAPDF. Four experimental datasets were derived from the ASTRAL database containing protein structures with different qualities. Datasets 1 to 3 were derived from X-ray diffraction structures indexed from the highest resolution to the lowest resolution and dataset 4 contains only structures solved using NMR. The performances of the four RAPDFs derived from these four datasets were evaluated by the log P_{B1}, log P_{B10}, FE and CC evaluation measures. Each evaluation measure quantifies the efficacy of RAPDFs at discriminating the eight decoy sets generated by different simulation methods.

Figure 1 shows the overall evaluation of RAPDFs on all decoy sets. The log P_{B1} estimates the likelihood of selecting a conformation of a particular cRMSD with the best score. The smaller the value, the greater the likelihood of assigning the best score to the structure with the lowest cRMSD. The overall evaluation by log P_{B1} suggests that the performance of RAPDF correlates with the quality of the dataset from which the RAPDF is derived. When the scores of the near-native decoy conformations are very close to each other, the evaluation by log P_{B10} is more effective. The overall sum of log P_{B10} indicates lower discrimination when the quality of the dataset is lower. An ideal discriminatory function has scores that are perfectly correlated with cRMSDs, allowing straightforward detection of the best-predicted conformations. The higher the CC, the better the discriminatory function at selecting near-native conformations. The overall average of CC increases consistently when using RAPDFs derived from the structural datasets of higher resolution. The FE captures the extent to

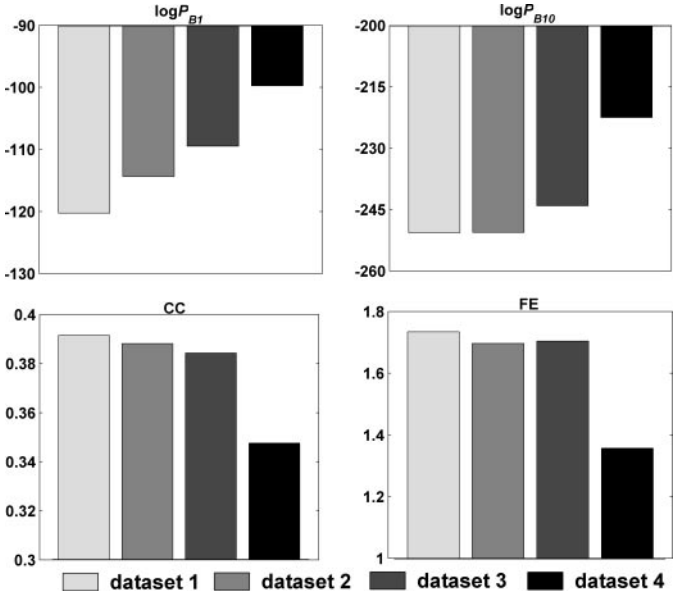


Fig. 1. Relationship between the experimental quality of the dataset and the discriminatory power of RAPDF as evaluated by the overall log P_{B1}, log P_{B10}, fraction enrichment (FE) and correlation coefficient (CC) for the 181 proteins from the eight decoy sets. The overall evaluations by the four measures suggest that the discriminatory power of RAPDF is enhanced as the resolution of X-ray diffraction dataset is improved. The lowest discriminatory power of RAPDF parameterized on the NMR dataset.

which the lowest cRMSD conformations are enriched by the subset of the best-scoring conformations. The overall FE evaluation indicates that RAPDF derived from the structural datasets of higher resolution has improved power of enriching the best-predicted conformations.

The non-parametric Wilcoxon Signed-Rank test was employed to estimate the statistical significance of the difference between the performances of RAPDFs derived from different datasets. For every two datasets, we hypothesize that the RAPDF compiled from the lower-quality dataset performs the same or better than the RAPDF compiled from the higher-quality dataset. For the evaluation by log P_{B1}, the P-value calculated using the Wilcoxon test between dataset 1 and that of dataset 2 is 0.0195, that between dataset 2 and dataset 3 is 0.0039 and that between dataset 3 and dataset 4 is 0.0039, well beyond the baseline significance of 0.05. Similar results were obtained using log P_{B1}, FE, and CC evaluation measures, indicating that the differences between the higher-resolution RAPDFs and the lower-resolution ones are statistically significant (P-values < 0.05).

In summary, the effectiveness of RAPDF correlates with the resolution of the three X-ray diffraction structural datasets from which the RAPDF is compiled in a statistically significant manner. The RAPDF derived from the NMR dataset shows the lowest discriminatory power. Regardless of the evaluation methods used, the discriminatory power of the knowledge-based function is improved by using a high-resolution dataset of experimentally determined structures.

The errors in conditional probabilities originate from the high frequencies of unfavorable contacts in low-resolution structures

The conditional probabilities derived from the four structural datasets were compared to determine the reason for the different discrimination occurrences of the RAPDFs. For each structural dataset, a table is compiled containing the scores $S(d_{ab})$ for all pairs between the 167 atom types in the 18 distance bins (ranging from 0 to 20 Å). Variances between the four equivalent scores for each pair were calculated and plotted against the corresponding atom types and the distance bin indexes. The largest variances were observed in the distance bins 1 for which the distance range is 0–3 Å (Figure 2). This indicates that the probabilities for atom pairs at close distances may contribute most to the errors in the conditional probabilities for RAPDF.

We further examined the specific contacts that may cause errors in the conditional probabilities compiled from the low-resolution and NMR datasets (datasets 3 and 4). All the 167 × 167 residue-specific interatomic contacts observed in the 18 distance bins were sorted by the difference between the score $S(d_{ab})$ compiled from the high-resolution X-ray diffraction structures (dataset 1) and the equivalent score compiled from the NMR structures (dataset 4). All interatomic contacts with the top 100 largest differences were found in distance bins 1 (0–3 Å), which is consistent with the observations in Figure 2. The scores of these contacts show a consistently decreasing trend along the dataset indices. This indicates a higher frequency of such contacts occurring in structural datasets with a lower average resolution, which is consistent with the fact that lower-resolution structures contain more inaccurate interatomic contacts (Kraulis, 1991; Murzin *et al*., 1995). In this study, these

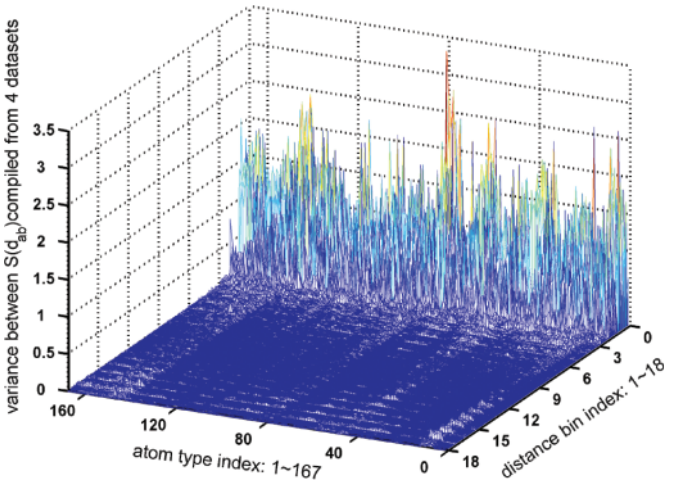


Fig. 2. Variance between four sets of log odds scores [$S(d_{ab})$] derived from four different datasets. Variances between the four equivalent $S(d_{ab})$ for each interatomic contact observed in each distance bin are calculated and plotted against the 167 residue-specific atom types and the 18 distance bins. The two horizontal axes represent the 167 residue-specific atom types and the 18 distance bins, respectively. The vertical axis indicates the variance between equivalent scores compiled from four different datasets. The largest variances between the four sets of scores are observed in distance bins 1(0–3 Å).

contacts, typically observed in the distance bins, are referred to as 'unfavorable contacts'.

The scores of 10 representative unfavorable contacts observed in the distance bin 1 (0–3 Å) are shown in Figure 4. They are VO-LO, VO-LCG, KO-VO, AO-LCB, TO-WCG, ACB-MO, ACB-RO, AO-DO, AO-VCG1 and VO-VCG2. Most of these interatomic pairs consist of one backbone oxygen with a steric clash to another atom, resulting in bad contacts in X-ray diffraction structures (Laskowski *et al*., 1998). However, given a high-resolution electron density map, these contacts could be removed during refinement. That is, unfavorable contacts are less frequent in high-resolution structures. In agreement with this, the scores for these unfavorable contacts compiled from the low-resolution dataset are lower than those from the high-resolution dataset, indicating that these contacts occur more frequently in the low-resolution structures (Figure 3).

Overall, the high frequencies of unfavorable contacts in protein structures reflect that the accuracy of the corresponding atom coordinates is not reliable. Overrepresentation of unfavorable contacts consequently results in errors in the conditional probabilities.

Unfavorable contacts in decoy conformations diminish the effectiveness of RAPDFs compiled from low-resolution or NMR structures

To explain how the overrepresentation of the unfavorable contacts in the low-resolution and NMR datasets results in the lower discrimination, decoy conformations in a specific decoy set were scrutinized. We asked the question: if a decoy also contains considerable numbers of unfavorable contacts, can an RAPDF compiled from any of the four structural datasets distinguish it effectively?

First, two decoy conformations from the decoy set `4state_reduced/1ctf` were compared with the native conformation `1ctf` (Figure 4A). The cRMSD of the two decoy conformations `1ctf.d9493` and `1ctf.g4353` are 0.8 and 5.3 Å, respectively. The RAPDF scores are compared

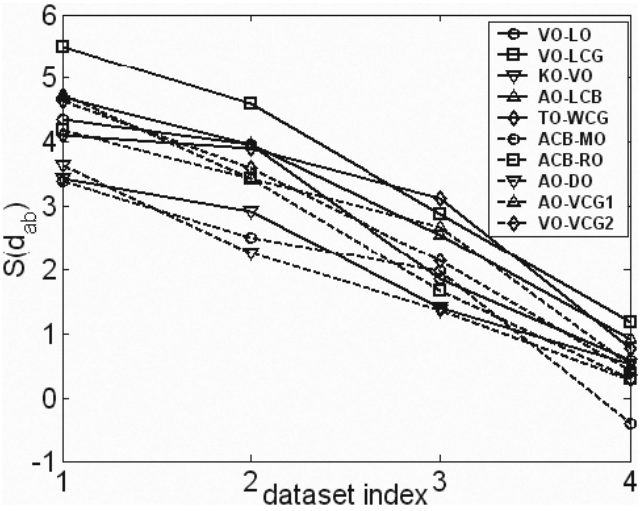
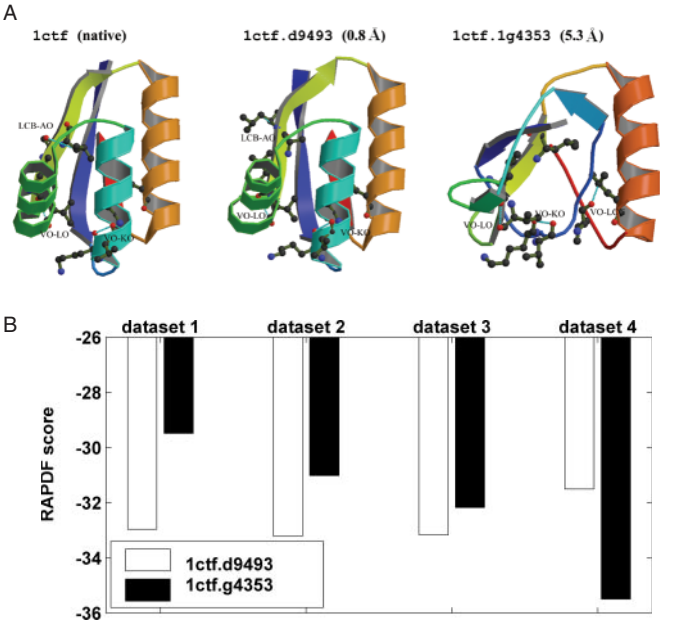


Fig. 3. The scores $S(d_{ab})$ of 10 representative unfavorable contacts observed in distance bin 1 (0–3 Å). For each of these contacts, the four scores consistently decrease as the dataset index increases, indicating that higher frequencies of these contacts occur in the lower-resolution and NMR datasets. All these unfavorable contacts contain a clashing backbone oxygen atom.



C

Interatomic contacts	Native structure	1ctf.d9493 cRMSD: 0.8 Å	1ctf.g4353 cRMSD: 5.3 Å
34LO-35VO	3.43	3.64	2.66
16VO-58LCG	6.06	5.39	2.98
13KO-14VO	3.23	3.53	2.87
41AO-42LCB	4.13	4.24	2.93

Fig. 4. Analysis of specific decoys and contacts in the 4state_reduced/1ctf set. (**A**) All conformations are colored from the N terminus (blue) to the C terminus (red). Four interatomic contacts are represented as ball-and-sticks, with the connection between the two paired atoms colored in cyan. The pattern of these contacts in the near-native decoy (1ctf.d9493) is similar to that observed in the native structure. The decoy conformation with high cRMSD (1ctf.g4353) has a significantly different distance pattern for these atoms. (**B**) RAPDF scores of 1ctf.d9493 and 1ctf.g4353 calculated using RAPDF derived from the four datasets are compared. The RAPDF score of 1ctf.g4353 reaches to a level closer to that of 1ctf.d9493 along the increasing dataset index. Using RAPDFs compiled from high-resolution datasets (dataset 1 and 2), these two decoys could be discriminated correctly. The RAPDF derived from NMR dataset (dataset 4) cannot discriminate between the near-native (1ctf.d9493) and non-native (1ctf.g4353) decoy. (**C**) Distances (Å) of four interatomic contacts containing backbone oxygens. Three of these contacts in the near-native decoy (1ctf.d9493) are observed in the same distance bin with equivalent contacts in the native conformation whereas those in the non-native decoy (1ctf.g4353) are not, resulting in different contributions to the final RAPDF scores for their respective decoy conformations. These contacts in 1ctf.g4353 are observed in distance bin 1 (0–3 Å) and thus represent unfavorable contacts.

in Figure 4B. The lower the RAPDF score, the higher the probability of a decoy conformation being native-like. Using RAPDFs compiled from high-resolution datasets (dataset 1 and 2), these two decoys could be discriminated correctly. The RAPDF score of 1ctf.g4353 reaches a level closer to that of the near-native conformation along the increasing dataset index. The RAPDF derived from the NMR dataset (dataset 4) could not discriminate these two decoys: the decoy 1ctf.g4353 with higher cRMSD has a better RAPDF score (−35.50) than the near-native decoy 1ctf.d9493 (−31.51) (Figure 4B).

To explain such a phenomenon, the distances of the interatomic contacts containing backbone oxygen in the native and the decoy conformations were then inspected. The distances of four such contacts, 34LO-35VO, 16VO-58LCG, 13KO-14VO and 41AO-42LCB, are shown in Figure 4C. These contacts in 1ctf.g4353, however, are observed in the distance bin 1 (0–3 Å) and represent unfavorable contacts. In contrast, equivalent contacts of the near-native decoy 1ctf.d9493 are observed within the acceptable distance range of 3–7 Å.

For each decoy conformation, the RAPDF score is obtained by summing up the scores of all the individual interatomic contacts that fall within a certain distance cutoff. As shown in Figure 3, scores of unfavorable contacts compiled from the low-resolution and the NMR structural datasets are lower because these contacts are overrepresented in those datasets. The contributions of these contacts to the final score of a decoy conformation are, therefore, enhanced, resulting in a more negative RAPDF score, thereby diminishing the effectiveness of the discriminatory function.

Practical rules for selecting experimentally determined structures for derivation of knowledge-based discriminatory functions

Our study points out the limitations of using protein structures uncritically for derivation of knowledge-based discriminatory functions. The discriminatory power of RAPDF is reduced as the resolution of the X-ray diffraction structural datasets decreases. The lower discriminatory power is caused by the overrepresentation of the unfavorable contacts.

To make use of experimentally determined structures for compiling knowledge-based discriminatory functions, we suggest two practical rules: First, the experimental resolution is a good measure of the quality of a structural dataset for extracting conditional probabilities. Second, eliminating unfavorable contacts reduces noise in the compilation of the conditional probabilities.

Most unfavorable contacts are observed as close carbon atom contacts within 0–3 Å (Figure 2). Ideally, if all unfavorable contacts could be distinguished from the close contacts

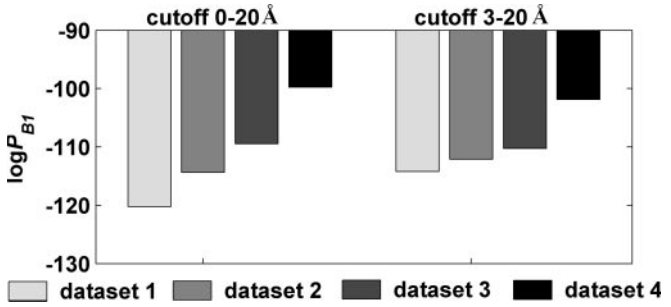


Fig. 5. Comparison of the influence of dataset quality on RAPDFs with distance cutoffs of 0–20 Å and 3–20 Å. The performance of RAPDF is evaluated by $S(d_{ab})$ over eight different decoy sets. The influence of the dataset quality is diminished when close contacts within 3Å are filtered out. For RAPDFs derived from high-resolution structures (dataset 1 and dataset 2), the discriminatory power decreases. The RAPDFs derived from the low-resolution structures (dataset 3) and NMR structures (dataset 2) show improved discrimination.

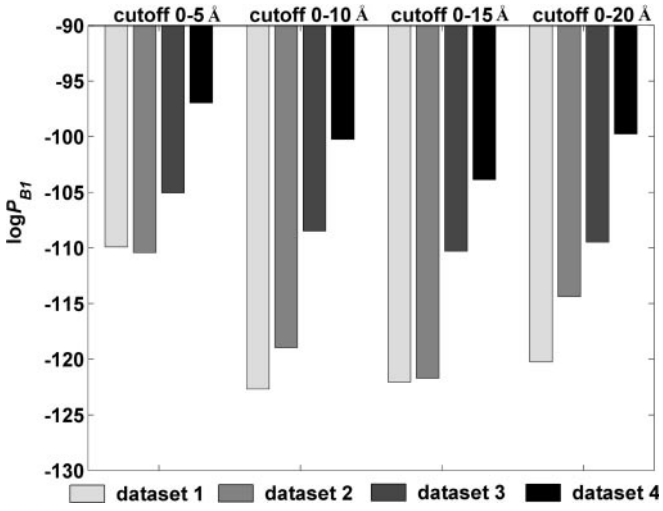


Fig. 6. Performance of RAPDF at different distance cutoffs as evaluated by log P_{B1} over eight different decoy sets. Generally, discrimination progressively improves with larger distance cutoffs, up to 15 Å. The lowest discrimination is always observed in the RAPDFs with a distance cutoff of 5 Å, regardless of the dataset that the RAPDF is parameterized on. The RAPDFs derived from high-resolution X-ray diffraction structures (dataset 1) achieve similar discrimination at cutoffs of 10 Å, 15 Å and 20 Å.

then the efficacy of the function could be improved. Figure 5 shows that the influence of the dataset quality is diminished when all the contacts within 3 Å are excluded for RAPDF compilation. However, for RAPDFs derived from high-resolution structures (dataset 1 and dataset 2), the discriminatory power decreases. This result suggests that some close contacts that are not unfavorable contacts are also crucial to the efficacy of the discriminatory function. Interestingly, the RAPDFs derived from the low-resolution structures (dataset 3) and NMR structures (dataset 2) show an improved discrimination, indicating that in low resolution or NMR structures the effect of unfavorable contacts dominates compared with other close contacts. These observations suggest specialized RAPDFs that are specifically designed to work well for decoy discrimination by eliminating consideration of atom pairs that lead to poor discrimination.

Our results also suggest that the RAPDF derived from the NMR structural dataset is not as powerful as those derived from the high-resolution X-ray diffraction structural datasets. An early study by Godzik *et al*. (1995) has suggested large differences between the parameters derived from X-ray diffraction structures and structures obtained by the NMR method. The origin of this difference is not yet understood. In addition, it is not possible to differentiate between reliable and unreliable NMR structures from the information given in the PDB conformation files. Thus, rules for using NMR structures to derive knowledge-based discriminatory functions are not yet available.

We used the thermal factor (called B-factor) as a gauge of the quality of structural datasets. B-factor is inversely proportional to the relative accuracy of a given atom and represents the thermal motions about the equilibrium structure (Bott and Frane, 1990). Any segment with large B-factors indicates more disorder in that region, which is less 'visible' by X-ray diffraction (Bott and Frane, 1990). To evaluate the effect of disordered regions in a protein conformation, atoms for which the B-factors were 2 SD greater than the average were filtered out to reconstruct our structural datasets. Evaluation on the 189 standard decoys shows similar results to those obtained from the original structural datasets, suggesting that excluding atoms with high B-factors does not affect the knowledge-based discriminatory functions.

In addition, we investigated two other functions developed previously, the residue-specific virtual-atom probability discriminatory function (RVPDF) and the non-residue-specific virtual-atom probability discriminatory function (NVPDF) [Samudrala and Moult, 1998]. These functions are also affected by the quality of the experimental datasets in a similar fashion to RAPDF (data not shown). However, the discriminatory power of RVPDF or NVPDF is lower than that of RAPDF across the different datasets.

The advantage of using a larger distance cutoff for distance-dependent knowledge-based discriminatory functions

Unfavorable contacts in an X-ray diffraction structure usually result from the incorrect interpretation of a poor electron density map. These contacts are the major origins of the errors compiled in the conditional probabilities. Most unfavorable contacts are observed as close contacts within 3 Å. For each set of RAPDFs compiled from different datasets, we compared their discriminatory power at distance cutoffs of 5, 10, 15 and 20 Å (Figure 6). Generally, discrimination progressively improves at a larger distance cutoff up to 15 Å. The RAPDFs derived from high-resolution X-ray diffraction structures (dataset 1) show similar discrimination at cutoffs of 10, 15 and 20 Å. The lowest discrimination is always observed in the RAPDFs with a distance cutoff of 5 Å, regardless of the dataset that the RAPDF is parameterized on. This suggests that including long-distance contacts compensates for the errors caused by unfavorable contacts. It also indicates that including long-distance interactions is necessary even while using high-resolution structures for compiling the RAPDF.

Conclusions

The discriminatory power of an RAPDF correlates with the quality of the structural dataset from which the RAPDF is derived. High-resolution structures for compilation of conditional probabilities improve the discriminatory power of RAPDF. In low-quality structures, overrepresentation of unfavorable contacts results in the errors in the conditional

probabilities. Such errors weaken the discriminatory power of the RAPDFs, especially when decoy conformations also contain considerable numbers of unfavorable contacts. It suggests that improving the current knowledge-based discriminatory functions is possible if the low-quality structures in an experimental dataset are filtered out.

The database dependence of a knowledge-based discriminatory function is difficult to avoid because of its theoretical defects. We, therefore, propose two practical rules to construct structural datasets for derivation of effective knowledge-based discriminatory functions. First, the experimental resolution is a good measure of the likely quality of a structural dataset. Second, eliminating unfavorable contacts reduces noise in the compilation of the conditional probabilities.

Current knowledge-based discriminatory functions do not perform adequately in selecting the most near-native conformations from an ensemble of decoys. Thus improvement in accuracy or effectiveness of discriminatory functions, even on a small scale, may contribute to improved structure prediction. The newly parameterized RAPDF on a high-resolution dataset is more effective at selecting near-native structures.

Acknowledgments

This work was supported in part by Searle Scholar Award, NSF grant DBI-0217241, NSF CAREER award and NIH grant GM068152. We thank Michal Guerquin and other members of the Samudrala group for helpful comments.

References

Ben-Naim,A. (1997) *J. Chem. Phys.*, **107**, 3698–3706.
Bott,R., and Frane,J. (1990) *Protein Eng.*, **3**, 649–657.
Chalaoux,F.R., O'Donoghue,S.I. and Nilges,M. (1999) *Proteins*, **34**, 453–463.
Chandonia,J.M., Hon,G., Walker,N.S., Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) *Nucleic Acids Res.*, **32**, 189–192.
Colovos,C., and Yeates,T.O. (1993) *Protein Sci.*, **2**, 1511–1519.
Cruickshank,D.W. (1999) *Acta Crystallogr.*, **D55**, 583–601.
EU3-D Validation Network (1998), *J. Mol. Biol.*, **276**, 417–436.
Furuichi,E., and Koehl,P. (1998) *Proteins*, **31**, 139–149.
Godzik,A., Kolinski,A. and Skolnick,J. (1995) *Protein Sci.*, **4**, 2107–2117.
Godzk,A. (1996) *Structure*, **4**, 363–366.
Jernigan,R.L., and Bahar,I. (1996) *Curr. Opin. Struct. Biol.*, **6**, 195–209.
Kraulis,P. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.
Laskowski,R.A., MacArthur,M.W. and Thornton,J.M. (1998) *Curr. Opin. Struct. Biol.*, **8**, 631–639.
Lazaridis,T., and Karplus,M. (2000) *Curr. Opin. Struct. Biol.*, **10**, 139–145.
Merritt,E., and Bacon,D.J. (1997) *Methods Enzymol.*, **277**, 505–524.
Moult,J. (1997) *Curr. Opin. Struct. Biol.*, **7**, 194–199.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Park,B.H., Huang,E.S. and Levitt,M. (1997) *J. Mol. Biol.*, **266**, 831–846.
Samudrala,R., and Levitt,M. (2000) *Protein Sci.*, **9**, 1399–1401.
Samudrala,R., and Moult,J. (1998) *J. Mol. Biol.*, **275**, 895–916.
Samudrala,R., Xia,Y., Levitt,M. and Huang,E.S. (1999) In Altman,R., Dunker,K., Hunter,L., Klein,T. and Lauderdale,K. (eds), *Proceedings of the Pacific Symposium on Biocomputing*. pp. 505–516.
Samudrala,R., Xia,Y., Levitt,M. and Huang,E.S. (1999) *Proteins*, **S3**, 194–198.
Sippl,M.J. (1990) *J. Mol. Biol.*, **213**, 859–883.
Sippl,M.J. (1993) *Proteins*, **17**, 355–362.
Sippl,M.J. (1995) *Curr. Opin. Struct. Biol.*, **5**, 229–235.
Thomas,P.D., and Dill,K.A. (1996) *J. Mol. Biol.*, **257**, 457–469.
Tsai,J., Bonneau,R., Morozov,A.V., Kuhlman,B., Rohl,C.A. and Baker,D. (2003) *Proteins*, **53**, 76–87.
Wang,K., Fain,B., Levitt,M. and Samudrala,R. (2004) *BMC Struct. Biol.*, **4**, 8–25.
Zhang,C., Liu,S., Zhou,H. and Zhou,Y. (2004) *Biophys. J.*, **86**, 3349–3358.

Received February 2, 2006; revised April 21, 2006; accepted April 30, 2006

Edited by P. Balaram