Published in final edited form as:

J Mol Model. 2011 November ; 17(11): 2831–2837. doi:10.1007/s00894-011-0976-9.

Systematic Assessment of Accuracy of Comparative Model of Proteins Belonging to Different Structural Fold Classes

Suvobrata Chakravarty,

Department of Chemistry & Biochemistry, Box 2202, South Dakota State University, Brookings, SD, 57007, USA. Suvobrata.chakravarty@sdstate.edu; Phone: 605-688-5694; Fax: 605-688-6364

Dario Ghersi, and

Department of Structural and Chemical Biology, Box 1677, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY 10029, USA

Roberto Sanchez

Department of Structural and Chemical Biology, Box 1677, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY 10029, USA. roberto.sanchez@mssm.edu; Tel: 212-659-8648. Fax: 1-212-659-8232

Abstract

In the absence of experimental structures, comparative modeling continues to be the chosen method for retrieving structural information of target proteins. However, models lack the accuracy of experimental structures. Alignment error and structural divergence (between target and template) influence model accuracy the most. Here, we examine the potential additional impact of backbone geometry, as our previous studies have suggested that the structural class (all- α , $\alpha\beta$, all- β) of proteins may influence the accuracy of their models. In the twilight zone (sequence identity $\leq 30\%$) and at a similar level of target-template divergence, model accuracy of proteins does indeed follow the trend all- $\alpha > \alpha\beta > \text{all-}\beta$. This is mainly a result of the alignment accuracy following the same trend (all- $\alpha > \alpha\beta > \text{all-}\beta$) with backbone geometry playing only a minor role. Differences in the diversity of sequences belonging to different structural classes leads to the observed accuracy differences thus enabling *a priori* accuracy estimates of alignments/models in a class-dependent manner. This study provides a systematic description and quantification of structural class-dependent effect in comparative modeling. The study also suggests datasets for large-scale sequence/structure analyses should have equal representation of different structural classes to avoid class-dependent bias.

Keywords

Homology Modeling; Model Accuracy; Sequence Alignment; Alignment Accuracy; Information Content; Secondary Structure

Introduction

Comparative (or homology) modeling uses experimentally determined protein structures (templates) to predict the 3D conformation of another protein with a similar amino acid sequence (target). As the number of known protein folds is potentially approaching completion [1] due to progress in structural genomics initiatives, comparative modeling will

continue to be the method of choice for building protein structure models [2-4]. Not only is comparative modeling the most accurate method for structure prediction [5] [6], but it also allows *a priori* estimation of the approximate quality of the models [7]. In addition, due to their added value [8], models are particularly suitable for comparative studies over complete protein families [9-11]. Despite significant progress in X-ray crystallography and high-field NMR spectroscopy for structure elucidation [12,13], structures of many protein, including therapeutically relevant targets remain unavailable. A number of studies have documented the usefulness of comparative modeling in therapeutic [14] and general applications such as identification of inhibitor/antagonists [15-17], virtual screening [18,19], molecular replacement [20], and function prediction [21].

In spite of successful applications, predicted structures generally contain errors and seldom reach the accuracy of experimental structures. Three variables influence accuracy of a comparative model [22]: (i) the structural similarity between target and template, (ii) the target-template alignment accuracy; and (iii) the ability to refine the model (i.e. loop modeling and general refinement). Combining these factors the quality of a model, measured in terms of errors, can be defined as [23]:

$$\text{Total Error} = \text{Structural Difference} + \text{Alignment Error Refinement} \quad [1]$$

Alignment error is frequently cited as the single most important variable influencing the quality of comparative models [24] [25]. We recently showed that not only do alignment errors affect model accuracy directly but they also affect it indirectly by depriving the models from the benefit of structural complementarity in multiple template modeling [23]. Thus a very strong correlation between the alignment accuracy and model accuracy is observed. Additionally, our previous work noted that the average model accuracy obtained at a certain level of alignment accuracy varied depending on the structural class of proteins (all- α , all- β and $\alpha\beta$) [23]. This suggested a role for backbone geometry, not discussed before, in influencing model accuracy. In the present work, we systematically investigate this relationship between model accuracy and protein structural class. This is relevant since most studies do not distinguish between structural classes when describing or estimating the accuracy of comparative models [26,27-29] [7]. Hence, we report here a comparison of the accuracy of comparative models of proteins belonging to the three main structural classes: all- α , all- β and $\alpha\beta$ proteins. The accuracy of models is measured by comparing them with their respective target experimental structure. As in our previous studies we make use of a reference set of single-template comparative models without an explicit refinement step (e.g. loop modeling) [7]. In this context the effect of the model building method is negligible [5], thus making the conclusions general and applicable to the use of any comparative modeling software.

Methods

Construction of dataset

Our previous dataset [7] was the starting point of construction (see Supplementary information). The selected pairs (size: 100-150 residues) were then sorted into sequence identity bins (10% to 60% with a bin size of 5%) such that each bin had the same number of all- α , all- β and $\alpha\beta$ pairs. The selected polypeptide chains of 100-160 residues were further sorted into 3 bins of 110 ± 10 , 130 ± 10 and 150 ± 10 residues. To avoid any observed difference resulting out of the differences in protein size each sequence identity bin had nearly same number of chains of 110 ± 10 , 130 ± 10 and 150 ± 10 residues from each of all- α , all- β and $\alpha\beta$ pairs. A total of 4,500 sequence pairs were finally selected by choosing 1,500 pairs for each of class. In addition, for each of the three structural classes, the template-

target pairs were selected such that the average STR model accuracy curve, a measure of structural divergence, of each class across the sequence identity bins was nearly identical (Fig 1b). We paid attention in the selection of pairs such that the overall structural deviation of selected pairs, irrespective of the class, was very similar.

Model Building

For each of the 4,500 selected pairs three different alignments (SEQ, PRO, and STR) and their corresponding models were constructed (see below), resulting in a total of 13,500 models for this study. The alignments and template structures were used as input to the program MODELLER version 6v2 [30]. **SEQ-Model:** The target sequence is aligned with the template using the ALIGN command of MODELLER. **STR-Model:** The structural alignment between the target and the template was generated using the ALIGN3D command of MODELLER. **PRO-Model:** The target-template profile-profile based alignment was generated in the following way. The non-redundant (NR) database was searched separately with each of the target and template sequences as query using PSI-BLAST with 20 iterations [31]. The PSI-BLAST hits were sorted in an ascending order of e-values for each query sequence. The first 1000 hits with the lowest e-values were selected to generate a multiple sequence alignment in a manner that did not alter the query-hits BLAST alignment. The two multiple alignments (one for each target-template pair) were simultaneously used as inputs to MODELLER 8v2 [30]. The profile-profile alignment was generated using the SALIGN command of MODELLER. The target and the template sequences were extracted from the profile-profile alignment without altering the alignment.

Overall Accuracy of models

Overall accuracy of models was computed by measuring the root mean square deviation between the equivalent C α atoms in the optimal superposition of target and model structures. Equivalent atoms are defined as those that are within 3.5Å of their corresponding atoms in the target after optimal superposition of the structures. The structural superposition was carried out by minimizing the root mean squared deviation of the equivalent C α atom coordinates. All the calculations are implemented in the SUPERPOSE command of the program MODELLER.

Alignment Accuracy

The quality of the target-template SEQ and PRO alignments was determined with respect to the target-template STR alignments. Alignment accuracy is defined as the ratio between the number of correctly aligned positions and the number of aligned positions of a given alignment. A target-template aligned residue pair of a given alignment is said to be a correctly aligned position if the particular pair is also aligned in the STR alignment [24]. The alignment accuracy used in this study refers to Q_{mod} [24]. Q_{mod} is defined as the number of amino acids correctly aligned in the sequence alignment divided by the total number of aligned residues in the sequence alignment. Residue neighborhood and pocket accuracy were determined as described in [7].

RESULTS

Comparison of overall accuracy

The average overall accuracy of models of individual protein domains (~150 residues) belonging to all three structural classes is shown in Fig. 1. As in previous studies [7] models were built on single templates using three alternative alignments: target-template pairwise sequence alignment (SEQ-Models), profile-profile alignment (PRO-Models) and structure-based alignment (STR-Models) between the target and the template (see Methods). To be

consistent with our earlier work and that of others [7], all comparisons of model accuracy are discussed as a function of template-target sequence identity. The respective average accuracy of SEQ, PRO and STR-models of the combined (all- α + $\alpha\beta$ + all- β) dataset is shown as reference (Fig. 1b). As previously shown, model quality increases in the order SEQ, PRO to STR (Fig. 1b), due to the increase in the alignment quality. Both SEQ (Fig. 1a, left) and PRO-Models (Fig. 1a, right) of domains belonging to the all- α class are observed to be more accurate than SEQ and PRO-Models of $\alpha\beta$ and all- β classes respectively. In general, the accuracy was observed to decrease in the order all- α , $\alpha\beta$ to all- β , indicating a possible contribution of backbone geometry in affecting model quality. The behavior of the $\alpha\beta$ category of models is close to that of the combined averages of Fig. 1b, whereas those of all- α and all- β are respectively better and worse than the group average. Since the length of modeled segments (domain size) in each of the three classes is very similar (~150 residues), the effect of protein size on the observed difference can be discounted. The distribution of RMSD of structural superposition of pairs irrespective of the class was also kept nearly identical for each sequence identity bin. This selection criterion was important for the study because we wanted to see the effect of alignment accuracy given the same of level of structural divergence (see Eqn 1). Hence, the origin of the difference in the model accuracy was not due to a bias such as selecting pairs structurally more similar in the all- α class than in the other classes. Furthermore, we selected a set of all- α pairs with larger structural divergence (larger than average RMSD of STR-Model curve, Fig. 1b) than the rest of the two classes and this still resulted in more accurate PRO and SEQ-Models than the rest, hence ruling out selection bias. The fact that the differences only appear below 30% sequence identity strongly suggests that sequence alignment accuracy plays a significant role in the observed difference in model accuracy.

Comparison of alignment accuracy

Protein structure of one class primarily differs from those of the others in the geometry of the backbone. In addition to alignment, the backbone geometry may also contribute the observed difference in model accuracy (see Supplementary Information for details). If this is indeed the case, models of proteins from different structural classes will show significant differences in model accuracy (C α -RMSD) at the same level of alignment accuracy (irrespective of target-template sequence similarity). Fig. 2 shows that in fact there is a difference between the model accuracy of proteins from different classes at the same level of alignment accuracy, indicating that the geometric factor contributes to the differences in model accuracy. However, this is only observed for lower quality alignments, below ~30% alignment accuracy and the fraction of models with this level of alignment accuracy is only ~15% (Fig. 2). This observation suggests that the geometric factor plays a role only in a small number of cases. It also indicates that, even though on average different small isolated structural segments (Supplementary Figure S1) respond differently (C α -RMSD) to the same level of alignment error, the contribution of the geometric factor over the entire structure is averaged out and, therefore, much smaller. The alignment factor is examined next. The histogram of fraction of models with certain level of alignment accuracy (Fig. 2) shows that proteins belonging to the all- α category have a relatively smaller proportion of models with poor alignments and relatively higher proportion of accurate alignments than the all- β class. The distribution indicates that the alignment accuracy of proteins of all- α category is higher than that of all- β proteins, while the $\alpha\beta$ category has an intermediate level of alignment accuracy (Fig 3a, only PRO-Models are shown). Like model accuracy, prediction accuracy of a number of structure-derived properties (SDPs) such as inter-residue distance, residue neighborhood, etc. varies substantially between the structural classes (see Supplementary Information for details).

Entropy of class specific sequences

The quality of alignment improves when the amount of information is increased in the alignment procedure. For example, PRO alignments are more accurate than SEQ alignments because additional information about general features in protein families such as position specific sequence conservation is utilized [32] even if the query sequence lacks some of these features. Similarly, STR alignments, by virtue of high-resolution structural information to establish equivalences between residues of the two sequences, have the highest quality. Due to the observed higher accuracy of alignments of all- α pairs, we anticipated that quality of information involving all- α sequences might be different than that of the rest. The information used in constructing an alignment, such as a profile-profile alignment, can be assessed by an entropy measure. The variability of amino acids in a profile column can be the basis of the entropy measure. Earlier studies have also reported entropy as a measure of variability in profiles, by counting the average number of different symbols in a profile column [33] [34] [35]. Identification of reliable regions in an alignment using a scoring scheme based on the sequence profile column has also been reported [36], justifying the relationship between entropy and alignment quality. We examined the average entropy in sequence profile columns to see if profiles generated from all- α class proteins differed in entropy from those of all- β proteins, thus explaining the observed difference in alignment accuracy. A distribution of the average column entropy (Shannon entropy) of sequence profiles for each category (Fig. 3c) shows that there is indeed a difference in entropy between profiles in each category. A higher average value of entropy implies that each column has more diverse set of amino acids in the profile than one with lower entropy. The average values of entropy computed from profiles show a clear trend all- α (2.059 ± 0.609) < $\alpha\beta$ (2.209 ± 0.557) < all- β (2.303 ± 0.517). The profiles of all- α proteins tend to have lower entropy, indicating there is on average less variability among amino acids in a profile column of all- α than in all- β case. An earlier study on protein families had highlighted the reduction of the alignment quality, in the twilight zone, due to sequence diversity in profiles [33,37]. The study indicated that families containing more diverse homologues in general produce less accurate profile-profile alignments [33,37]. Hence, our observation could simply be capturing this general caveat of protein families categorized in the structural classes: all- α , all- β and $\alpha\beta$; where $\alpha\beta$ and all- β are in general more diverse in sequence space than the all- α class.

Discussion

It is intriguing to ask if there is any structural basis for sequence diversity in a class dependent manner, a question that is closely related to the topological basis of protein ‘designability’ studied earlier [38,39] which addressed the quantity of sequence space associated with a given structure/topology. Koehl & Levitt clearly demonstrated that both protein size and topology influenced the sequence space that can be accommodated in a structure fold [38]. A fixed set of designed sequences, energetically compatible with two different structural folds of similar size, was used to compute a per residue entropy measure. In the study, ribosomal protein L7/L12 C-terminal domain (1ctf, $\alpha\beta$ fold) was compatible with less diverse sequences (low entropy) than that of SH3 domain (2hsp, all- β fold)[38] very similar to the naturally occurring sequences of these proteins in the HSSP database. The higher helical content of ribosomal protein L7/L12 C-terminal domain restrained the choice of amino acids in its sequence (low entropy) as there is a clearer pattern of amino-acid preferences for helices (helix propensities) than for strands [40]. Unlike α -helix formation, β -sheet propensity is modulated strongly by the tertiary context and not by intrinsic secondary structure preference [40] as β -sheets are elements of both secondary and tertiary structure. This may translate to a weaker signal of preferred amino acids for a strand-forming segment than for a helix-forming segment. Hence, the increase in the entropy

of profiles of all- β and $\alpha\beta$ proteins compared to the all- α class could be explained by a wider spectrum of different amino acids that can be accommodated in a strand than a helix. An earlier study mapping local sequence to local structure by clustering segments of 3-15 residues of HSSP structural alignments of protein families also showed that sequence patterns were found more frequently in helices (such as Amphipathic helix, Less amphipathic helix, Helix N-cap, Schellman helix C-cap, Schellman helix-turn-sheet etc) than in strands (Amphipathic strands, α L Strand C-cap, Buried strand) [41-43]. Even a simple binary pattern (hydrophobic-polar, HP) analysis had showed a stronger sequence-structure correlation for α -helices than for β -strands [44]. The lower entropy values of helices probably aids in aligning more recognizable patterns in the sequence resulting in a better alignment. This is true even for SEQ alignments as we observe the accuracies of them to follow the same order all- α > $\alpha\beta$ > all- β proteins.

Shakhnovich et al. 2005 showed that the number of sequences in a domain family depended on the contact density (CD) of the structure [45]. They reasoned that, the higher the proportion of favorably interacting residues that stabilize a fold or topology, the higher is the chance for the rest of the sequence to change without destabilizing the fold. Simply, if the CD for a given topology is high there are more sequences adopting the topology [45]. A distribution of the CDs for all- α class structures shows smaller values compared to that of the $\alpha\beta$ class (supplementary figure S4), indicating that CD may play a role. However, all- β class proteins showed the smallest CD, which is not consistent with our observation. It could be that family size may not translate to higher sequence entropy when similar sequences populate a family.

Re-examining the accuracy of models of small, medium and large proteins in our earlier study (Fig. 1a of [7]) revealed that the much superior accuracy of small proteins compared to those larger ones was due to over representation of all- α proteins in the small proteins set, whereas the medium and large sized protein sets were equally represented by all the three groups. This underscores the need for large-scale analyses of models and alignments [46,33] to consider equal representation of proteins of all three classes to avoid class dependent artifacts. The structural class dependent effect described here appears to be a general property of protein sequences and structures. The present study provides a systematic description and quantification of this effect in comparative modeling. However a precise deconvolution of the complex interplay of various entangled factors such as sequence identity, information content, alignment accuracy, and class specific substitution matrices, requires further study at an even larger scale. It is also important to note that enzymes are dominated by the α/β (TIM barrel) and $\alpha+\beta$ (RRM-like fold) class, along with β -class members such as the double-stranded β -helix (DSBH) domain and the β -propellers [47]. The α -class does not include dominant enzymatic folds [47]. This suggests that use of modeled enzyme structures for high-resolution structural feature analyses, in absence of experimental structures, are to be performed with greater caution due to higher errors associated with non-helical folds.

Even though a single program, MODELLER [30], was used here for model building, the results obtained are software-independent as the study is based on simple comparative models [7]. The modeling procedure involves no explicit refinement protocol such as loop modeling or any other refinement making the model accuracy dependent only on alignment accuracy and target-template structural difference (see Eqn 1). In order to examine the influence of software on alignment accuracy, alternate alignments using program MUSCLE [48] were constructed, and also showed clear class-dependent alignment accuracy (data not shown), further supporting the software-independent nature of our results.

Conclusions

In continuity of our characterization of simple comparative models [7] [8] [23], the results of this large-scale (~15,000 models) study show that, in the twilight zone of sequence similarity and given the same level of target-template structural divergence, proteins belonging to different structural classes tend to be modeled with different accuracy thus enabling *a priori* accuracy estimates of alignments/models in a class-dependent manner. As the geometry of the polypeptide backbone contributes negligibly to this observation, it is the dependence of the alignment accuracy on structural class that is the major contributing factor resulting in the observed differences in model accuracy. This study once again reiterates the importance of alignment accuracy in comparative modeling and that its causes and effects are not simply a function of sequence similarity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Sucheta Godbole for helping us with the profile-profile alignments. We thank Zhanwen Li of the Godzik laboratory at the Burnham Institute for helping us with the FFAS server to the test cases of profile-profile alignments. SC thanks Prof Ming-Ming Zhou for encouragement. The study was supported by National Institute of General Medicine at the National Institutes of Health grant 1R01GM081713 (RS), and SDSU Agricultural Experimentation and Biological Control And Analysis by Applied Photonics (BCAAP), the South Dakota 2010 center grant 3SG163 (SC).

References

1. Taylor WR. Evolutionary transitions in protein fold space. *Curr Opin Struct Biol.* 2007; 17(3):354–361. [PubMed: 17580115]
2. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci U S A.* 1998; 95(23):13597–13602. [PubMed: 9811845]
3. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A. Protein structure modeling for structural genomics. *Nat Struct Biol.* 2000; 7(Suppl):986–990. [PubMed: 11104007]
4. Stevens RC, Yokoyama S, Wilson IA. Global efforts in structural genomics. *Science.* 2001; 294(5540):89–92. [PubMed: 11588249]
5. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. *Proteins.* 2003; 53(Suppl 6):352–368. [PubMed: 14579324]
6. Lushington GH. Comparative modeling of proteins. *Methods in molecular biology* (Clifton, NJ. 2008; 443:199–212.
7. Chakravarty S, Wang L, Sanchez R. Accuracy of structure-derived properties in simple comparative models of protein structures. *Nucleic Acids Res.* 2005; 33(1):244–259. [PubMed: 15647507]
8. Chakravarty S, Sanchez R. Systematic analysis of added-value in simple comparative models of protein structure. *Structure (Camb).* 2004; 12(8):1461–1470. [PubMed: 15296739]
9. Kiel C, Wohlgemuth S, Rousseau F, Schymkowitz J, Ferkinghoff-Borg J, Wittinghofer F, Serrano L. Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations. *J Mol Biol.* 2005; 348(3):759–775. [PubMed: 15826669]
10. Liu T, Rojas A, Ye Y, Godzik A. Homology modeling provides insights into the binding mode of the PAAD/DAPIN/pyrin domain, a fourth member of the CARD/DD/DED domain family. *Protein Sci.* 2003; 12(9):1872–1881. [PubMed: 12930987]
11. Murray PS, Li Z, Wang J, Tang CL, Honig B, Murray D. Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. *Structure.* 2005; 13(10):1521–1531. [PubMed: 16216583]

12. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*. 2002; 58(Pt 6 No 1):899–907. [PubMed: 12037327]
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28(1):235–242. [PubMed: 10592235]
14. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. *Drug discovery today*. 2004; 9(15):659–669. [PubMed: 15279849]
15. Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc Natl Acad Sci U S A*. 1993; 90(8):3583–3587. [PubMed: 8475107]
16. Evers A, Klabunde T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *Journal of medicinal chemistry*. 2005; 48(4):1088–1097. [PubMed: 15715476]
17. Evers A, Klebe G. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of medicinal chemistry*. 2004; 47(22):5381–5392. [PubMed: 15481976]
18. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *Journal of medicinal chemistry*. 2003; 46(13):2656–2662. [PubMed: 12801229]
19. Lengauer T, Lemmen C, Rarey M, Zimmermann M. Novel technologies for virtual screening. *Drug discovery today*. 2004; 9(1):27–34. [PubMed: 14761803]
20. Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr D Biol Crystallogr*. 2001; 57(Pt 10):1373–1382. [PubMed: 11567148]
21. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. *Nature biotechnology*. 2000; 18(3):283–287.
22. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*. 2005; 15(3):285–289. [PubMed: 15939584]
23. Chakravarty S, Godbole S, Zhang B, Berger S, Sanchez R. Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC structural biology*. 2008; 8:31. [PubMed: 18631402]
24. Sauder JM, Arthur JW, Dunbrack RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*. 2000; 40(1):6–22. [PubMed: 10813826]
25. Dunbrack RL Jr. Sequence comparison and protein structure prediction. *Curr Opin Struct Biol*. 2006; 16(3):374–384. [PubMed: 16713709]
26. Baker D, Sali A. Protein structure prediction and structural genomics. *Science*. 2001; 294(5540):93–96. [PubMed: 11588250]
27. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*. 2007; 69(Suppl 8):38–56. [PubMed: 17894352]
28. Nayeem A, Sitkoff D, Krystek S Jr. A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. *Protein Sci*. 2006; 15(4):808–824. [PubMed: 16600967]
29. Rayan A. New tips for structure prediction by comparative modeling. *Bioinformatics*. 2009; 3(6):263–267. [PubMed: 19255646]
30. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993; 234(3):779–815. [PubMed: 8254673]
31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25(17):3389–3402. [PubMed: 9254694]
32. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci*. 2000; 9(2):232–241. [PubMed: 10716175]

33. Sadreyev RI, Grishin NV. Estimates of statistical significance for comparison of individual positions in multiple sequence alignments. *BMC bioinformatics*. 2004; 5:106. [PubMed: 15296518]
34. Panchenko AR. Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res*. 2003; 31(2):683–689. [PubMed: 12527777]
35. Casbon J, Saqi MA. S4: structure-based sequence alignments of SCOP superfamilies. *Nucleic Acids Res*. 2005; 33(Database issue):D219–222. [PubMed: 15608181]
36. Tress ML, Jones D, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol*. 2003; 330(4):705–718. [PubMed: 12850141]
37. Sadreyev RI, Grishin NV. Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. *Bioinformatics*. 2004; 20(6):818–828. [PubMed: 14751996]
38. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci U S A*. 2002; 99(3):1280–1285. [PubMed: 11805293]
39. England JL, Shakhnovich EI. Structural determinant of protein designability. *Physical review letters*. 2003; 90(21):218101. [PubMed: 12786593]
40. Minor DL Jr, Kim PS. Context is a major determinant of beta-sheet propensity. *Nature*. 1994; 371(6494):264–267. [PubMed: 8078589]
41. Han KF, Baker D. Recurring local sequence motifs in proteins. *J Mol Biol*. 1995; 251(1):176–187. [PubMed: 7643386]
42. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci U S A*. 1996; 93(12):5814–5818. [PubMed: 8650175]
43. Bystroff C, Simons KT, Han KF, Baker D. Local sequence-structure correlations in proteins. *Current opinion in biotechnology*. 1996; 7(4):417–421. [PubMed: 8768900]
44. West MW, Hecht MH. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci*. 1995; 4(10):2032–2039. [PubMed: 8535239]
45. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. Protein structure and evolutionary history determine sequence space topology. *Genome research*. 2005; 15(3):385–392. [PubMed: 15741509]
46. Edgar RC, Sjolander K. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*. 2004; 20(8):1301–1308. [PubMed: 14962936]
47. Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Current opinion in chemical biology*. 2003; 7(1):12–20. [PubMed: 12547421]
48. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*. 2004; 5:113. [PubMed: 15318951]

Fig. 1.

Accuracy of comparative models: **(a)** Comparison of Cα RMSD between experimental and modeled structure between all-α, all-β and αβ proteins (SEQ-models, left and PRO-models, right). **(b)** Comparison of Cα RMSD between modeled experimental structures SEQ, PRO and STR-models.

Fig. 2. Comparison of the relationship between model accuracy (Cα RMSD) and alignment accuracy between PRO-models of all-α (dark filled circle), all-β (gray filled circle) proteins. The percent of models with certain degree of alignment accuracy is shown as histogram (black bar, all-α; gray bars, all-β).

Fig. 3.

Alignment accuracy and profile entropy: **(a)** Comparison of the relationship between alignment accuracy and sequence identity between all- α (black), all- β (gray) proteins. Inset is the comparison between all- α , $\alpha\beta$ (dotted), and all- β proteins. **(b)** Procedure for computing entropy of a sequence profile. **(c)** Comparison of the distribution of entropy of profiles between PRO-models of all- α (dark filled circle), $\alpha\beta$ (empty circle), and all- β (gray filled circle) proteins.