

# Systematic Analysis of Added-Value in Simple Comparative Models of Protein Structure

Suvobrata Chakravarty and Roberto Sanchez\*

Structural Biology Program  
Department of Physiology and Biophysics  
Mount Sinai School of Medicine  
New York, New York 10029

## Summary

**Added-value is the additional information that a model carries with respect to the template structure used for model building. Thousands of single-template models, corresponding to proteins of known structure, were analyzed. The accuracy of structure-derived properties, such as residue accessibility, surface area, electrostatic potential, and others, was determined as a function of template:target sequence identity by comparing the models with their corresponding experimental structures. Added-value was determined by comparing the accuracy in models with that from templates. Geometry-dependent properties such as neighborhood of buried residues and accessible surface area showed low added-value. Properties that also depend on the protein sequence, such as presence of polar areas and electrostatic potential, showed high added-value. In general added-value increases when template:target sequence identity decreases, but it is also affected by alignment errors. This study justifies the use of models instead of the use of templates to estimate structure-derived properties of a target protein.**

## Introduction

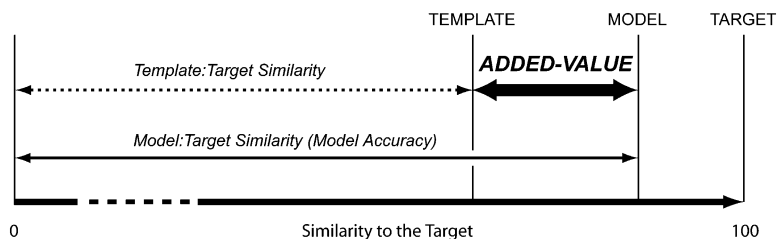
The discovery of large numbers of new gene sequences, by genome sequencing projects and by more traditional methods, presents an opportunity and a challenge to understand the function of the encoded proteins individually and in the context of each other. Full understanding of the biological role of these proteins requires knowledge of their three-dimensional (3D) structure and biochemical function. The 3D structure of a protein generally provides more information about its function than its sequence alone because patterns in space are frequently more recognizable than patterns in sequence (Sanchez and Sali, 1998). Different types of information can be derived from protein structures, such as overall shape and volume; electrostatic and hydrophobic properties of the surface; presence of pockets and cavities; residue-residue contacts; and salt bridges, disulphide bridges, and others (S. Chakravarty et al., submitted). These structure-derived properties describe the similarities and differences between proteins and therefore are valuable in understanding how they function. This type of analysis is critical when we try to understand differences between homologs in different tissues or organ-

isms, or between polymorphisms of a particular gene. Ideally, comparative studies of protein structure should use a complete set of proteins. For example, if one wants to understand differences in substrate specificity through the study of structural differences in a family of enzymes, one would use structures for all members of the family. Similarly, if properties of proteins from thermophiles and mesophiles are being compared to identify the structural basis of thermal adaptation, one would like to compare structures of as many pairs as possible of orthologs from mesophiles and thermophiles (Chakravarty and Varadarajan, 2002). Unfortunately, because the number of known protein sequences is an order of magnitude larger than the number of known protein structures, in most cases complete sets of experimental structures are not available to answer such questions. In such a situation using protein structure prediction is necessary to obtain the kind of structure-derived information described above.

Many approaches to protein structure prediction methods have been developed, of which comparative modeling is currently the most accurate (Fischer et al., 2003; Tramontano and Morea, 2003; Venclovas et al., 2003). Comparative modeling (CM) uses experimentally determined protein structures (templates) to predict the 3D conformation of another protein with a similar amino acid sequence (target). Its applicability is limited by the requirement of a template structure, but in spite of this limitation, it is possible to model at least one domain in more than half of the known protein sequences (Pieper et al., 2002). CM is particularly well suited for the kind of studies described before, where comparison of similar proteins is the focal point. Since CM uses one or more experimental structures as templates to model a target protein of unknown structure, it is by definition a method that can be used to leverage experimental information to extend structural information to complete families of proteins. As CM predicts the structure of a protein based on a template of known structure, it generally does not provide new information about the primary function of a protein. The reason for this is that conservation of the fold usually implies conservation of function (Thornton et al., 2000; Wilson et al., 2000), and conservation of the fold is a requirement for CM. Hence, it is valid to ask what new information a comparative model can provide about a protein. The additional information that a model carries with respect to the template structure used for model building is defined as the *added-value* of the model. The added-value can be calculated by comparing the model and template structures with the experimental target structure. The model:target comparison determines the model accuracy and the template:target comparison determines the template:target similarity. The difference between model accuracy and template:target similarity is the added-value of the model (Figure 1).

Several anecdotal examples have shown that comparative models sometimes contain features that are not present in their templates, providing new information

\*Correspondence: roberto@sanchezlab.org



$$ADDED-VALUE = (Model Accuracy) - (Template:Target Similarity)$$

Figure 1. Definition of Added-Value

The accuracy of a model is defined by how close it is to its target. This is measured by comparing the model and target structures (model:target similarity). The template used to build the model also shows some level of similarity with respect to the target (template:target similarity, dotted line). The added-value of a model indicates how much closer it is to the target when compared to the template. This is calculated by subtracting the template:target similarity from the model accuracy.

that is not easily or directly derivable from the template structure (Sali et al., 1993; Sanchez and Sali, 1998; Xu et al., 1996), but no systematic study has been carried out to address this question. The present work systematically addresses the question of added-value in comparative modeling.

## Results and Discussion

The added-value for the following structure-derived properties (SDPs) was analyzed in single-template models: (i) overall accuracy, (ii) exposure state of residues, (iii) neighborhood of residues, (iv) accessible surface area, (v) identification of surface pockets, (vi) composition of surface pockets, and (vii) electrostatic potential. When measuring the *accuracy* of a property in a model, the value of the property derived from the model is compared with the value obtained from its corresponding experimental structure (target). The template:target difference is determined by comparing the value of the property derived from the template with the value obtained from target (see Experimental Procedures). The *added-value* of the models is determined by comparing the accuracy of the models with the template:target difference (Figure 1). Thus, added-value always has the same units as the accuracy. The comparison of models based on template:target pairwise sequence alignments (SEQ models) and structure-based alignments (STR models) provides an indication of the effect of alignment errors on added-value. Models built using the structure-based alignments represent the accuracies (overall or SDP based) of a model in the absence of alignment errors. The accuracy and added-value of SDPs is shown as a function of sequence identity of the template:target alignment as it is the most commonly referred variable in CM (Marti-Renom et al., 2000). Although models of three size classes were constructed and analyzed (see Experimental Procedures), only the results for medium sized proteins (8208 models) are shown here as they are the most representative and generally do not show large differences with the models in the other two classes.

### Overall Accuracy

Overall accuracy is measured by the percentage of equivalent atoms within 3.5 Å of each other in the optimal superposition of the model and the target experimental structure. This is a simple and common evaluation that has been performed systematically for comparative

models (Eyrich et al., 2001; Marti-Renom et al., 2002; Sanchez and Sali, 1998). Figure 2A shows the change in percentage of equivalent C $\alpha$  atoms as a function of template:target sequence identity for models built using sequence-based alignments (SEQ models, closed circles) and their corresponding templates (open triangles). As expected, a trend of increasing accuracy with higher template:target sequence identity is observed. There is a sharp decrease in the accuracy below 35% sequence identity, as previously reported (Eyrich et al., 2001; Sanchez and Sali, 1998). There is no difference between the accuracy of SEQ models and their templates, indicating that there is no added-value when measuring the accuracy of the models by the percentage of equivalent C $\alpha$  atoms. The same result, lack of added-value, is observed in Figure 2B when models built using structure-based alignments (STR models, open circles) are compared with their templates (open triangles). When all heavy atoms are included in the calculation of percentage equivalent atoms (Figures 2C–2E), there is a range of template:target sequence identities for which added-value is observed. This is explained by the fact that the model has the same sequence as the target. At high template:target sequence identity there is little sequence difference between the template and the model, and since unrefined models follow their templates closely, it is not possible for the model to provide new information. As the template:target sequence identity decreases, the template contains less heavy atoms that are identical with the target's atoms. The same is not true for the model, resulting in more equivalent heavy atoms in the model than the template. As the template:target sequence identity decreases even more (<30%), the added-value decreases again (Figure 2E). Because there is virtually no difference between the added-value of models built using sequence-based (SEQ) and structure-based (STR) alignments, alignment errors can not be the reason for the decrease in added-value. Thus the decrease in added-value must be due to the larger structural differences between template and target below 30% sequence identity. At this level of sequence similarity the packing of residues changes significantly and the template is not such a good representation of the target's backbone anymore (Chung and Subbiah, 1996).

### Residue Exposure State

Exposure state of a residue, i.e., if a residue is exposed, intermediate, or buried, is decided based on its solvent

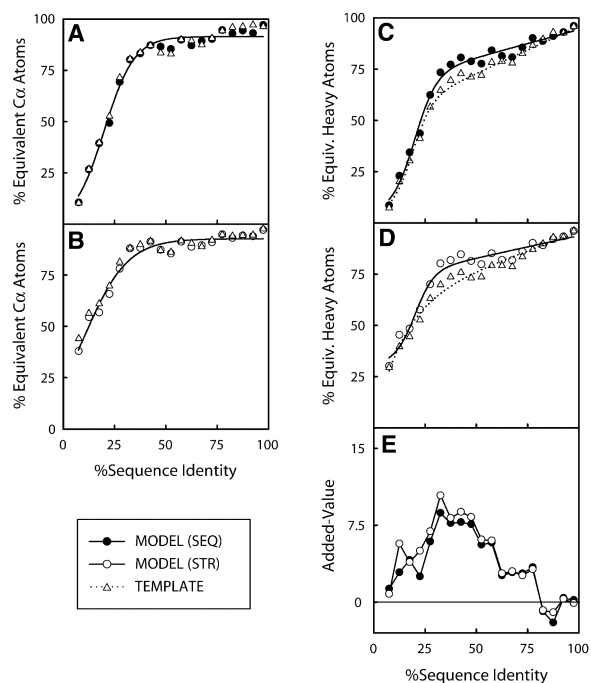


Figure 2. Overall Accuracy

Overall model accuracy is shown as a function of template:target sequence identity. The overall accuracy is measured by the percentage of equivalent atoms between the model (or template) and the target after superposition of their structures. The structural superposition follows the alignment used in model building.

(A) Percentage of equivalent C $\alpha$  atoms for models built using sequence-based alignments (SEQ, closed circles) and their templates (open triangles).

(B) Percentage of equivalent C $\alpha$  atoms for models built using structure-based alignments (STR, open circles) and their templates (open triangles).

(C) Percentage of equivalent heavy atoms for SEQ models (closed circles) and their templates (open triangles).

(D) Percentage of equivalent heavy atoms for STR models (open circles) and their templates (open triangles).

(E) Added-value of percentage equivalent heavy atoms for SEQ and STR models. The added-value is calculated by subtracting the percentage equivalent atoms of the template from the percentage equivalent atoms of the model in (C) and (D).

accessibility (see Experimental Procedures). Residues accessible to the solvent are generally responsible for the interactions of a protein with other molecules, thus determining its biochemical function. For this reason, protein structures are frequently used to determine which residues are exposed to the solvent and that information is used in applications such as site-directed mutagenesis, subcellular localization prediction, and protein design. The prediction accuracy of exposure state, which represents the probability that a residue that is exposed (or buried) in the model is also exposed (or buried) in the experimental target structure, increases with template:target sequence identity (Figures 3A–3D). Exposure state predicted from models built on sequence-based alignments (SEQ models) is only slightly more accurate than the exposure state calculated using the templates, both for exposed and buried residues (Figures 3A and 3B). When comparing models built on structure-based alignments (STR models) with

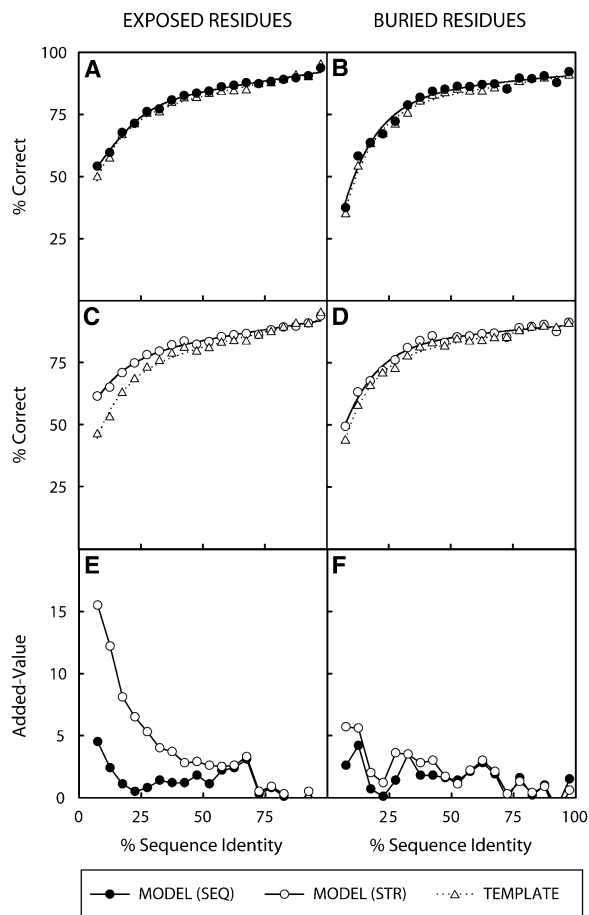


Figure 3. Residue Exposure State

Exposure state indicates if a residue is exposed or buried.

(A) Percentage of correctly predicted exposed residues for models built using sequence-based alignments (SEQ, closed circles) and their templates (open triangles).

(B) Percentage of correctly predicted buried residues for SEQ models (closed circles) and their templates (open triangles).

(C) Percentage of correctly predicted exposed residues for models built using structure-based alignments (STR, open circles) and their templates (open triangles).

(D) Percentage of correctly predicted buried residues for STR models (open circles) and their templates (open triangles).

(E and F) Added-value of exposure state for exposed (E) and buried (F) residues in SEQ and STR models. The added-value is calculated by subtracting the percentage correctly predicted exposed residues of the template from the percentage correctly predicted exposed residues of the model in (A) and (C) (E), and (B) and (D) (F).

their templates, the behavior for exposed and buried residues is different. While exposed residues show added-value, particularly below 50% sequence identity (Figures 3C and 3E), for buried residues the added value is almost the same as in SEQ models (Figures 3D and 3F). Exposed residues of STR models show a clear trend of increasing added-value with decreasing template:target sequence identity (Figure 3E).

### Residue Neighborhood

Information about neighborhood of a particular residue is obtained from the contacts it makes with its neigh-

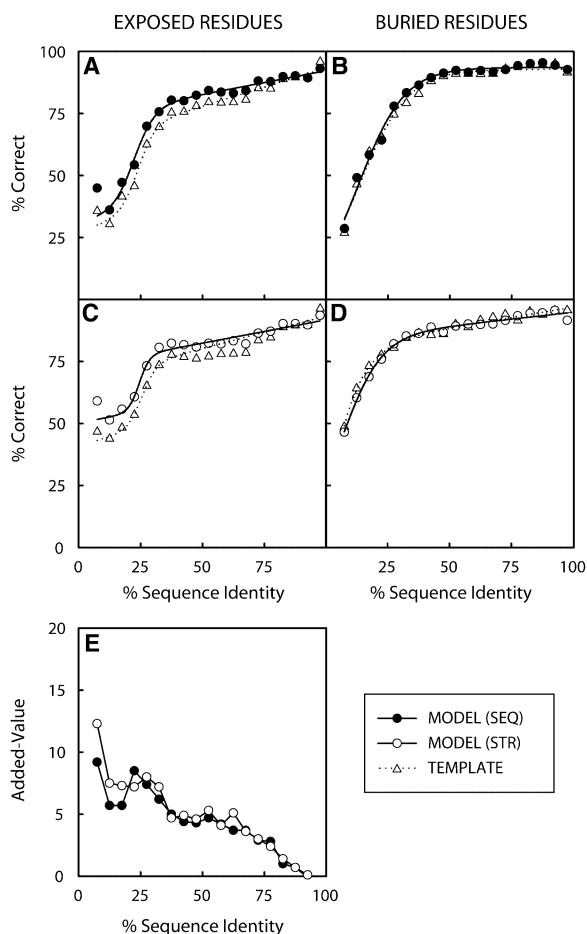


Figure 4. Residue Neighborhood

Neighborhood of a residue is defined as the list of residues in van der Waals contact with it.

(A) Percentage of correctly predicted neighbors of exposed residues for models built using sequence-based alignments (SEQ, closed circles) and their templates (open triangles).

(B) Percentage of correctly predicted neighbors of buried residues for SEQ models (closed circles) and their templates (open triangles).

(C) Percentage of correctly predicted neighbors of exposed residues for models built using structure-based alignments (STR, open circles) and their templates (open triangles).

(D) Percentage of correctly predicted neighbors of buried residues for STR models (open circles) and their templates (open triangles).

(E) Added-value of neighborhood for exposed residues in SEQ and STR models. The added-value is calculated by subtracting the percentage correctly predicted neighbors of the template from the percentage correctly predicted neighbors of the model in (A) and (C).

bors. Neighborhood of residues is routinely used for rational design of mutants, biochemical labeling (attaching a fluorophore or a spin label), incorporation of disulphide bridges, and in protein design. The list of neighbors of each residue in the model was compared with that of its corresponding residue in the experimental target structure (see Experimental Procedures). For buried residues there is no added-value for neighborhood (Figures 4B and 4D) irrespective of the alignment used, while for exposed residues the models clearly provide added-value (Figures 4A, 4C, and 4E). The added-value for neighborhood of exposed residues in-

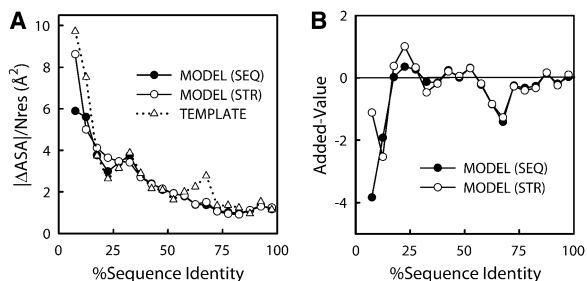


Figure 5. Accessible Surface Area

The error in the accessible surface area (ASA) of models and templates is determined by comparing the model and template ASA with that of the corresponding target experimental structure.

(A) Per residue ASA error,  $|ASA_{model} - ASA_{exp}|/N$ , as a function of template:target sequence identity for models built on sequence-based alignments (SEQ, closed circles), models built on structure-based alignments (STR, open circles), and templates (open triangles).

(B) Added-value of ASA in SEQ models (closed circles) and STR models (open circles). The added-value is calculated by subtracting the template ASA error from the model ASA error. Note that for ASA error a negative added-value corresponds to an improvement in the ASA estimate (decrease of the error).

creases with decreasing template:target sequence identity for SEQ and STR models (Figure 4E). A possible explanation for the added-value observed for exposed residues is that the identification of neighbors depends on the actual side chain, particularly its size. Because the model and its corresponding experimental structure have identical residues at every position, there would not be an effect of residue size, which is not the case with the templates. For example, a pair of neighboring bulky residues in the target structure may correspond to a pair of "nonneighbor" small residues in the template structure. The model, by virtue of having identical bulky residues, would have a better chance of detecting them as neighbors. This effect is smaller for buried residues because they are more conserved than exposed residues, and probably also because the interior is better packed than the surface.

#### Accessible Surface Area

The value of the total accessible surface area (ASA) of a protein is frequently used in calculation of protein stability or oligomerization state (Livingstone et al., 1991; Spolar and Record, 1994). We use the average per-residue ASA difference  $\Delta ASA/N$  ( $\text{\AA}^2$ ) =  $|ASA_E - ASA_M|/N$  as a measure of the error (Figure 5A);  $ASA_M$  and  $ASA_E$  are the total surface area of a model and its corresponding experimental structure, respectively; and  $N$  is the number of residues in the protein. Above 20% sequence identity there is little difference between the ASA values calculated from models and templates. Below 20% sequence identity the models provide some added-value for ASA (Figure 5B), probably due to differences in size between homologs at this level of sequence similarity.

#### Surface Pockets

Protein function, such as binding of a ligand, is frequently mediated by surface pockets. The comparison of detection and volume of surface pockets in models

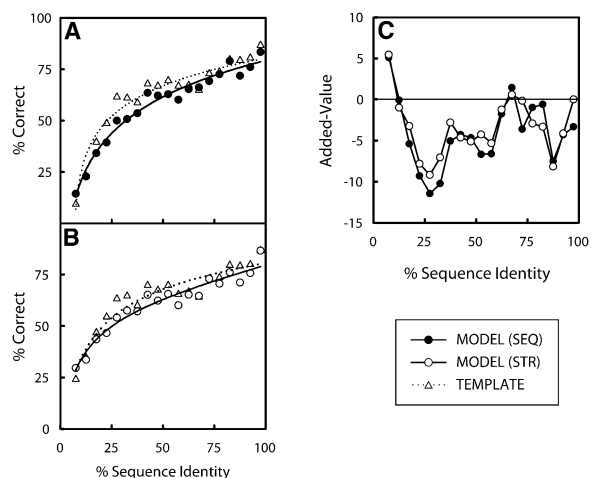


Figure 6. Surface Pockets

Accuracy of detection of surface pockets is shown as a function of template:target sequence identity.

(A) Percentage of correctly predicted pockets is compared between SEQ models (closed circles) and their templates (open triangles). (B) Percentage of correctly predicted pockets is compared between STR models (open circles) and their templates (open triangles). (C) Added-value of surface pocket detection in SEQ (closed circles) and STR (open circles) models. The added-value is calculated by subtracting the percentage correctly predicted pockets of the template from the percentage correctly predicted pockets of the model in (A) and (B).

with respect to the experimental structures was attempted. Preliminary data showed that volumes of identical pockets of even very close homologs show a large variation. For example the volume of the central lipid binding cavity in the family of fatty acid binding protein (FABP) showed large variation even between very close homologs (data not shown). This difference is however not due to changes in substrate specificity, but due to widening or narrowing of the mouth of the central pocket as a result of side chain orientation, indicating that the estimate of pocket volumes is intrinsically noisy. Hence, for this study only the identification and location of pockets are dealt with and comparison of pocket size is avoided. PASS (putative active sites with spheres) was used for identification of pockets (see Experimental Procedures). PASS reports coordinates of grid points representing putative active site ligands. Residues in contact with these grids define the pocket boundary. Identity of a pocket in a model with that of an experimental structure is established by comparing the list of boundary residues of the pockets (see Experimental Procedures). One-third of the pockets were large, with ten or more boundary residues; the remaining two-thirds had fewer than ten boundary residues. We looked at large pockets because the largest pocket in a protein is most often the biological active site (Liang et al., 1998). The accuracy of detection of a pocket is the ratio between the number of identical pockets (see Experimental Procedures) and the total number of pockets in a model (see above). Over most of the template:target sequence identity range the template-based estimate of pockets is better than that of models (Figures 6A and 6B) indicating that there is negative added-value in models for this property (Figure

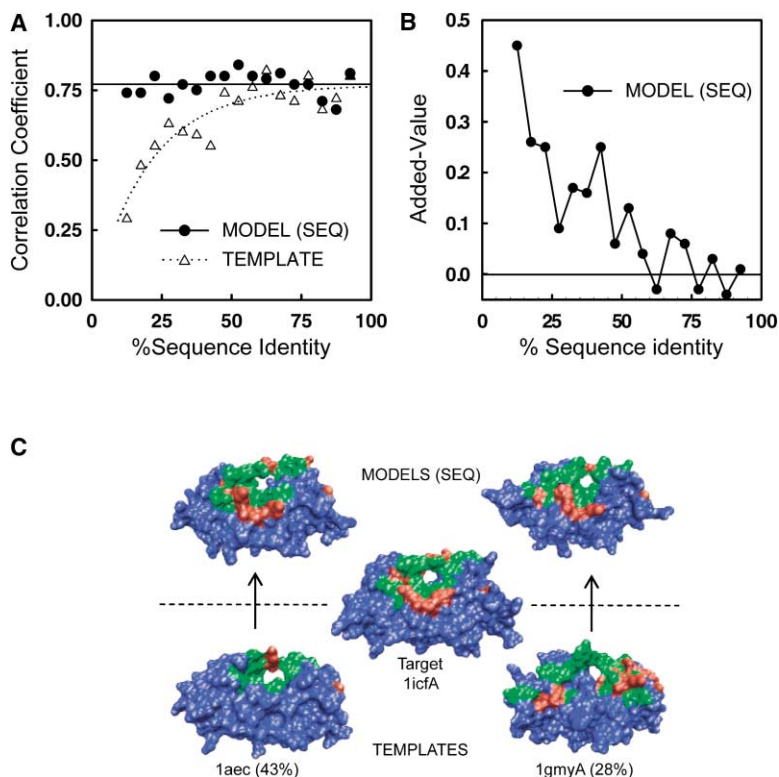
6C). At very low sequence identity the models show some added-value (Figure 6C). The main reason for the low accuracy and added-value of pocket detection is the large number of false pockets present in models, which increases the total number of pockets (S. Chakravarty et al., submitted). Alignment accuracy has very little effect on the added-value of pocket detection as illustrated by the SEQ versus STR comparison (Figure 6C).

### Composition of Surface Pockets

Since the residues of models are identical to those of their corresponding experimental structures, the physicochemical properties of a pocket in a model should be very close to those of the experimental target structure, but those of a template would be different due to dissimilarity between residue types. This feature is qualitatively highlighted by looking at the correlation coefficient between the number of charged atoms of pockets in models and experimental structures for equivalent pockets (Figure 7A). The correlation coefficient has a constant value for models but varies significantly with sequence identity for templates. This would qualitatively mean that the location of charged atoms in models is very similar to that of the experimental target structures in case of equivalent pockets; hence, the properties of pockets derived from a model are expected to be more accurate than those derived from the template. Most properties discussed so far showed that unrefined single-template models in general did not provide much added-value over the template, but this feature, though qualitative, certainly establishes a large accuracy advantage of models over templates, as illustrated in Figure 7B. This is not surprising since physicochemical properties such as electrostatics and hydrophobicity depend not only on the conformation of the protein (geometry) but also on the specific chemical groups that are present, which depend on the protein sequence. As the template:target sequence similarity drops, the template becomes a worse representation of the chemical groups present in the target protein. By combining the geometry provided by the template with the correct chemical groups (amino acids) from the target sequence, comparative modeling adds value to the template.

### Electrostatic Potential

Calculations of electrostatic potential (EP) in protein structures are frequently used to identify regions of positive or negative charge that may represent binding pockets or active sites (Nicholls et al., 1991; Sali et al., 1993). We calculated the accuracy of the electrostatic potential by comparing the three-dimensional grid resulting from the EP calculation of models and templates with the grid obtained from the experimental target structure (see Experimental Procedures). The EP similarity is measured by the correlation of the EP values in a pair of grids. Figure 8A shows the correlation coefficient obtained when comparing EP of models built on sequence-based alignments (SEQ, closed circles), models built on structure-based alignments (STR, open circles), and templates (open triangles). The EP accuracy drops with decreasing template:target sequence identity, but the



**Figure 7. Composition of Surface Pockets**  
**(A)** Linear correlation coefficient between the number of charged atoms per pocket (only for identical pockets) in the target experimental structure and model (closed circle) or template (open triangle).  
**(B)** Added-value of pocket composition for SEQ models. The added-value is calculated by subtracting the template correlation coefficient from the model correlation coefficient in (A).  
**(C)** Charged residue distribution in templates (bottom) and models (top) of Thyroglobulin type-1 domain binding pocket of cathepsin heavy chain (11cfA, middle). Charged residues are shown in red and the binding pocket in green. Sequence identities and PDB codes (with chain ID) of templates are indicated.

accuracy for models is always clearly higher than that of templates, indicating that EP is a property with high added-value in models. Figure 8B compares the added-value of SEQ and STR models. As the template:target sequence identity decreases, the added-value increases; this is a continuous trend for models built on structure-based alignments (STR). For models built on sequence-based alignments (SEQ) the added-value peaks at around 30% sequence identity and then starts to fall off (Figure 8B). This indicates that alignment errors affect the accuracy and added-value of the EP potential at low (< 30%) template:target sequence identity. The large added-value observed for EP is similar to the observation made for the composition of surface pockets in the previous section. The source of the added-value is the same in both cases, namely the combination of the geometry provided by the template with the correct target sequence. In this case the sequence provides the correct charges for the calculation of the electrostatic potential.

### Conclusions

In the absence of refinement (e.g., loop modeling) and multiple templates, the differences in accuracy between comparative modeling approaches is very small (Tramontano and Morea, 2003). As such, this study is representative of all comparative modeling methods in spite of using a single program (MODELLER) to construct the set of models. This study represents a baseline of added-value for comparative models against which more elaborate modeling procedures can be compared. It is also representative of the types of models produced by large-scale fully automated methods which usually

rely on automated alignments, single templates, and no refinement (Peitsch et al., 2000; Pieper et al., 2002; Sanchez et al., 2000; Sanchez and Sali, 1998).

The added-value of models is not the same for different structure-derived properties (SDPs). For SDPs that depend mostly on position of residues, such as exposure state and neighborhood of buried residues, and number of surface pockets, models do not provide added-value with respect to the template. At least that is the case for the simple set of models used here. For other SDPs, such as exposure state and neighborhood of exposed residues, and total ASA of low sequence identity models, models show some added-value. Finally, for properties that strongly depend on the physicochemical characteristics of the amino acids in the sequence, such as composition of pockets and electrostatic potential, models show large added-value. The lack of added-value for the first set of SDPs is not surprising, since in the absence of model refinement or multiple templates it is not possible for comparative modeling to deviate much from the template structure (Marti-Renom et al., 2000; Sanchez and Sali, 1998). Thus, the template geometry remains largely unchanged in the model.

The origin of added-value in models appears to have two sources. Some properties depend not only on the position of residues but also on the size of the residues (e.g., exposure state and neighborhood of exposed residues); because the model contains the same residues as the target it adds information on top of the positions provided by the template. Other properties depend more on the physicochemical characteristics of the side chains (e.g., composition of pockets and electrostatic

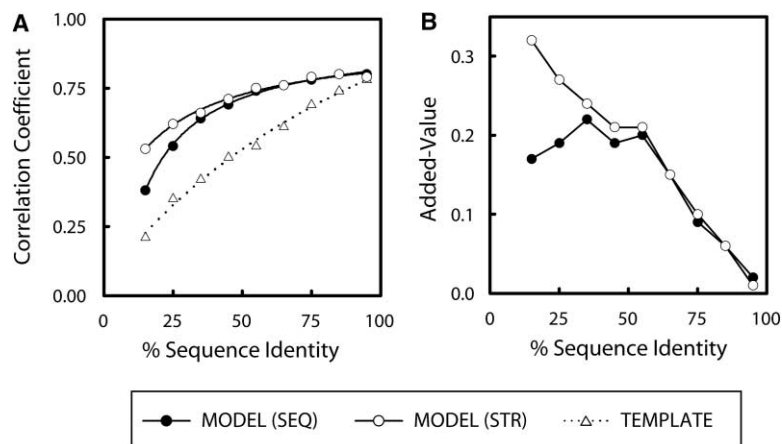
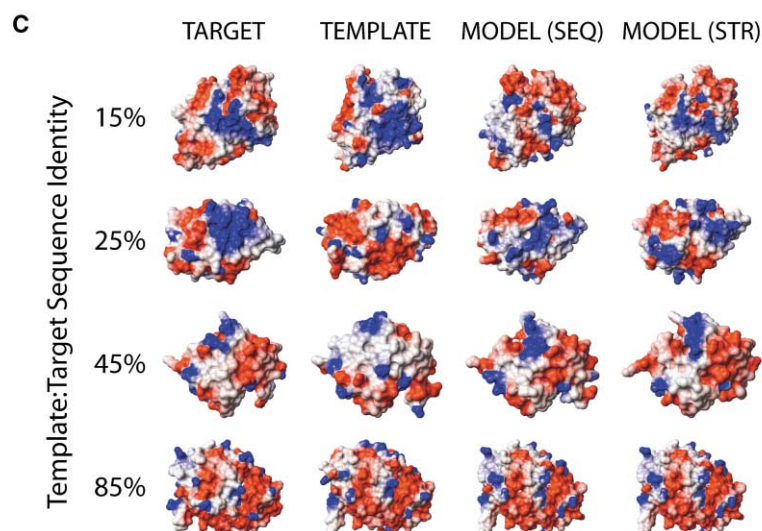


Figure 8. Electrostatic Potential

(A) Accuracy of electrostatic potential as a function of template:target sequence identity. The accuracy is measured by the rank correlation coefficient between the values of the electrostatic potential of the target and models (SEQ, closed circles; STR, open circles) or template (open triangles).

(B) Added-value of electrostatic potential for SEQ (closed circles) and STR (open circles) models. The added-value is calculated by subtracting the template correlation coefficient from the model correlation coefficient in (A).

(C) Electrostatic potential colored surfaces for selected targets, templates, and models. The examples were selected such that their correlation coefficients fall on the curves of (A).



potential); in this case the model provides the correct physicochemical properties to go along with the positions provided by the template. As template:target sequence similarity decreases, the template provides less information about the size and physicochemical properties of the residues in the target, explaining why added-value increases with decreasing template:target sequence identity. This illustrates the difference between accuracy and added-value. While the accuracy of all SDPs decreases with lower template:target sequence identity, their added-value generally increases, making the models relatively more informative in spite of their lower accuracy. This underlines the importance of improving the accuracy of models based on low (<30%) sequence identity templates, because it is where the most new information is generated. This overlaps with the fact that most real modeling cases fall in the 20%–30% identity range (Sanchez et al., 2000; Sanchez and Sali, 1998).

The accuracy and added-value of single-template models such as the ones used here can be improved by using multiple templates, which allows comparative modeling to select the best parts from different struc-

tures to build the target model (Sanchez and Sali, 1997). Anecdotal evidence indicates that this is the case (Sanchez and Sali, 1997), but no systematic study has been performed. This study is under way and preliminary data indicates that there is at least a small improvement when using multiple templates. Refinement of models, in the form of loop modeling (Fiser et al., 2000), should also provide an improvement over the simple models presented here. How much loop modeling affects the accuracy and added-value of different SDPs is not clear and is a question that will be addressed elsewhere as the computational cost of proper loop modeling is orders of magnitude larger than that of building the models used here. Because most loop modeling cases correspond to solvent-exposed insertions, it is expected that it will have an impact on properties related to the protein surface (exposure state, ASA).

This study shows that comparative models provide added-value by combining the “right sequence” with the “right template.” With the exception of the detection of pockets, even in the worst cases the models are at least as good as the templates, and for most properties they show some level of added-value. This justifies the

use of a model instead of the use of a template to estimate structure-derived properties of a target protein. The more a given property depends on the sequence of the protein the more useful a model will be in estimating the value of that property.

## Experimental Procedures

### Data Set

Chains of X-ray structures with resolution better than 2.5 Å were selected from the Protein Data Bank (PDB) (Berman et al., 2002). A representative set of these chains was selected by doing an all-against-all comparison of their sequences using BLAST (Altschul et al., 1997) and clustering into groups that had alignments with >95% sequence identity to each other and that covered at least 85% of the chain sequence. The highest resolution member of each group was retained. The representative chains were structurally aligned with each other using CE (Shindyalov and Bourne, 1998). Only alignments with a CE Z-score higher than 4.5 and covering at least 85% of one of the chains were retained for model building. The aligned segments were accepted as having the same fold. The aligned sequences were then sorted based on size into three nonoverlapping groups: small (50–100 residues), medium (150–200 residues), and large (≥250 residues). The alignments were classified into 18 groups based on sequence identity ranging from 10% to 100% with a bin size of 5%. The number of alignments for groups with lower sequence identity outnumbered those of groups with higher sequence identity. In order to have a relatively uniform distribution of number of alignments across all the 18 groups, approximately 200 alignments were selected at random from the groups with lower sequence identity so that each group or bin had a little over 5% of the total set of alignments. There are 1564, 911, and 856 unique chains of small, medium, and large proteins, respectively. The alignments of small, medium, and large proteins are respectively 4912, 4104, and 3716 in our data set.

### Model Building

The structure-based alignments produced by CE were used as input to program MODELLER version 6v2 (Eswar et al., 2003; Sali and Blundell, 1993) to construct a 3D model of the target sequence. The set of models based on CE alignments is called STR. A second set of models based on “poor” alignments were generated by realigning the sequences of the CE alignment using the ALIGN command of MODELLER; these are called SEQ models. This resulted in alignments based exclusively on sequence information, as opposed to the structure-based alignments generated by CE. Models were constructed using the default “model” routine in MODELLER. All alignments contained a single template and no loop modeling was performed. A total of 25,464 models were calculated; half of them based on SEQ alignments and the other half based on STR alignments.

### Model Accuracy, Template:Target Similarity, and Added-Value

When measuring the accuracy of a property in a model, the value of the property derived from the model is compared with the value obtained from its corresponding experimental structure (target). For most properties the accuracy is expressed as the percentage of the cases observed in the model that are also observed in the target. Let  $\{M\}$  be the set that consists of all predicted cases in a model and let  $\{E\}$  be the corresponding set consisting of actual cases in the experimental structure. Accuracy would then be the ratio between the number of elements of  $\{M \cap E\}$  and  $\{M\}$ :

$$\text{Accuracy} = \frac{\{M\} \cap \{E\}}{\{M\}}$$

For some properties (accessible surface area, pocket composition, electrostatic potential) this way of expressing accuracy is not convenient. The accuracy measurement for each of these properties is described in their corresponding sections. The template:target similarity is expressed in the same way as the model accuracy. But in this case  $\{M\}$  corresponds to all cases predicted using the template structure. This is done by combining the use of the template struc-

ture and the template:target alignment. For example, to predict the exposure state of a residue in the target by using its template, first the residue in the template that is equivalent to the target's residue must be identified. This equivalence is defined by the template:target alignment; if two residues are aligned, they are considered equivalent. The exposure state for the equivalent template residue is then calculated using the template structure, and the resulting value is assigned to the equivalent residue in the target. The template:target alignment was used to define the equivalences that allow the assignment of template residue measurements to the target residues in the following properties: overall accuracy, residue exposure state, residue neighborhood, pocket detection, and pocket composition. For the remaining properties (accessible surface area and electrostatic potential), the template structures alone are used to compute the property values. The added-value of the models is determined by comparing the accuracy of the models with the template:target similarity (Figure 1). The added-value always has the same units as the model accuracy. For residue-based properties, such as exposure state, the template:target similarity depends on the template:target alignment (see above). In these cases the added-value for a model is calculated using the same template:target alignment that was used to build the model (i.e., SEQ or STR alignment).

### Overall Accuracy

Overall accuracy was computed by determining the percentage of equivalent atoms between the model (or template) structure and the target structure. Equivalent atoms are defined as those atoms that are within 3.5 Å of their corresponding atom in the target after superposition of the structures. The superposition of the structures is carried out by minimizing the root-mean-square deviation of the coordinates of corresponding C $\alpha$  atoms. The correspondence of C $\alpha$  atoms for both the model and the template is defined by the alignment used to build the model. All calculations are implemented in the SUPERPOSE command of program MODELLER (Eswar et al., 2003).

### Accessible Surface Area and Residue Exposure State

Accessible surface area, ASA, of a protein was computed using the method of Lee and Richards (Lee and Richards, 1971) as implemented in the program NACCESS (Hubbard and Thornton, 1993) with a probe radius of 1.4 Å. Accessibility of a residue X to a solvent probe is the ratio of the ASA of X in the folded state of the protein to that of Gly-X-Gly tripeptide. Residues with solvent accessibility  $\geq 0.4$  (Holbrook et al., 1990) are considered exposed, and residues with solvent accessibility  $< 0.05$  are considered buried. The remaining residues are considered to have an intermediate level of exposure. For a model with  $Nm\_E$  exposed residues, the accuracy of prediction of exposed state is defined as

$$\frac{Nm\_E \cap Ne\_E}{Nm\_E}$$

where  $Nm\_E$  and  $Ne\_E$  are the set of exposed residues in the model and the experimental structure respectively.

### Residue Neighborhood

A pair of residues (with a sequence separation,  $K \geq 3$ ) is considered to be neighbors if at least one interresidue atomic distance  $D \leq D_0$ ,  $D_0 = vWr_a + vWr_b + 1$ ; where  $vWr_a$ ,  $vWr_b$  are the van der Waals radii (Chothia, 1976) of atom a and b respectively, measured in angstroms. For residue  $i$  in the model the list of its neighbors ( $Nm\_i$ ) is compared with the list of neighbors ( $Ne\_i$ ) of the corresponding residue in the experimental structure. The accuracy of neighborhood prediction of a model with  $Nres$  residues is

$$\sum_i^{Nres} \frac{Nm\_i \cap Ne\_i}{Nm\_i}$$

### Pockets

Surface pocket analysis was carried out using the PASS software (Brady and Stouten, 2000). PASS's utility as a predictive tool for binding site identification has been tested by predicting known binding sites of proteins in the PDB using both complexed macromole-

cles and their corresponding apo-protein structures. PASS reports coordinates of grid points occupying each pocket. Residues in contact with grid points (protein atoms within 4.5 Å of each grid point) were taken as boundary atoms/residues. Pockets with ten or more boundary residues (large pockets) were considered for this analysis. The initial calculation of molecular surface enclosed volumes of pockets of FABP and 183 SEQ models of our set were carried out with the CASTp web server (<http://cast.engr.uic.edu>) with a default probe radius of 1.4 Å. Though CASTp (Liang et al., 1998) is most widely used for detection and identification of pocket/cavity, its web-based nature made it impractical for use with thousands of models.

#### Electrostatic Potential

The electrostatic potential (EP) is calculated using the algorithms of Nicholls and Honig (Nicholls et al., 1991) for solving the Poisson-Boltzmann equation, as implemented in the command CalcPot of program MOLMOL (Koradi et al., 1996). EP is calculated using partial charges as provided in the MOLMOL libraries, with a dielectric constant of 80 for the solvent and 2 for the protein, and a salt concentration of 150 mM. The output of the calculation is a 3D grid containing the values of the EP at each grid point. The size of the grid is such that no protein atom is closer than 10 Å to the boundaries of the grid. The comparison of EP between two proteins was carried out by superposing the protein structures before the calculation of the EP grid, such that the resulting grids are aligned. For each point in one grid the equivalent point in the second grid is identified (i.e., the closest point in the second grid) resulting in a list of pairs of equivalent grid points. The similarity of the two grids is defined by the correlation between the EP values of the pairs. The Spearman rank correlation coefficient (Press, 1992) was used to compute the correlation. A value of 0 indicates no correlation; a value of 1 indicates complete correlation. Only pairs of equivalent grid points in which one of the potential values is larger than 0.1 or smaller than -0.1 were considered for the calculation of the correlation coefficient. Using smaller limit values (i.e., using more points) did not significantly change the results, but did increase the computation time. Values of 0.5, 1.0, and 2.0 Å for the grid spacing of the EP calculation were tested on a small set of proteins resulting in similar correlation values. A grid spacing of 1.0 Å was used for the final calculation. Because of the larger calculation time, a subset of 1000 models of medium size was used for the EP calculations. These models were divided into nine template:target sequence identity bins.

#### Curve Fitting

Curve fitting was done only for illustration, to show the trend the data follows. No particular attention was paid to the function used in fitting other than selecting one that closely follows the trend shown by the data. In cases where the trend is not very clear, no curve fitting was done. All curve fitting was done using program SigmaPlot version 8.0. For Figures 2–4 a modified sigmoidal function was used. Figure 6 uses a modified logarithmic function. Figures 7 and 8 use a hyperbolic function.

#### Acknowledgments

We thank Carlos Madrid for general assistance with hardware and software, and Bing Zhang, Dr. Marc Ceruso, and Dr. Ming-Ming Zhou for useful comments and carefully reading the manuscript. This work was supported by Mount Sinai School of Medicine start-up funds and grant 1P01GM066531-01 from the National Institutes of Health.

Received: February 26, 2004

Revised: April 29, 2004

Accepted: May 18, 2004

Published: August 10, 2004

#### References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST:

a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. (2002). The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* 58, 899–907.

Brady, G.P., Jr., and Stouten, P.F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* 14, 383–401.

Chakravarty, S., and Varadarajan, R. (2002). Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry* 41, 8152–8161.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105, 1–12.

Chung, S.Y., and Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure* 4, 1123–1127.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B., et al. (2003). Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* 31, 3375–3380.

Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* 17, 1242–1243.

Fischer, D., Rychlewski, L., Dunbrack, R.L., Jr., Ortiz, A.R., and Elofsson, A. (2003). CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins Suppl.* 6 53, 503–516.

Fiser, A., Kinh Gian Do, R., and Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci.*, 1753–1773.

Holbrook, S.R., Muskal, S.M., and Kim, S.H. (1990). Predicting surface exposure of amino acids from protein sequence. *Protein Eng.* 3, 659–665.

Hubbard, S.J., and Thornton, J. (1993). NACCESS (computer program). Department of Biochemistry and Molecular Biology, University College London.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* 14, 51–55, 29–32.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 7, 1884–1897.

Livingstone, J.R., Spolar, R.S., and Record, M.T., Jr. (1991). Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area. *Biochemistry* 30, 4237–4244.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325.

Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., and Sali, A. (2002). Reliability of assessment of protein structure prediction methods. *Structure (Camb)* 10, 435–440.

Nicholls, A., Sharp, K.A., and Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11, 281–296.

Peitsch, M.C., Schwede, T., and Guex, N. (2000). Automated protein modelling—the proteome in 3D. *Pharmacogenomics* 1, 257–266.

Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A., and Sali, A. (2002). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* 30, 255–259.

Press W.H. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition (Cambridge, UK: Cambridge University Press).

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815.

Sali, A., Matsumoto, R., McNeil, H.P., Karplus, M., and Stevens, R.L.

- (1993). Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes. *J. Biol. Chem.* **268**, 9023–9034.
- Sanchez, R., and Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl.* **7**, 50–58.
- Sanchez, R., and Sali, A. (1998). Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602.
- Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M.A., Madhusudhan, M.S., Mirkovic, N., and Sali, A. (2000). Protein structure modeling for structural genomics. *Nat. Struct. Biol. Suppl.* **7**, 986–990.
- Shindyalov, I.N., and Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
- Spolar, R.S., and Record, M.T., Jr. (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777–784.
- Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N., and Orengo, C.A. (2000). From structure to function: approaches and limitations. *Nat. Struct. Biol. Suppl.* **7**, 991–994.
- Tramontano, A., and Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins Suppl.* **6** 53, 352–368.
- Venclovas, C., Zemla, A., Fidelis, K., and Moutl, J. (2003). Assessment of progress over the CASP experiments. *Proteins Suppl.* **6** 53, 585–595.
- Wilson, C.A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
- Xu, L.Z., Sanchez, R., Sali, A., and Heintz, N. (1996). Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.* **271**, 24711–24719.