PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Predicting structurally conserved contacts for homologous proteins using sequence conservation filters

Mayako Michino[1] and Charles L. Brooks III[2*]

[1] Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037
[2] Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT

The prediction of intramolecular contacts has a useful application in predicting the three-dimensional structures of proteins. The accuracy of the template-based contact prediction methods depends on the quality of the template structures. To reduce the false positive predictions associated with using the entire set of template-derived contacts, we develop selection filters that use sequence conservation information to predict subsets of contacts more likely to be structurally conserved between the template and the target. The method is developed specifically for protein families with few available templates such as the G protein-coupled receptor (GPCR) family. It is validated on a test set of 342 template-target pairs from three protein families, and applied to one template-target pair from the GPCR family. We find that the filter selection method increases the accuracy of contact prediction with sufficient coverage for structure prediction.

Proteins 2009; 77:448–453.

Key words: structural homology; sequence conservation; contact prediction; intramolecular contacts; template-based structure prediction; G protein-coupled receptors.

INTRODUCTION

G protein-coupled receptors (GPCRs) constitute a large and functionally diverse family of integral membrane proteins, with about 800 genes encoded by the human genome and only a handful of high-resolution three-dimensional structures.[1–4] The GPCR superfamily is characterized by a generally low sequence identity of below 30% and a structural similarity of ~2.0–3.0 Å Cα RMSD in the transmembrane core region among members of the family.[3,5] In predicting the structures for GPCRs of the unknown structure, template-based modeling is arguably the most reliable route, with one caveat that the prediction method needs to accurately refine the model closer to the native and away from the template. Most recent progress in protein structure refinement has been the development of the hybrid knowledge-based and physics-based potentials combined with spatial restraints from templates.[6–8] The spatial restraints confine the conformational search to a smaller phase space and facilitate a faster convergence towards near-native conformations.[9] Within this framework of the refinement approach, predicted intramolecular contacts can be usefully incorporated as tertiary restraints.

Among methods of protein contact prediction, template-based methods are more accurate than sequence-based methods.[6,10–13] The accuracy of the template-based methods depends on the quality of the template structures; the number of false positive contacts is larger for more structurally divergent templates. Because the false positive contacts are a potential source of inaccuracies in structure calculation,[6,14] we sought a method to reduce the number of false positives, especially in the case of templates that are up to 3 Å Cα RMSD from the target in the core region. It is well established that the level of sequence conservation is often indicative of structural and functional importance.[15,16] Hence, we decided to design a series of filters based on sequence conservation information to select the subsets of contacts that would more likely be structurally conserved between the template and the target.

In this article, we first describe the selection filters, then validate the method on a test set of 342 template-target pairs from three protein families. Although the method was developed specifically for application in the structure prediction of GPCRs, we use other protein families to test the method due to the paucity of known structures in the GPCR family. The three protein families in the test set have multiple experimentally determined high-resolution structures, and share

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: NIH; Grant number: RR012255; Grant sponsor: Novartis; Grant number: SFP_1562.

*Correspondence to: Prof. Charles L. Brooks III, Department of Chemistry and Biophysics Program, University of Michigan, 930 N University Ave, Ann Arbor, MI 48109. E-mail: brookscl@umich.edu

Received 2 November 2008; Revised 21 February 2009; Accepted 29 March 2009
Published online 20 April 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22456

similar properties to the GPCR family, that is, large and diverse yet reliable multiple sequence alignment, less than 30% sequence identity and less than 3 Å Cα RMSD structural divergence in the core region among members of the family. The template-target pairs in the test set collectively span a wide range in the sequence and structural similarity space, with ~10–85% sequence identity and ~0.3–2.7 Å Cα RMSD (see Fig. 1). The selection filter method is applied to one template-target pair from the GPCR family to demonstrate the applicability of the method to the protein family for which the method was originally developed. We find that the filter selection method increases the accuracy of common contact prediction with sufficient coverage for structure prediction, and more importantly, reduces the fraction of severely violated contacts.

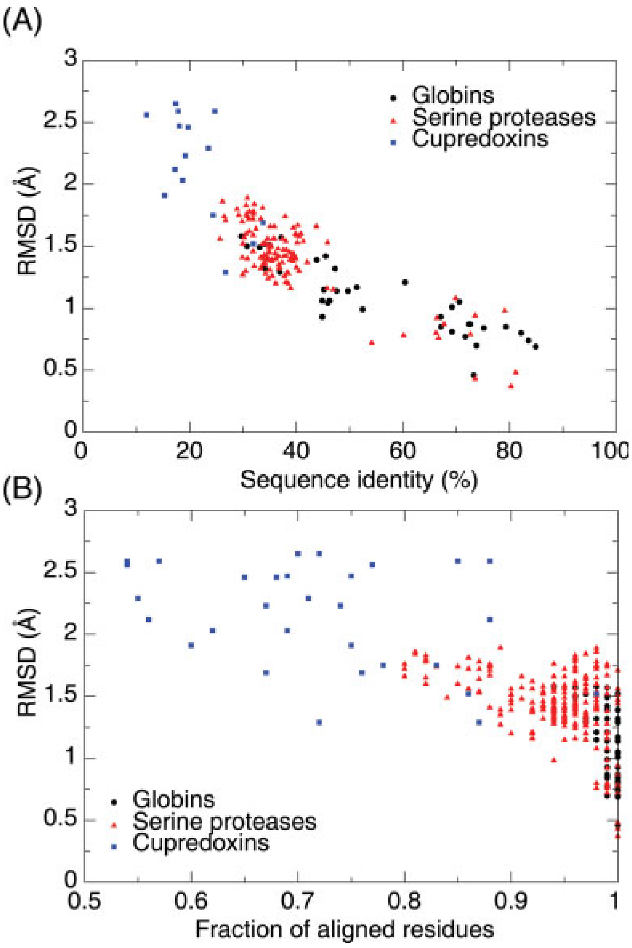

Figure 1

(**A**) Cα RMSD vs. sequence identity in the structurally aligned region of common core for the template-target pairs in the test set. Structures are superimposed by the LGA structural alignment program with a cutoff distance of 5 Å.[17] (**B**) Cα RMSD vs. fraction of aligned residues. The fraction of aligned residues is defined as N_{align}/L, where N_{align} is the number of residues in the structurally aligned region and L is the length of the target sequence. Our test set shows a strong negative correlation between sequence identity and Cα RMSD separation. The fraction of aligned residues is greater than 0.5 for all template-target pairs, and there is a weak negative correlation to Cα RMSD separation.

METHODS

Selection filters are defined based on sequence conservation metrics and applied to all residue pairs in the target that are aligned to the contact-forming residue pairs in the template. Given a known template structure, and a sequence alignment in which a contact-forming residue pair (i, j) in the template is aligned to a residue pair (i', j') in the target, the filters are used to predict if the residue pair (i', j') forms a common contact. The code implementing this method for the class A GPCRs will be made available through the MMTSB website (http://www.mmtsb.org) with the next release of the MMTSB Tool Set.[18]

Definition of contacts

A residue pair (i, j) is defined to form a contact if i and j are at least four residues apart, and the minimum inter-residue distance of any pair of heavy atoms in i and j is less than 4.2 Å. The list of contacts is obtained from 3D coordinates with the contact.pl utility in the MMTSB Tool Set (http://www.mmtsb.org).[18] This contact definition is close to the definition used by Skolnick and Kihara,[19] and it is more physically meaningful than the definition based on the C_β atoms, commonly used in CASP.[20]

Test sets and multiple sequence alignments

The performance of the selection filters is tested on template-target pairs from three protein families: globins (PDB ID codes: 2MHB:B, 1A4F:B, 1A9W:E, 1CG5:B, 1FDH:G, 1HBH:B, 1SPG:B, 2HHB:B, 2PGH:B), chymotrypsin-class serine proteases (PDB ID codes: 2PTN, 1MCT, 1A0J, 1AZZ, 1LMW, 1A5I, 1A5H, 1FUJ, 1HNE, 3RP2, 1DFP, 2TBS, 3EST, 1KLT, 1A7S, 1TRM), and monomeric cupredoxins (PDB ID codes: 1PLC, 1PAZ, 1AAC, 2CBP, 1JER, 1RCY). Within each family, all possible template-target pairs are used, thus the test set consists of a total of 342 template-target pairs (72 globin pairs, 240 serine protease pairs, and 30 cupredoxin pairs).

The sequence alignment between the template and the target is obtained from a multiple sequence alignment (MSA) of the protein family. We use MSAs reported in the literature.[21,22] The globin family MSA consists of 880 sequences (304 representative sequences at the level of 90% sequence identity); the serine protease family MSA consists of 616 sequences (402 representative sequences at the level of 90% sequence identity); the cupredoxin family MSA consists of 77 sequences (75 representative sequences at the level of 90% sequence identity). To avoid statistical bias from over-represented sequences, sequences with more than 90% identity to another sequence in the alignment are removed from the MSA using the program CD-HIT.[23]

Application to the GPCR superfamily

The selection filters are applied to the GPCR superfamily for the template-target pair of bovine rhodopsin and human β_2-adrenergic receptor (PDB ID codes: 1U19:A, 2RH1:A). We use an MSA of the transmembrane region of class A GPCRs reported in Suel GM *et al.*[22] The MSA consists of 940 sequences (550 representative sequences at the level of 90% sequence identity).

Selection filters

Sequence conservation is quantified by several different metrics and applied as selection filters to all template-derived residue pairs (i', j') in the target. The contacts selected by each selection filter are pooled to give the final set of predicted common contacts for the target protein.

Filter 1: sequence conservation at positions i and j

A six-letter reduced amino acid alphabet is used to detect sequence conservation between the residue pairs (i, j) and (i', j'), and across the MSA. The 20 standard amino acids are classified into six groups by their hydrophobicity (H-hydrophobic, A-amphipathic, P-polar) and size (S-small, L-large): {Gly, Ala, Val, Pro} (H,S), {Phe, Met, Ile, Leu} (H,L), {Thr, Cys} (A,S), {Trp, Tyr, Arg, Lys} (A,L), {Ser, Asn, Asp} (P,S), and {Glu, Gln, His} (P,L). The degree of conservation at each position n in the MSA is calculated by a stereochemically sensitive Shannon's entropy score.[24]

$$V_n = -\sum_{i}^{\kappa} p_i \ln p_i,$$

where p_i is the frequency of amino acid type i at position n in the MSA, and $\kappa = 6$ for the six groups of amino acids in the reduced alphabet. This score takes into account of the stereochemical properties of the amino acids to recognize that mutations between amino acids from the same physicochemical group are more conservative than those from different groups. Lower values of V indicate higher conservation.

The residue pair (i', j') is selected by filter 1 if the following two criteria are satisfied: (1) the residues i' and j' belong to the same amino acid groups as the residues i and j, respectively; (2) the sum of conservation score V at positions i and j is less than 1.0. The arbitrary threshold of $V_i + V_j < 1.0$ was found to be optimal in the test set [Supporting Information Fig. 1(A)].

Filter 2: sequence conservation of fragments around (i, j)

A fragment-based sequence similarity score is used to detect sequence conservation in short fragments surrounding the residue pairs (i, j) and (i', j'). Fragments of 13-residues, centered around i, i', j, j' are specified by the residues at positions $[i - 6{:}i + 6]$ and $[j - 6{:}j + 6]$. The sequence similarity between the aligned fragments is calculated by

$$S = \sum_{r_1=i-6}^{i+6} B(r_1, r_1') + \sum_{r_2=j-6}^{j+6} B(r_2, r_2'),$$

where B is the Blosum62 substitution matrix, and r and r' are residues from the template and the target, respectively.[25] S is standardized to a z-score S_z by

$$S_z = \frac{S - \bar{S}}{\sigma},$$

where $\bar{S}$ is the mean and σ is the standard deviation of S scores obtained for the particular template-target pair. A higher value of S_z indicates greater similarity.

The residue pair (i', j') is selected by filter 2 if S_z is greater than 0.8. The arbitrary threshold of $S_z > 0.8$ was found to be optimal in the test set [Supporting Information Fig. 1(B)].

Prediction accuracy and coverage

The filter selection method is evaluated by accuracy and coverage. Prediction accuracy and coverage are defined as Acc $= N_{\mathrm{corr}}/N_{\mathrm{pred}}$ and Cov $= N_{\mathrm{corr}}/N_{\mathrm{homo}}$, respectively, where N_{corr} is the number of correctly predicted common contacts, N_{pred} is the total number of contacts selected by the filters, and N_{homo} is the total number of true common contacts between the template and the target structures. The false positive rate is defined as FP $= 1 -$ Acc.

RESULTS AND DISCUSSION

Prediction accuracy and coverage for the test set

The prediction accuracy and coverage for the test set are shown in Figure 2. Accuracy varies between 0.67 and 0.98, with an average of ~0.85 across all RMSD ranges (mean ± SD for <1.0 Å Cα RMSD: 0.87 ± 0.07; 1.0–2.0 Å Cα RMSD: 0.84 ± 0.05; >2.0 Å Cα RMSD: 0.83 ± 0.06). Coverage varies between 0.24 and 0.56, with each protein family forming a cluster about its mean (mean ± SD for globins: 0.47 ± 0.07; serine proteases: 0.32 ± 0.03; cupredoxins: 0.33 ± 0.05). It is notable that both accuracy and coverage remain nearly constant across most of the RMSD range and do not show significant decreases at higher structural divergence.

Contacts selected by filter 1 (conservation of i, j) and filter 2 (conservation of fragments around i, j) are largely nonoverlapping, with ~20% of the final predicted con-

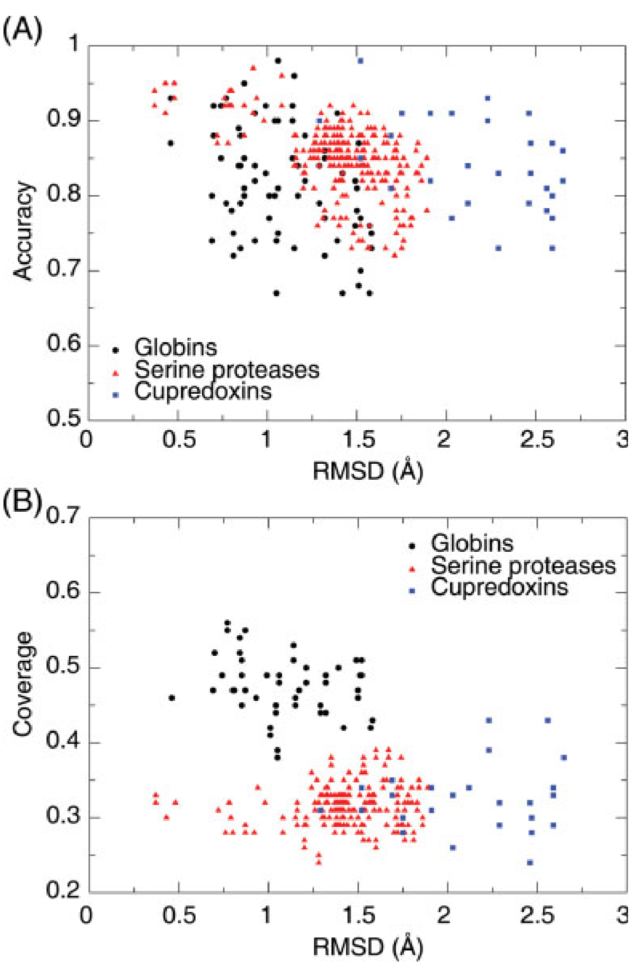

Figure 2

Prediction accuracy (**A**) and coverage (**B**) for the template-target pairs in the test set. Accuracy and coverage are defined as Acc = N_{corr}/N_{pred} and Cov = N_{corr}/N_{homo}, respectively, where N_{corr} is the number of correctly predicted common contacts, N_{pred} is the total number of contacts selected by the filters, and N_{homo} is the total number of true common contacts between the template and the target structures.

tacts selected by both (data not shown). Depending on whether the filters operate on just the contact-forming residues or the fragments around them, they select different subsets of contacts from the template structure. Filter 1 selects for contacts formed by highly conserved residues and filter 2 selects for contacts formed by residues located in regions of high sequence conservation. It is expected that accuracy may be maximized for contacts selected by both filters, at the cost of decreased coverage. We use the combined set of contacts selected by either filters, to maximize coverage so that a sufficient level is reached for application in structure prediction.

Alternative filters as possible improvements

Although there is no universally applicable amino acid groupings for a reduced alphabet, it was recently suggested that the optimal reduced alphabet for GPCR classification may consist of a larger number of groups than the classical three groups based on hydrophobicity, first introduced by Chothia and Finkelstein,[26] and may lie in the 7–11 amino acid region.[27] In our selection filter, we use a six-letter alphabet based on hydrophobicity and size. Although this grouping is more complex than the three-letter alphabet based on hydrophobicity alone, there is a possibility that another alphabet may be better suited for detecting sequence similarity among GPCR sequences.

Among the numerous metrics available to quantify the site-specific residue conservation in an MSA, we use an entropy-like score in our selection filter. Although this score is simple to calculate, more sophisticated scores based on probabilistic evolutionary models and phylogenetic tree such as the scores calculated by the ConSeq and ConSurf servers may provide better estimations of site-specific conservation in the MSA.[16,28]

Our selection filters are solely based on sequence conservation information. However, other sequence information such as correlated mutation, first introduced by Gobel *et al.*, is weakly correlated with inter-residue contact formation, and hence may be incorporated as additional selection filters to improve both prediction accuracy and coverage.[12,29]

Use of filtered contacts in structure prediction

High accuracy and sufficient coverage are necessary for the predicted contacts to be usefully incorporated into structure prediction methods.[9,11] False positive contacts not only mislead structure calculations toward the template structure and away from the target structure but also may not allow the models to converge. It has been suggested that for single-domain proteins with up to 200 residues, the number of distance restraints necessary to deduce a low-resolution fold is ∼L/8, where L is the length of the target sequence.[9,30]

Our filter selection method improves the accuracy of contact prediction by an average of ∼0.1 across all RMSD range, with increasing improvements at higher RMSD (mean ± SD for <1.0 Å Cα RMSD: 0.05 ± 0.04; 1.0–2.0 Å Cα RMSD: 0.11 ± 0.04; >2.0 Å Cα RMSD: 0.20 ± 0.08) [Fig. 3(A)]. The number of predicted contacts, assessed by the percentage relative to the length of the target sequence (N_{pred}/L), is greater than the minimum requirement of 1/8 (0.125) for all template-target pairs in the test set (mean ± SD for globins: 0.41 ± 0.04; serine proteases: 0.57 ± 0.06; cupredoxins: 0.38 ± 0.09) [Fig. 3(B)]. Furthermore, our selection filters are effective in reducing the fraction of severely violated contacts [Fig. 3(C)]. A predicted contact (i', j') is defined to be severely violated if the minimum inter-residue distance of all pairs of heavy atoms is greater than 8.0 Å in

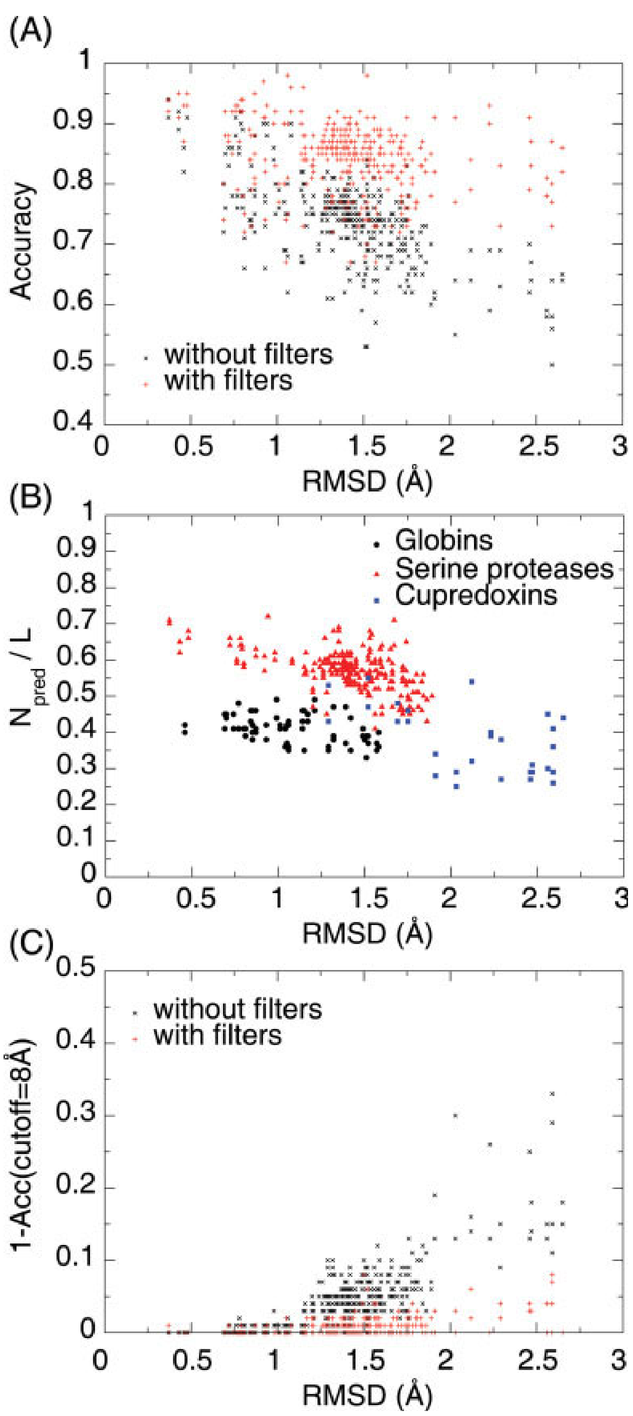

Figure 3

(**A**) Comparison of prediction accuracy with and without selection filters applied to the template-derived contacts (i', j'). (**B**) The percentage of the number of predicted contacts with the selection filters relative to the length of the target sequence. (**C**) Comparison of the fraction of severely violated contacts in the predicted contacts with and without selection filters applied. A severely violated contact is defined as a template-derived contact (i', j') that has a minimum inter-residue heavy atom distance of greater than 8 Å.

the target structure; these contacts would contribute most towards the inaccuracies in structure calculation, as the contact restraints in structure prediction protocols are typically imposed with harmonic functions and the penalty is larger for greater deviations. The fraction of severely violated contacts is reduced by an average of ~0.15 for template-target pairs in the >2.0 Å Cα RMSD region. Taken together, our selection filter method predicts sufficient numbers of contacts with high accuracy for the template-target pairs in the test set.

Filtered contacts for GPCR structure prediction

The selection filters were applied to the template-target pair of bovine rhodopsin and human β_2-adrenergic receptor (β2AR). Of the 183 interhelical contacts in the transmembrane region of the rhodopsin structure, the filters selected a total of 45 contacts, with an accuracy of 0.69 (31/45), and coverage of 0.27 (31/113) (see Fig. 4). The accuracy is improved from 0.62 (113/183) and the

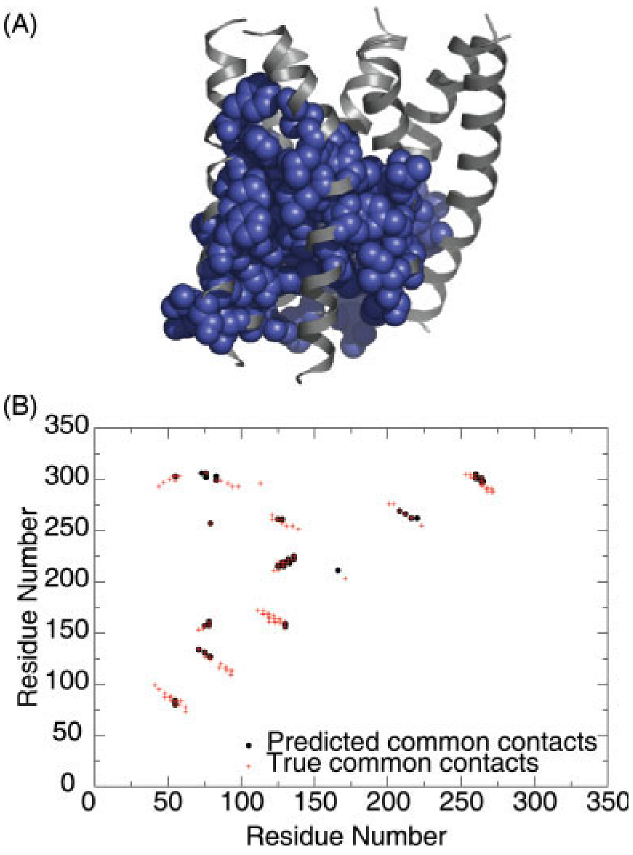

Figure 4

(**A**) The predicted common contacts for the β2AR are mapped onto the rhodopsin structure and shown in blue van der Waals representation. (**B**) The predicted and true common contacts between rhodopsin and β2AR are shown on the contact map of rhodopsin.

fraction of severely violated contacts is reduced to 0 from 0.03 (6/183) when compared with the unfiltered set of contacts. The number of selected contacts is sufficient for use in structure calculation ($N_{\text{pred}}/L \sim 45/194 = 0.23$).

The selected contacts have recently been used to predict the structure of the transmembrane region of β2AR with some success in accurately modeling the structural divergence between rhodopsin and β2AR (Michino et al., in preparation). Furthermore, we show that the β2AR models generated with the selected contacts are overall more accurate than the models generated with the unfiltered set of contacts. The inaccuracies in the latter models, especially in the tilt angle of transmembrane helix I, can be attributed to the three severely violated contacts that are with respect to the extracellular side of helix I.

CONCLUSIONS

We have developed a filter selection method for improving the accuracy of template-based contact prediction. The selection filters are based on sequence conservation information. The method is validated on a test set of 342 template-target pairs from three protein families, and applied to one template-target pair from the GPCR family. When compared with the unfiltered set of template-derived contacts, the selected subset of contacts is more accurate and has a reduced fraction of severely violated contacts. The selected set of contacts is expected to be usefully incorporated into structure prediction methods.

ACKNOWLEDGMENTS

We thank Jianhan Chen for helpful discussions and feedback.

REFERENCES

1. Pierce KL, Premont RT, Lefkowitz RJ. Seven-transmembrane receptors. Nat Rev Mol Cell Biol 2002;3:639–650.
2. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: A G protein-coupled receptor. Science 2000;289:739–745.
3. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 2007;318: 1258–1265.
4. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF. Structure of a beta1-adrenergic G-protein-coupled receptor. Nature 2008;454:486–491.
5. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 2003;63:1256–1272.
6. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.
7. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 2005;102:1029–1034.
8. Chen J, Brooks CL, III. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 2007; 67:922–930.
9. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
10. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 2008;24:924–931.
11. Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69(Suppl 8):152–158.
12. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
13. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
14. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
15. Godzik A, Sander C. Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng 1989;2:589–596.
16. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 2005;33(Web Server issue):W299–W302.
17. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
18. Feig M, Karanicolas J, Brooks CL, III. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model 2004;22:377–395.
19. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR–a new approach to threading. Proteins 2001;42:319–331.
20. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61(Suppl 7):214–224.
21. Gough J, Chothia C. The linked conservation of structure and function in a family of high diversity: the monomeric cupredoxins. Structure 2004;12:917–925.
22. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
23. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.
24. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
25. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
26. Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–1039.
27. Davies MN, Secker A, Freitas AA, Clark E, Timmis J, Flower DR. Optimizing amino acid groupings for GPCR classification. Bioinformatics 2008;24:1980–1986.
28. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N. ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 2004;20:1322–1324.
29. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D. Co-evolving residues in membrane proteins. Bioinformatics 2007;23:3312–3319.
30. Li W, Zhang Y, Skolnick J. Application of sparse NMR restraints to large-scale protein structure prediction. Biophys J 2004;87:1241–1248.