2-8-2011

Systematic assessment of accuracy of comparative model of proteins belonging to different structural fold classes

Subrata Chakrabarty
South Dakota State University

Dario Ghersi
University of Nebraska at Omaha, dghersi@unomaha.edu

Roberto Sanchez
Mount Sinai School of Medicine

Systematic assessment of accuracy of comparative model of proteins belonging to different structural fold classes

By: Suvobrata Chakravarty, Dario Ghersi, and Roberto Sanchez

Abstract

In the absence of experimental structures, comparative modeling continues to be the chosen method for retrieving structural information on 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, αβ, all-β) of a protein may influence the accuracy of its model. In the twilight zone (sequence identity ≤ 30%) and at a similar level of target-template divergence, the accuracy of protein models does indeed follow the trend all-α > αβ > all-β. This is mainly because the alignment accuracy follows the same trend (all-α > αβ > all-β), 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 the accuracy of alignments/models to be estimated a priori in a class-dependent manner. This study provides a systematic description of and quantifies the structural class-dependent effect in comparative modeling. The study also suggests that datasets for large-scale sequence/structure analyses should have equal representations 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], the structures of many proteins, 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 the identification of inhibitor/antagonists [15–17], virtual screening [18, 19], molecular replacement [20], and function prediction [21].

In spite of these successful applications, predicted structures generally contain errors and seldom reach the accuracy of experimental structures. Three variables influence the accuracy of a comparative model [22]: (i) the structural similarity between the 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]:

$$\text{Total Error} = \text{Structural Difference} + \text{Alignment Error} - \text{Refinement} \quad (1)$$

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 of 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 αβ) [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 [7, 26–29] . 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 χ/β 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 the “Electronic supplementary material”). The selected pairs (size: 100–150 residues) were then sorted into sequence identity bins (10–60% with a bin size of 5%) such that each bin had the same number of all-α, all-β and αβ 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 from 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 the all-α, all-β and αβ pairs. A total of 4,500 sequence pairs were finally selected by choosing 1,500 pairs for each 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 when selecting pairs to ensure that the overall structural deviations of the selected pairs, irrespective of the class, were 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 nonredundant (NR) database was searched separately with each of the target and template sequences as a query using PSI-BLAST with 20 iterations [31]. The PSI-BLAST hits were sorted in 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

The overall accuracy of models was computed by measuring the root mean square deviation between the equivalent Cα atoms in the optimal superposition of target and model structures. Equivalent atoms were defined as those that were within 3.5 Å 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 Cα atom coordinates. All the calculations are implemented with 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 was defined as the ratio of the number of correctly aligned positions to the number of aligned positions of a given alignment. A target– template aligned residue pair of a given alignment was said to be a correctly aligned position if the particular pair was also aligned in the STR alignment [24]. The alignment accuracy used in this study refers to Q_{mod} [24]. Q_{mod} was 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 accuracies of models of individual protein domains (~150 residues) belonging to all three structural classes are 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 other studies [7], all comparisons of model accuracy are discussed as a function of template–target sequence identity. The respective average accuracies of the SEQ, PRO and STR models of the combined (all-α + αβ + all-β) dataset are shown for reference purposes (Fig. 1b). As previously shown, model quality increases in the order SEQ < PRO < STR (Fig. 1b), due to the increase in the alignment quality. Both the SEQ (Fig. 1a, left) and PRO (Fig. 1a, right) models of domains belonging to the all-α class are observed to be more accurate than the SEQ and PRO models of the αβ and all-β classes, respectively. In general, the accuracy was observed to decrease in the order all-α > αβ > all-β, indicating a possible influence of backbone geometry on model quality. The behavior of the αβ category of models is close to that of the combined averages of Fig. 1b, whereas those of the all-α and all-β models are respectively better and worse than the group average. Since the lengths of the modeled segments (domain size) in each of the three classes are very similar (∼150 residues), the effect of protein size on the observed difference can be discounted. The RMSD distributions for the structural superposition of pairs, irrespective of the class, were also nearly identical for all sequence identity bins. This selection criterion was important for the study, because we wanted to see the effect of alignment accuracy given the same level of structural divergence (see Eq. 1). Hence, the origin of the difference in the model accuracy was not due to a bias such as selecting pairs that were 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 the STR model curve, Fig. 1b) than the other 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 The 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 to the observed difference in model accuracy (see “Electronic supplementary material” for details). If this is indeed the case, models of proteins from different structural classes will show significant differences in model accuracy (Cα-RMSD) at the same level of alignment accuracy (irrespective of target–template sequence similarity). Figure 2 shows that in fact there is a difference between the model accuracies 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 in only a small number of cases. It also indicates that, even though on average different small isolated structural segments (see Fig. S1 of the “Electronic supplementary material”) respond differently (Cα-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 the fraction of models with a certain level of alignment accuracy (Fig. 2) shows that proteins belonging to the all-α category have a relatively small proportion of models with poor alignments and a relatively high proportion of accurate alignments than the all-β class. The distribution indicates that the alignment accuracy of proteins of the all-α category is higher than that for the all-β proteins, while the αβ category has an intermediate level of alignment accuracy (Fig. 3a, only PRO models are shown). Like model accuracy, the accuracy of prediction for a number of structure-derived properties (SDPs), such as inter-

residue distance, residue neighborhood, etc., varies substantially between the structural classes (see the "Electronic supplementary material" 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 the quality of information involving all-α sequences might be different than that of the rest. The information used when 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 utilized entropy as a measure of variability in profiles by counting the average number of different symbols in a profile column [33–35]. The 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. The 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 the profiles in each category. A higher average value of entropy implies that each column has a 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) < αβ (2.209 ± 0.557) < all-β (2.303 ± 0.517). The profiles of all- α proteins tend to have lower entropy, indicating that there is on average less variability among amino acids in a profile column of all-α than in the all-β case. An earlier study on protein families 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 homologs 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 αβ; αβ 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 and 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 a similar size, was used to compute a per residue entropy measure. In the study, ribosomal protein L7/L12 C-terminal domain (1ctf, αβ 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 the 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 profile entropies of all- β and αβ proteins compared to that for the all-α class could be explained by a wider spectrum of different amino acids than 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, αL strand C-cap, buried strand) [41–43]. Even a simple binary pattern (hydrophobic polar, HP) analysis 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 > αβ > all-β proteins.

Shakhnovich et al. showed that the number of sequences in a domain family depend 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 greater the chance that the rest of the sequence can change without destabilizing the fold. Simply put, if the CD for a given topology is high, there are more sequences that adopt the topology [45]. The distribution of the CDs for all-α class structures show smaller values compared to that for the αβ class (see Fig. S4 of the “Electronic supplementary material”), indicating that CD may play a role. However, all-β class proteins showed the smallest CD, which is not consistent with our observation. This could be because 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]), we noted that the much superior accuracies for small proteins compared to those for larger ones was due to the overrepresentation of all-α proteins in the small proteins set, whereas the three groups were equally well represented in the medium- and large-sized protein sets. This underscores the need for large-scale analyses of models and alignments [33, 46] to provide equal representations of proteins from all three classes in order to avoid classdependent 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 α+β (RRM-like fold) classes, along with β-class members such as the doublestranded β-helix (DSBH) domain and the β-propellers [47]. The α-class does not include dominant enzymatic folds [47]. This suggests that, in the absence of experimental structures, modeled enzyme structures should be used for highresolution structural feature analyses with great caution due to the high errors associated with nonhelical 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 model accuracy dependent only on alignment accuracy and target–template structural

difference (see Eq. 1). In order to examine the influence of software on alignment accuracy, alternate alignments were constructed using the program MUSCLE [48], and these also showed clear class-dependent alignment accuracy (data not shown), further supporting the software-independent nature of our results.

Conclusions

As a continuation of our characterization of simple comparative models [7, 8, 23], the results of this largescale (~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 accuracies, 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 factor contributing to 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.

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 Fold and Function Assignment (FFAS) server when investigating the test cases of profile–profile alignments. SC thanks Prof. Ming-Ming Zhou for encouragement. The study was supported by the National Institute of General Medicine at the National Institutes of Health [grant 1R01GM081713 (RS)], and South Dakota State University's (SDSU) Agricultural Experiment Station and Center for Biological Control and Analysis by Applied Photonics (BCAAP) [grant 3SG163 (SC)].

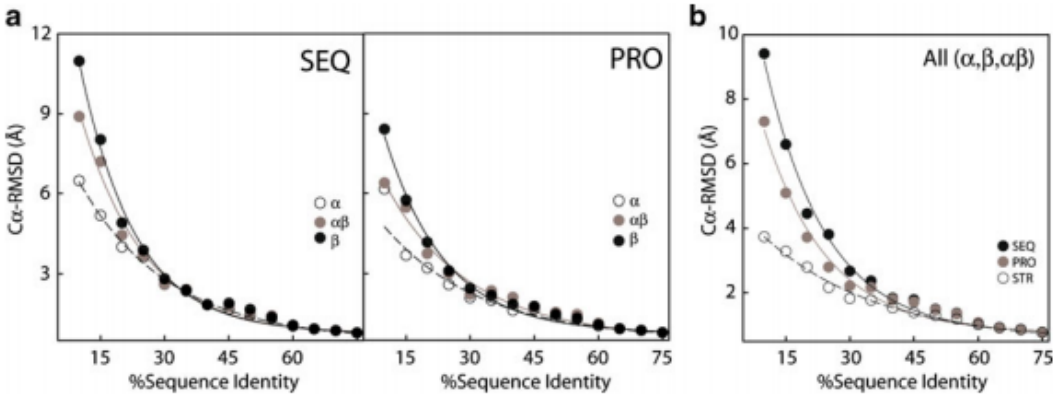

Fig. 1 Accuracies of comparative models: **a** comparison of Cα RMSD between experimental and modeled structure for all-α, all-β and αβ proteins (SEQ models on the *left* and PRO models on the *right*). **b** Comparison of Cα RMSDs between modeled experimental structures for SEQ, PRO and STR models

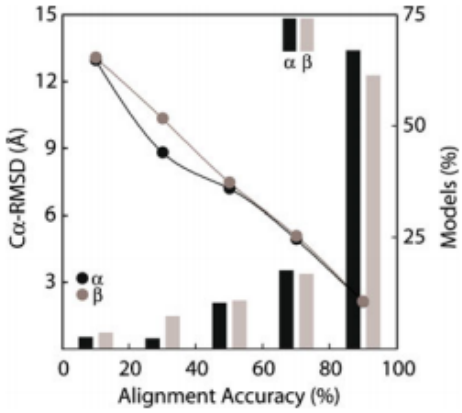

Fig. 2 Comparison of the relationship between model accuracy (Cα RMSD) and alignment accuracy between PRO models of all-α and all-β proteins. The percentage of models with certain degree of alignment accuracy is shown as a histogram

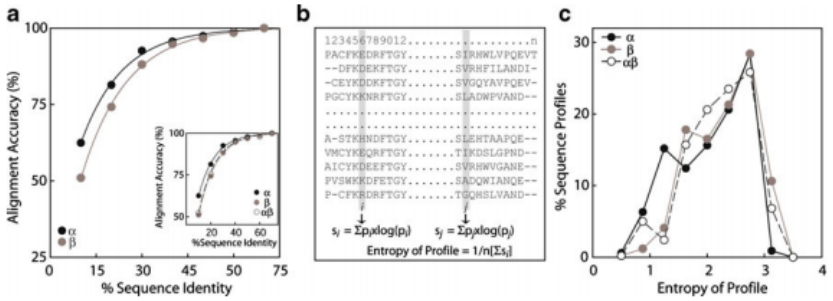

Fig. 3 Alignment accuracy and profile entropy. **a** Comparison of the relationship between alignment accuracy and sequence identity between all-α (*black*) and all-β (*gray*) proteins. The *inset* shows the comparison between all-α, αβ (*dotted*), and all-β proteins. **b** Procedure for computing the entropy of a sequence profile. **c** Comparison of the profile entropy distributions of PRO models of all-α (*dark filled circles*), αβ (*empty circles*), and all-β (*filled gray circles*) proteins

References

1. Taylor WR (2007) Evolutionary transitions in protein fold space. Curr Opin Struct Biol 17:354–361

2. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci USA 95:13597–13602

3. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A (2000) Protein structure modeling for structural genomics. Nat Struct Biol 7(Suppl 1):986–990

4. Stevens RC, Yokoyama S, Wilson IA (2001) Global efforts in structural genomics. Science 294:89–92

5. Tramontano A, Morea V (2003) Assessment of homology-based predictions in CASP5. Proteins 53(Suppl 6):352–368

6. Lushington GH (2008) Comparative modeling of proteins. Meth Mol Biol Clifton NJ 443:199–212

7. Chakravarty S, Wang L, Sanchez R (2005) Accuracy of structurederived properties in simple comparative models of protein structures. Nucleic Acids Res 33:244–259

8. Chakravarty S, Sanchez R (2004) Systematic analysis of addedvalue in simple comparative models of protein structure. Struct Camb 12:1461–1470

9. Kiel C, Wohlgemuth S, Rousseau F, Schymkowitz J, FerkinghoffBorg J, Wittinghofer F, Serrano L (2005) Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations. J Mol Biol 348:759–775

10. Liu T, Rojas A, Ye Y, Godzik A (2003) 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 12:1872–1881

11. Murray PS, Li Z, Wang J, Tang CL, Honig B, Murray D (2005) Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. Structure 13:1521–1531

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 (2002) The Protein Data Bank. Acta Crystallogr D 58:899–907

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

14. Hillisch A, Pineda LF, Hilgenfeld R (2004) Utility of homology models in the drug discovery process. Drug Discov Today 9:659–669

15. Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc Natl Acad Sci USA 90:3583–3587

16. Evers A, Klabunde T (2005) Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J Med Chem 48:1088–1097

17. Evers A, Klebe G (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. J Med Chem 47:5381–5392

18. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J Med Chem 46:2656–2662

19. Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Novel technologies for virtual screening. Drug Discov Today 9:27–34

20. Read RJ (2001) Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D 57(Pt 10):1373–1382

21. Skolnick J, Fetrow JS, Kolinski A (2000) Structural genomics and its importance for gene function analysis. Nat Biotech 18:283–287

22. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15:285–289

23. Chakravarty S, Godbole S, Zhang B, Berger S, Sanchez R (2008) Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. BMC Struct Biol 8:31

24. Sauder JM, Arthur JW, Dunbrack RL Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40:6–22

25. Dunbrack RL Jr (2006) Sequence comparison and protein structure prediction. Curr Opin Struct Biol 16:374–384

26. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96

27. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 69(Suppl 8):38–56

28. Nayeem A, Sitkoff D, Krystek S Jr (2006) A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. Protein Sci 15:808–824

29. Rayan A (2009) New tips for structure prediction by comparative modeling. Bioinformation 3:263–267

30. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

31. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

32. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9:232–241

33. Sadreyev RI, Grishin NV (2004) Estimates of statistical significance for comparison of individual positions in multiple sequence alignments. BMC Bioinf 5:106

34. Panchenko AR (2003) Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Res 31:683–689

35. Casbon J, Saqi MA (2005) S4: structure-based sequence alignments of SCOP superfamilies. Nucleic Acids Res 33:D219–222

36. Tress ML, Jones D, Valencia A (2003) Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 330:705–718

37. Sadreyev RI, Grishin NV (2004) Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. Bioinformatics 20:818–828

38. Koehl P, Levitt M (2002) Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 99:1280– 1285

39. England JL, Shakhnovich EI (2003) Structural determinant of protein designability. Phys Rev Lett 90:218101

40. Minor DL Jr, Kim PS (1994) Context is a major determinant of beta-sheet propensity. Nature 371:264–267

41. Han KF, Baker D (1995) Recurring local sequence motifs in proteins. J Mol Biol 251:176–187

42. Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA 93:5814–5818

43. Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence– structure correlations in proteins. Curr Opin Biotechnol 7:417–421

44. West MW, Hecht MH (1995) Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci 4:2032–2039

45. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E (2005) Protein structure and evolutionary history determine sequence space topology. Genom Res 15:385–392

46. Edgar RC, Sjolander K (2004) A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 20:1301–1308

47. Anantharaman V, Aravind L, Koonin EV (2003) Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 7:12–20

48. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf 5:113