**Published online January 12, 2005**

***244–259*** *Nucleic Acids Research, 2005, Vol. 33, No. 1*
*doi:10.1093/nar/gki162*

# Accuracy of structure-derived properties in simple comparative models of protein structures

**Suvobrata Chakravarty, Lei Wang and Roberto Sanchez***

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

Received October 29, 2004; Accepted December 13, 2004

## ABSTRACT

**The accuracy of comparative models of proteins is addressed here. A set of 12 732 single-template models of sequences of known high-resolution structures was built by an automated procedure. Accuracy of several structure-derived properties, such as surface area, residue accessibility, presence of pockets, electrostatic potential and others, was determined as a function of template:target sequence identity by comparing models with their corresponding experimental structures. As expected, the average accuracy of structure-derived properties always increases with higher template:target sequence identity, but the exact shape of this relationship can differ from one property to another. A comparison of structure-derived properties measured from NMR and X-ray structures of the same protein shows that for most properties, the NMR/X-ray difference is of the same order as the error in models based on ~40% template: target sequence identity. The exact sequence identity at which properties reach that accuracy varies between 25 and 50%, depending on the property being analyzed. A general characteristic of simple comparative models is that their surface has increased area as a consequence of being more rugged than that of experimental structures. This suggests that including solvent effects during model building or refinement could significantly improve the accuracy of surface properties in comparative models.**

## INTRODUCTION

An enormous progress in our ability to discover gene sequences, both by genome sequencing projects and by more traditional methods, presents an opportunity, and a challenge, to understand the function of those genes 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 (1). Different types of information can be derived from protein structures, such as overall shape and volume; the amino acid composition of the surface and its electrostatic and hydrophobic properties; and the presence of pockets and cavities, salt bridges, disulphide bridges, etc. Knowledge of the 3D structure also allows us to look at the properties of functionally relevant subsets of the structure like the binding/active sites (2–4) as well as the exploration of molecular interactions through docking calculations for protein–ligand and protein–protein interactions. Structure-derived information describes the similarities and differences among proteins, and therefore is valuable in understanding how they function (5). This type of analysis is critical when we try to understand differences between homologs in different tissues or organisms, or between polymorphisms of a particular gene (6). Ideally, comparative studies of structure-derived information should use a complete set of proteins for the particular question being asked. For example, if one wants to understand differences in substrate specificity through the study of structural differences in a family of enzymes, ideally one would need 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 (7). 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, the use of predicted protein structures is necessary to obtain the kind of structure-derived information described above.

*To whom correspondence should be addressed. Tel: +1 212 659 8648; Fax: +1 212 849 2456; Email: roberto@sanchezlab.org
Present address:
Lei Wang, Structural Biology Program, J. G. Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA

Comparative modeling uses experimentally determined protein structures (templates) to predict the 3D conformation of another protein with a similar amino acid sequence (target). Currently, it is the most accurate approach to protein structure prediction (8–10). 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 (11). Comparative modeling is particularly well suited for the kind of studies described above, where comparison of similar proteins is the focal point. Since comparative modeling 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. Even when a single template is used in comparative modeling, the resulting models contain more information about the target than the template used to build them; this added-value justifies the use of comparative models (12).

Although comparative models can be very accurate, in general they contain errors and do not reach the level of accuracy of high-resolution experimental structures. The overall accuracy of comparative models depends mainly on the structural similarity between the target and the template, and the accuracy of the alignment of the two sequences; both are a function of template:target sequence similarity (1). Although it is useful to know the overall accuracy of comparative models [commonly measured by root mean square deviation (RMSD) between model and the experimental structure], this measure does not always correlate with, or is not always a good indicator of the accuracy of a particular type of structure-derived property (SDP). For example, the amino acid composition of the surface of a protein can be reliably calculated from models, but as the overall accuracy drops so does the reliability of estimation of surface composition. The exact relationship between overall accuracy (or a parameter of the model like its sequence similarity with the template structure) and the accuracy of the surface composition or any SDP, calculated from it, is not known. This is particularly problematic if the model is to be used to compare its SDPs with those from a related, but different, protein because it would not be possible to know if the observed differences are real or just a consequence of errors in the model. Hence, the need arises to address the accuracy of SDPs separately. This is particularly important because the relevant accuracy of a model may differ depending on the application. If comparative models are to be used for structure-based drug design or protein–ligand docking, the accuracy of the pockets (binding site) and properties such as electrostatic potential of the binding site are far more relevant than the accuracy of any other region of the protein. If comparative models are used to aid in the structure-based prediction of sub-cellular localization (13), the accuracy of surface composition is the most relevant feature. There is a need to look at the accuracy of models from the perspective of the application. Though seemingly trivial, systematic large-scale analysis of comparative modeling to estimate the average accuracy of various structural features has not been carried out.

This work analyzes the accuracy of structure-derived properties of comparative models. Specifically we ask four questions. (i) What is the relationship between the template:target sequence identity and the accuracy of SDPs in the model? (ii) At what level of template:target sequence identity is the accuracy of SDPs in models equal to the difference between NMR and X-ray structures? (iii) Is there a difference in the relationship between template:target sequence identity and accuracy for different SDPs? (iv) What is the influence of template:target alignment errors on the accuracy of SDPs in models?

## METHODS

### Dataset

Chains of X-ray structures with resolution better than 2.5 Å were selected from the Protein Data Bank (PDB) (14). A representative set of these chains was selected by an all-against-all comparison of their sequences using BLAST (15) 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 program CE (16). 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 protein size into three non-overlapping groups: small (50–100 residues), medium (150–200 residues) and large ($\geqslant$250 residues). The alignments were classified into 18 groups based on the sequence identity ranging from 10 to 100% with a bin size of 5%. Sequence identity was defined as the ratio between the number of identical aligned residues and the number of target residues in the template:target alignment. The number of alignments for groups with lower sequence identity outnumbered those of groups with higher sequence identity. Approximately, 200 alignments were selected at random from the groups with lower sequence identity so that the contribution of each group or bin to the total set of alignments is approximately equal. There are 1564, 911 and 856 unique chains of small, medium and large proteins, respectively. The number of alignments for small, medium and large proteins is respectively 4912, 4104 and 3716 in our dataset.

### Model building

The structure-based alignments produced by CE were used as inputs to program MODELLER version 6v2 (17,18) 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 was based on 'simple' alignments generated by realigning the sequences of the CE alignments using the ALIGN command of MODELLER; these are called SEQ. SEQ alignments are based exclusively on sequence information, as opposed to the structure-based STR alignments. Models were constructed from the respective alignments using the default 'model' routine in MODELLER (17,18). All alignments contained a single template, and no loop modeling was performed. A total of 25 464 models were calculated for small, medium and large proteins; half of them based on SEQ alignments and the other half on STR alignments.

### NMR/X-ray and X-ray/X-ray pairs

NMR structures were obtained from the PDB (14). NMR structures showing 100% sequence identity [using BLAST (15)] with high-resolution (≤2.5 Å) X-ray structures, as well as a CE *Z*-score ≥ 4.5 for every model of the NMR ensemble were selected. For an NMR structure with more than one corresponding X-ray structure, the one with highest resolution and most similar ligand or hetero-atom composition was chosen. This resulted in 48 pairs of structures with a size range of 100–200 residues (close to the medium size), out of these, 34 NMR structures are represented by 10 or more structural models. All the structural models of an NMR ensemble are considered equivalent having the same probability of representing the structure of the protein. To avoid differences arising due to disorders in the N- and C-termini of the NMR structure, the termini of each NMR/X-ray pair of structures were removed in the following steps: (i) for an NMR structure with $N$ number of models, the model whose CE alignment ($Z$-score ≥ 4.5) with the corresponding X-ray structure had the highest number of equivalent residues was chosen; (ii) only the continuous stretch of residues in the CE alignment of the selected model was retained; and (iii) all the remaining $N - 1$ models as well as the corresponding X-ray structure were truncated to look like the selected model in sequence. The above-mentioned chain lengths refer to truncated polypeptides. The quality of NMR structures can be described by the completeness of nuclear overhauser effects (NOEs) (ratio between observed and expected NOEs) (19). This value can be estimated from the fraction of residues in the Ramachandran Core regions (19,20). On average, 69 ± 12% of residues in the NMR structures of this set fall into the core regions, this would correspond to structures with ∼50% completeness of NOEs at 4 Å cut-off, which is close to the average observed for NMR structures solved after 1996 in the PDB (19). The size of proteins in this set is slightly smaller than the models in our Medium set (above); hence, for comparison of properties that are size dependent, such as number of pockets and salt-bridges per protein, the 'NMR' values are appropriately scaled in Figures 7A and 8A.

Pairs of X-ray structures of a particular protein determined either in different space groups or under different conditions such as pH, or having different resolution (0–2.5 Å) were used for comparison to put an upper limit to the accuracy. These X-ray/X-ray pairs were selected from alignments with 100% sequence identity (covering at least 95% of the chain length) in an all-against-all BLAST (15) comparison of sequences of high-resolution X-ray structures. There are 1341 such pairs of structures. No filtering was done to prevent comparison of structures bound to different ligands or structures with and without ligand. Although ligand-induced conformational changes may contribute to observed differences between X-ray structures or NMR/X-ray pairs, a comparison of the deviations between pairs of proteins with and without ligand showed that there is no significant difference between the two groups. Therefore, all examples were retained for this analysis.

### Model accuracy

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

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

For some properties [accessible surface area (ASA) and electrostatic potential], this way of expressing accuracy is not convenient. The accuracy definition for each of these properties is described in their corresponding sections.

### Solvent accessible surface, solvent excluded surface, exposure state and fractal dimension

Accessible surface area of a protein was computed using the method of Lee and Richards (21) as implemented in the program NACCESS (22) 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 in a Gly–X–Gly tripeptide. Residues with solvent accessibility ≥40% (23) are considered to be exposed, and residues with solvent accessibility <5% 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 exposed state assignment is defined as

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

where $\{Nm\_E\}$ and $\{Ne\_E\}$ are the sets of exposed residues in the model and the experimental structure, respectively.

Solvent excluded surface area (SES) is computed using program MSMS (24) and was used for computing fractal dimension (FD). FD measures the rate of change of SES as the probe size increases (25). By definition,

$$\text{FD} = \frac{1}{n}\sum_{i}^{n} D_i \quad \text{and} \quad D_i = 2 - \left[\frac{d\ log(\text{SES})}{d\ log(\text{rad})}\right]_i$$

SES and rad are the *i*-th solvent excluded surface area and probe radius, respectively (25) and $n$ is 25 (see below). For convenience $D_i$ was computed as

$$D_i = 2 - \left[\frac{log\left(\frac{\text{SES}_i}{\text{SES}_{i-1}}\right)}{log\left(\frac{\text{rad}_i}{\text{rad}_{i-1}}\right)}\right]$$

$\text{SES}_i$ and $\text{SES}_{i-1}$ are solvent excluded surface area for *i*-th and $(i - 1)$th probe radius, respectively. Probe radius ranged between 1.0 and 3.5 Å with an interval of 0.1 Å, hence, $n$ is 25. Smoothed atomic FD (SAFD) describes the roughness around each atom by smoothing over its neighborhood (26). By definition,

$$f_i = 2 - \left[\frac{d\ log\sum_j (A_j)}{d\ log(\text{rad})}\right] \quad \text{and} \quad S_i = \left[\frac{r_i}{r_i + r_p}\right]^2$$

where $f_i$ is the SAFD value for atom $i$, $A_j$ is the contact area (27) of atom $j$, and the sum is over all neighbor atoms $j$ within

5 Å of atom *i*. The contact area of atom *i* is obtained by multiplying the scaling factor $S_i$ to the ASA of the atom, where $r_i$ and $r_p$ are the radii of atom *i* and probe, respectively. SAFD was computed in the same way as FD (see above), i.e. in place of the SES of the whole protein, the sum of the contact area of a few select atoms was used. SAFD was used to measure the roughness of pockets.

### Residue neighborhood and inter-residue distance

Two residues (with a sequence separation, $K \geqslant 3$) are considered to be interacting or neighbors if at least one inter-residue atomic distance ($D$) is smaller than $D_o$, where $D_o = \mathrm{vWr}_a + \mathrm{vWr}_b + 1$ and $\mathrm{vWr}_a$, $\mathrm{vWr}_b$ are the van der Waals radii (28) of atoms '*a*' and '*b*', respectively. 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 the residue neighborhood in a model with $N_{res}$ residues is

$$\sum_{i}^{N_{res}} \frac{\{Nm\_i\} \cap \{Ne\_i\}}{\{Nm\_i\}}.$$

An interacting pair of residues is considered either buried or exposed only when both the partners are buried or exposed, respectively. In all other cases, the interacting pair is considered to be of intermediate exposure. Neighborhood is a qualitative measure of distance between a pair of residues. Hence, for inter-residue distance we resort to a more quantitative measure. The inter-residue distance is computed from the geometric centers of side-chain atoms. For residue *i* in the model, the distances of the center of mass of *n_i* residues ($\mathrm{A}i_1, \mathrm{A}i_2, \ldots, \mathrm{A}i_{n\_i}$) within a sphere of 6 Å from its geometric center is calculated. Let $\mathrm{A}i_1', \mathrm{A}i_2', \ldots, \mathrm{A}i_{n\_i}'$ be the corresponding *n_i* residues in the experimental structure. Let $D\_Ai_1$ be the distance between residues *i* and $\mathrm{A}i_1$. Then $\Delta D_i$, the set of all the differences $|D\_Ai_1 - D\_Ai_1'|$, $|D\_Ai_2 - D\_Ai_2'|$, $\ldots |D\_Ai_{n\_i} - D\_Ai_{n\_i}'|$ for residue *i*, is used to obtain the combined distribution of $\Delta D$ (combined from $N_{res}$ residues of a model). The fraction of $\Delta D \leqslant 2$ Å is used as a measure of the accuracy of contact distances.

### Pockets

Surface pocket analysis was carried out with program PASS (29). 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 10 or more boundary residues (large pockets) were considered for this analysis. A pocket in the model was considered identical to one in the experimental structure if at least 60% of its boundary residues were identical.

### Packing density

Packing density of a residue is defined as

$$\text{Packing density} = \frac{\text{volume}_{\text{vdW}}}{\text{volume}_{\text{Voronoi}}}$$

where $\text{volume}_{\text{vdW}}$ and $\text{volume}_{\text{Voronoi}}$ are the van der Waals and Voronoi volume of residues, respectively (30). The van der Waals volumes were taken from Creighton (31). For an infinite set of arbitrary points/atoms in space, the Voronoi procedure (32) divides up space with a unique volume assigned to each point or atom. Residue Voronoi volumes were obtained from the sum of the constituent atom-volumes. Only residues with solvent accessibility = 0.0 were considered for packing calculations (33) as only volumes of interior atoms are possible to calculate with the Voronoi construction. The program for computing the Voronoi volume of interior atoms was obtained from http://www.molmovdb.org/geometry (33).

### Volume

Volume enclosed by the solvent accessible surface of a protein was computed using the ProShape suite of programs (34) with a probe radius of 1.4 Å.

### Salt-bridge

We considered an ionic interaction to be a salt-bridge when the distance between the center of positive charge (Arg, Lys, His) and the center of negative charge (Glu, Asp) as well as at least a pair of oppositely charged atoms are within 4 Å (35). The following atoms were considered for the salt-bridge calculations NE, CZ, NH1, NH2 of Arg; NZ of Lys; HD1, NE2 of His; OD1, OD2 of Asp; and OE1, OE2 of Glu. The distance between the centers of positive and negative charge was varied from 4 to 8 Å at an interval of 0.5 Å for models to check the efficiency of salt-bridge detection in models. The accuracy was highest at 4 Å.

### Electrostatic potential

The electrostatic potential is calculated using the algorithms of Nicholls *et al*. (36) for solving the Poisson–Boltzmann equation, as implemented in the command CalcPot of program MOLMOL (37). Partial charges provided in the MOLMOL libraries are used, dielectric constants 80 and 2 for solvent and protein, respectively, and salt concentration of 150 mM. The output of the calculation is a 3D grid containing the values of the electrostatic potential 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 electrostatic potential between two proteins was carried out as described previously (12). Because of the larger calculation time, a subset of 1000 models of medium size was used for these calculations. These models were divided into nine template:target sequence identity bins. Also, the smaller set of 30 X-ray/X-ray pairs of Jacobson *et al*. (38) was used.

## RESULTS

The following structure-derived properties (SDPs) were analyzed for 12 732 single-template models: (i) overall accuracy, (ii) inter-residue distance, (iii) exposure state of residues, (iv) neighborhood of residues, (v) ASA, (vi) identification of surface pockets, (vii) salt-bridges and (viii) electrostatic potential. Models were built on single templates using two alternative alignments: template:target pairwise sequence alignment (SEQ models) and structure-based alignment (STR models) of the experimental structure of target and template (see Methods). An STR model illustrates the accuracy of a given property in the absence of alignment errors. In other words, comparison of accuracy of SDPs between SEQ models and STR models provides an indication of the effect of alignment errors. The accuracy of SDPs is discussed here as a function of

sequence identity of the template:target alignment as it is the most commonly referred variable in comparative modeling (39). Because the significance of sequence identity as a measure of similarity depends on the alignment length (40) models were further divided into three classes based on sequence length (see Methods). When measuring accuracy, the model is compared with its corresponding experimental structure. High-resolution X-ray structures were selected as the accuracy reference for the following reasons: X-ray structures are generally accepted to have higher information content than NMR-structures; a much larger number of high-resolution X-ray structures than NMR structures are available, thus allowing for a larger-scale comparison; and finally, most of the work done on assessment of comparative models has been done using X-ray structures as templates and targets, thus facilitating the comparison of our results with those obtained in other studies. Comparisons of SDP accuracy are carried out in three distinct modes: (i) comparison of SEQ models among Small, Medium and Large sets (Protein Size Effect). (ii) Comparison of SEQ versus STR model (Alignment Error Effect). (iii) Comparison of SEQ model versus NMR/X-ray pairs and X-ray/X-ray pairs (Experimental Variation) (see Methods). Except for (i), in the other two modes of comparison only medium-sized models were used.

### Overall accuracy of models

Overall accuracy was measured by the coordinate RMSD of all $C_\alpha$ atoms, and by the percentage of equivalent $C_\alpha$ atoms within 3.5 Å of each other in the optimal superposition of the model and the target experimental structure. These are common measurements that have been performed systematically for comparative models (41,42). Figure 1A and B (RMSD and percentage of equivalent $C_\alpha$ atoms, respectively) shows the change in overall accuracy as a function of template:target sequence identity for Small, Medium and Large models. As expected, the measures show a trend of increasing accuracy (decreasing error) with higher template:target sequence identity. The apparent higher accuracy for the Small set is due to the size dependence of measures such as RMSD; it is easier to achieve lower RMSD with a smaller number of atoms. At 35% identity, the average percentage of equivalent $C_\alpha$ atoms and the RMSD in the models is similar to the average difference observed for the NMR/X-ray comparison (dotted line). There is a sharp decrease in the accuracy below 35% sequence identity, as reported previously (1,41). Two factors contribute to this; the divergence in the structural similarity between homologous proteins (43,44) and the increase in the template:target alignment errors. The comparison of SEQ and STR models (Figure 1C) deconvolutes these two contributions. At 40% sequence identity, the difference in accuracy between SEQ and STR models starts to become visible and it sharply increases below 35% sequence identity, showing the effect of alignment error on the overall model accuracy. But even in the absence of alignment errors (STR), the accuracy drops sharply below 35% indicating that the target and template structures start to diverge.

### Inter-residue distance

One of the simplest measures that can be derived from a protein structure is the distance between pairs of residues or atoms. This measurement is common in determining the presence of a potential disulfide bridge or salt-bridge. It is also the basis for the development of statistical potentials used in many aspects of computational structural biology (45). The fraction of distances with $\Delta D \leqslant 2$ Å, where $\Delta D$ is the deviation in the inter-residue distance, is used as a measure of the accuracy (see Methods). Although this measure can be used to estimate overall accuracy of a model, it is used here to illustrate the difference in accuracy between exposed and buried residues. The accuracy of the inter-residue distances decreases as the

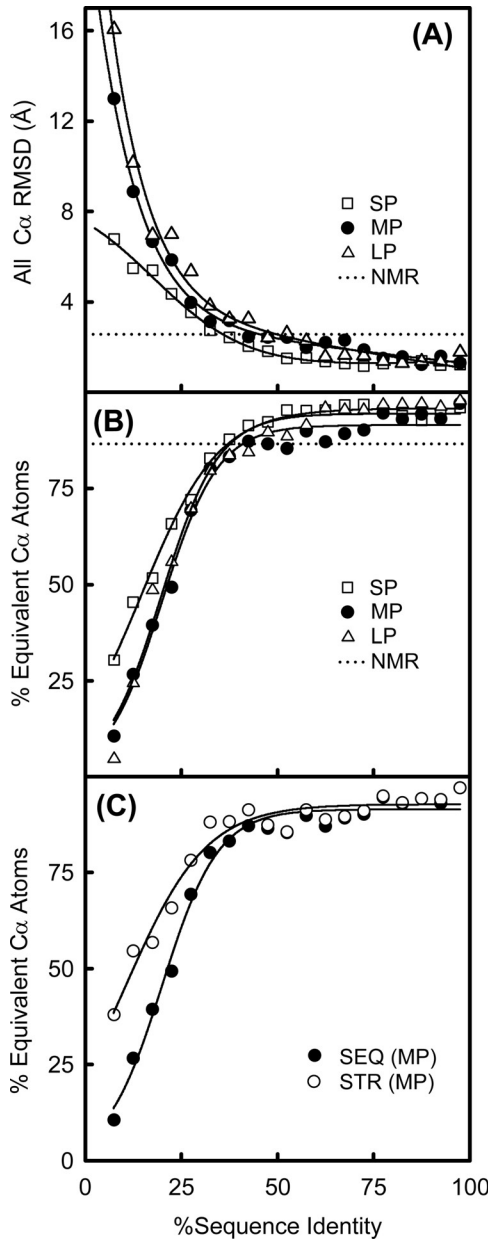


**Figure 1.** Overall accuracy is shown as a function of template:target sequence identity. (**A**) RMSD (all $C_\alpha$) between SEQ models and their corresponding experimental structures is compared among Small (SP), Medium (MP) and Large (LP) proteins. The horizontal dotted line represents the RMSD (all $C_\alpha$) between NMR structures and their corresponding X-ray structures for medium-sized proteins (MP). (**B**) Percentage of equivalent $C_\alpha$ atoms of SEQ models is compared among SP, MP and LP. The dotted line corresponds to the NMR/X-ray difference. (**C**) Comparison of percentage of equivalent $C_\alpha$ atoms between models built on sequence-based (SEQ, closed circles) and structure-based (STR, open circles) alignments for MP.

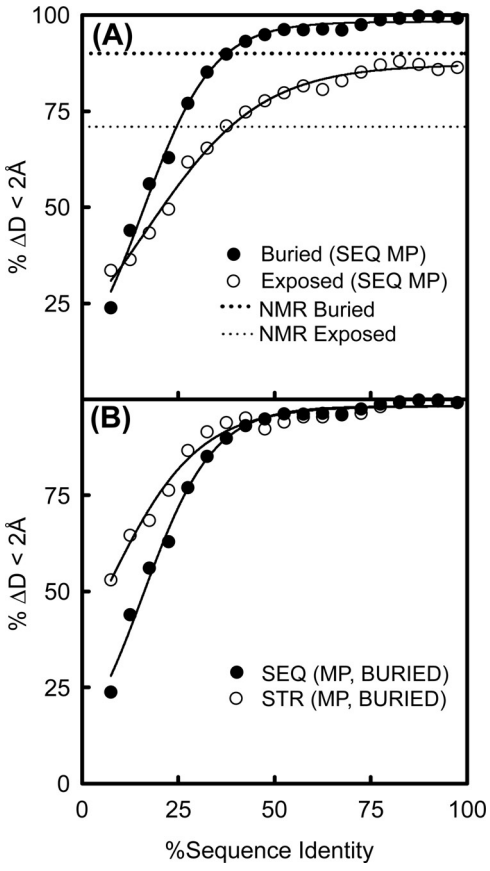


**Figure 2.** Inter-residue distance. Accuracy is defined as the fraction of residue pairs with $\Delta D \leq 2$ Å (see text) and is shown as a function of template:target sequence identity. (**A**) Accuracy of exposed (open circles) and buried (closed circles) residues in SEQ models of medium-sized proteins (MP). The dotted lines represent NMR/X-ray differences. (**B**) Comparison between models built on sequence-based (SEQ, closed circles) and structure-based (STR, open circles) alignments for MP.

residue becomes more exposed (Figure 2A). The accuracy of the solvent exposed residues reaches a plateau at 80% while that of buried residues at 100%, this is probably owing to the less constrained environment of surface residues compared with that of interior residues. This is further confirmed by the observation that the deviation in the inter-residue distances between NMR and X-ray structures is dependent on the exposure state of the residues (Figure 2A, dotted lines). At 40% sequence identity, inter-residue distances for exposed and buried residues reach accuracies equivalent to the difference between NMR and X-ray structures. Below 40% identity, the errors due to misalignments and the template:target structural divergence start to appear (Figure 2B). The marked difference in the accuracy of exposed and buried residues indicates that SDPs that depend on surface residues may have markedly different accuracy from SDPs that depend on buried residues at the same level of template:target sequence identity.

### Residue exposure state

Exposure state of a residue, i.e. if a residue is exposed, intermediate or buried, is decided based on its solvent accessibility (see Methods). Residues accessible to the solvent are generally responsible for the interaction 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 the information is used in applications such as site-directed mutagenesis, sub-cellular localization prediction and protein design. The accuracy of exposure state for exposed residues, which represents the probability that a residue that is exposed in the model is also exposed on the experimental target structure, decreases with protein size (Small > Medium > Large) and increases with template:target sequence identity (Figure 3A and D). The NMR/X-ray difference is comparable to the error in SEQ models at 30% sequence identity (Figure 3A, dotted line). The higher accuracy with decreasing protein size is probably owing to the increase in the surface/volume ratio (i.e. a larger proportion of exposed residues) as protein size decreases (Figure 3A, inset). Thus, in a smaller protein, the probability of randomly assigning an exposed state is higher. Another indication of this effect is that the comparison of accuracy for exposed and buried residues in Medium proteins (Figure 3C), which shows that below 30% identity exposed residues are assigned with higher accuracy than buried residues in spite of the higher conservation of buried residues. As the quality of models decreases, the assignment of exposure state becomes more random and accuracy approaches a value similar to the fraction of residues in the corresponding exposure state (e.g. in medium-sized models ~40% exposed and ~18% buried; see Figure 3A, inset). Exposure state accuracy seems to be slightly more robust than overall accuracy with respect to alignment error (Figure 3B). Only below 30% sequence identity do the errors in the alignment have an impact on the accuracy of exposure state.

### Residue neighborhood

Information about neighborhood of a particular residue is obtained from the contacts it makes with its neighbors. Information about neighborhood of residues is routinely used in 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 those in the experimental structure (see Methods). Neighborhood accuracy shows no clear dependence on protein size (Figure 4A). At 30% sequence identity, the neighborhood accuracy of SEQ models is comparable to the difference observed between NMR and X-ray structures (Figure 4A, dotted line) and drops rapidly at lower values. Above 30% sequence identity, the alignment error is small enough not to have effect on the accuracy (Figure 4B). Below 30% identity, the effect of the alignment error becomes apparent and the magnitude of the difference observed between SEQ and STR models in this case (Figure 4B) is larger than that observed for exposure state (Figure 3B). The accuracy of neighborhood for exposed residues is clearly lower than that for buried residues over the whole range of sequence identities (Figure 3C), but the difference is not as large as that observed for inter-residue distance (Figure 2A).

### Accessible surface area

The value of the total ASA of the protein is frequently used in the calculation of protein stability and binding (46) or

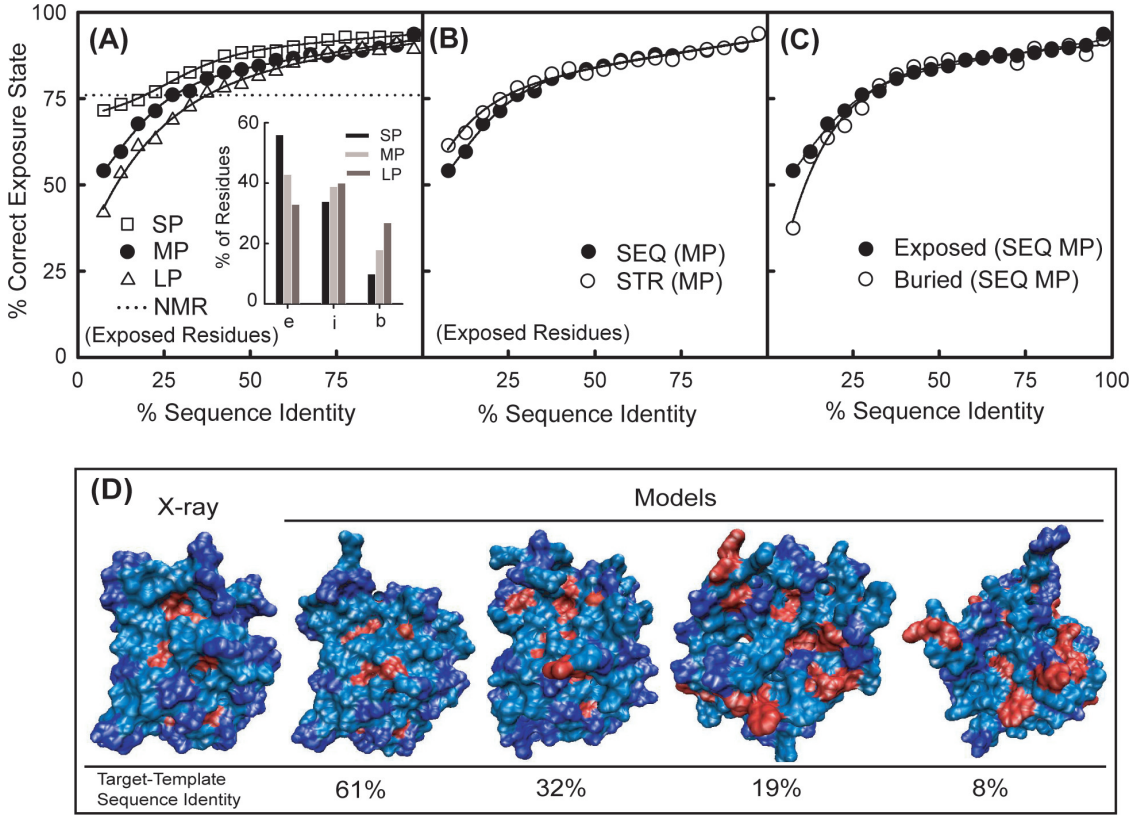


**Figure 3.** Residue exposure state. Exposure state indicates if a residue is exposed to the solvent or buried in the interior of the protein. The fraction of correctly predicted exposed (A–C) and buried (C) residues of models is shown as a function of template:target sequence identity: (**A**) Comparison among SEQ models of Small (SP), Medium (MP) and Large (LP) proteins. The dotted line indicates the NMR/X-ray difference. The inset shows percentage of exposed (e), intermediate (i) and buried (b) residues in SP, MP and LP. (**B**) Comparison between models built on sequence-based (SEQ, closed circles) and structure-based (STR, open circles) alignments in medium-sized proteins (MPs). (**C**) Comparison between exposed and buried residues of medium-sized SEQ models. (**D**) Surface of models *Streptomyces griseus* protease B colored by exposure state of residues in the corresponding experimental structure (1sgp): buried (red), exposed (magenta) and intermediate (blue). The sequence identities of the models are indicated below. The accuracy of the models is such that they would fall on the MP curve in (A).

oligomerization state (47,48). We use the average per-residue ASA difference $|\Delta ASA|/N_{res}$ (Å$^2$) = $|ASA_E - ASA_M|/N_{res}$ as a measure of the error (Figure 5A and B); $ASA_M$ and $ASA_E$ are the total surface area of a model and its corresponding experimental structure, respectively; and $N_{res}$ is the number of residues in the protein. Below 50% sequence identity, there is little difference in per residue based area estimation across protein size (Figure 5A). This indicates that the error in total ASA estimates would linearly increase with protein size. Above 50% identity, the Small models show a larger $|\Delta ASA|/N_{res}$ than Medium and Large models. $|\Delta ASA|/N_{res}$ is comparable to the difference between NMR and X-ray structures when sequence identity reaches 40% (Figure 5A, dotted line). The effect of misalignments on ASA accuracy is very small compared to their effect on the accuracy of other properties discussed so far (Figure 5B). At very low-sequence identity, STR models have slightly larger error than SEQ models (Figure 5B). This is contrary to what we have seen so far where STR models are always more accurate than SEQ models (Figures 1–4). This is probably a consequence of the number and size of insertions and deletions present in the models at this level of sequence similarity. Since no loop modeling or any other type of refinement is being used, each insertion in the model adopts a relatively random conformation which tends to be extended, thus exaggerating the total ASA value. At very low-sequence identity, the SEQ alignment in general presents a smaller number of gaps than the STR alignment because of the effect of the gap penalty function (data not shown). Although the SEQ alignment is less accurate, the fact that it contains fewer gaps at very low-sequence identity is an advantage when estimating the total ASA for a protein.

To further characterize the error in ASA, the distribution of the relative change in ASA, $\Delta ASA/ASA = (ASA_E - ASA_M)/ASA_E$, was plotted for models, templates and NMR structures (Figure 5C). $\Delta ASA/ASA$ for models is mostly in the negative region (ranging between −20 and +5%) indicating that the majority of models have larger surface area than their corresponding experimental structure. The template, on the other hand, shows symmetrical distribution around 0 ranging roughly between −15 and +15%. This indicates that corrected ASA can be computed from models as their $\Delta ASA/ASA$ distribution is skewed (Figure 5C). In most cases, the ASA estimate in the models should be decreased by the value indicated in Figure 5A. This is relevant, for example, in applications where ASA of models is used to calculate change in heat capacity ($\Delta C_p$) of binding (49) and also in computation of

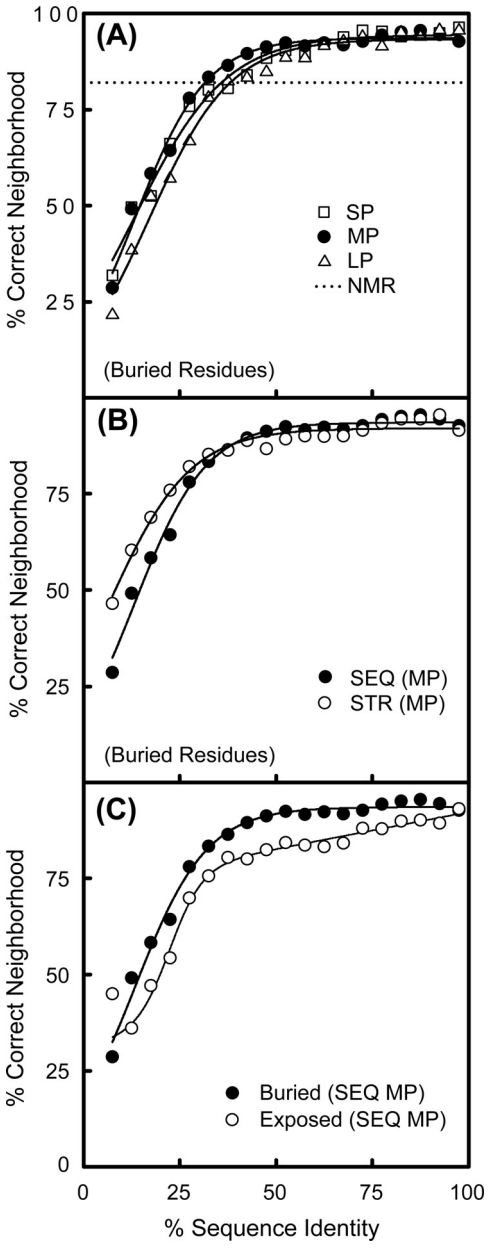


**Figure 4.** Residue neighborhood. Neighborhood of a residue is defined as the list of residues in van der Waals contact with it. The fraction of correctly predicted neighbors of buried (A–C) and exposed (C) residues of models is shown as a function of template:target sequence identity. (**A**) Comparison among SEQ models of SP, MP and LP. The dotted line indicates the NMR/X-ray difference. (**B**) Comparison between models built on sequence-based (SEQ, closed circles) and structure-based (STR, open circles) alignments for medium-sized proteins (MP). (**C**) Comparison between exposed and buried residues of SEQ models of MP.

solvent–protein interaction energy from ASA (50). NMR structures show a broad symmetric ΔASA/ASA distribution centered around −5%, indicating a tendency to have increased ASA with respect to X-ray structures. To investigate whether the ASA increase in models and NMR structures is due to an increase in the molecular volume, ASA enclosed volume ($V$) was computed using ProShape (34) and the distribution of the relative change in volume, $\Delta V/V = (\text{Volume}_E - \text{Volume}_M)/\text{Volume}_E$, was examined (Figure 5D). For models, the distribution of $\Delta V/V$ shows a narrow near-symmetric distribution around 0 indicating that an increase in volume is not the main cause of the increase in surface area. In contrast, NMR structures do show an increase in volume with respect to the corresponding X-ray structure. A possible explanation for ASA increase in the models is that the surface of a model is more rugged than that of a natural protein. We examined surface roughness by computing FD (see Methods). FD measures the rate of change of SES as probe size increases (25). For a rough surface, the rate of change would be faster than that of a smooth surface, i.e. its FD is higher. FD of a modeled protein is not only always larger than that of a natural protein, but also shows a strong dependence on sequence identity (Figure 6A). The value of FD obtained in our study is slightly lower than that of Lewis and Rees (25), but close to that calculated by Timchenko *et al*. (51) from a power law relationship between number of probe bodies covering the surface and the probe radius. Because FD ranges between 2 and 3, FD-2 (the lower limit) is used here as reference to measure any relative change. The relative change in FD, ΔFD/(FD-2), of models with respect to their corresponding experimental structures ranges between −80 and +40% (Figure 6B). The increase in the FD of a model may be due to a larger number of surface pockets or internal cavities with respect to the experimental structures (Figure 6C). As the probe size increases surface pockets and internal cavities become inaccessible to the probe resulting in a large change of surface area accounting for higher FD of the modeled protein (Figure 6C). These results suggest that the quality of comparative models could be improved by explicit refinement of the surface. Because ∼80% of the total area of proteins comes from side chains, the observed increase in ASA and FD could be attributed to inaccurate side-chain modeling. Refinement of surface residues by side-chain modeling with SCWRL3 (52) did not result in any improvement for surface features (data not shown). Even though methods such as SCWRL3, which are based on backbone-dependent rotamer libraries, are more accurate than other methods for predicting $\chi_1$ and $\chi_2$ side-chain dihedral angles (53), the observed surface artifacts may still persist because of long side chains where incorrect assignment of dihedral angles beyond $\chi_1$ and $\chi_2$ still results in incorrect positioning of the bulk of the side-chain atoms. This suggests that additional steps such as including solvent effects and electrostatic interactions would probably be needed to improve the accuracy of surface features. NMR structures also show an increased FD with respect to the X-ray structure with ΔFD/(FD-2) ranging between −60 and +50% (Figure 6A and B). In NMR structure determination, there are fewer interproton distance constraints for side chain atoms than for main chain atoms, hence, side-chain conformations are less defined than the backbone (44). The observed increase in FD is a manifestation of this.

## Surface pockets

Protein function, such as binding a ligand, is frequently mediated by surface pockets. The accuracy of detection and volume of surface pockets in models was measured. 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 Fatty Acid

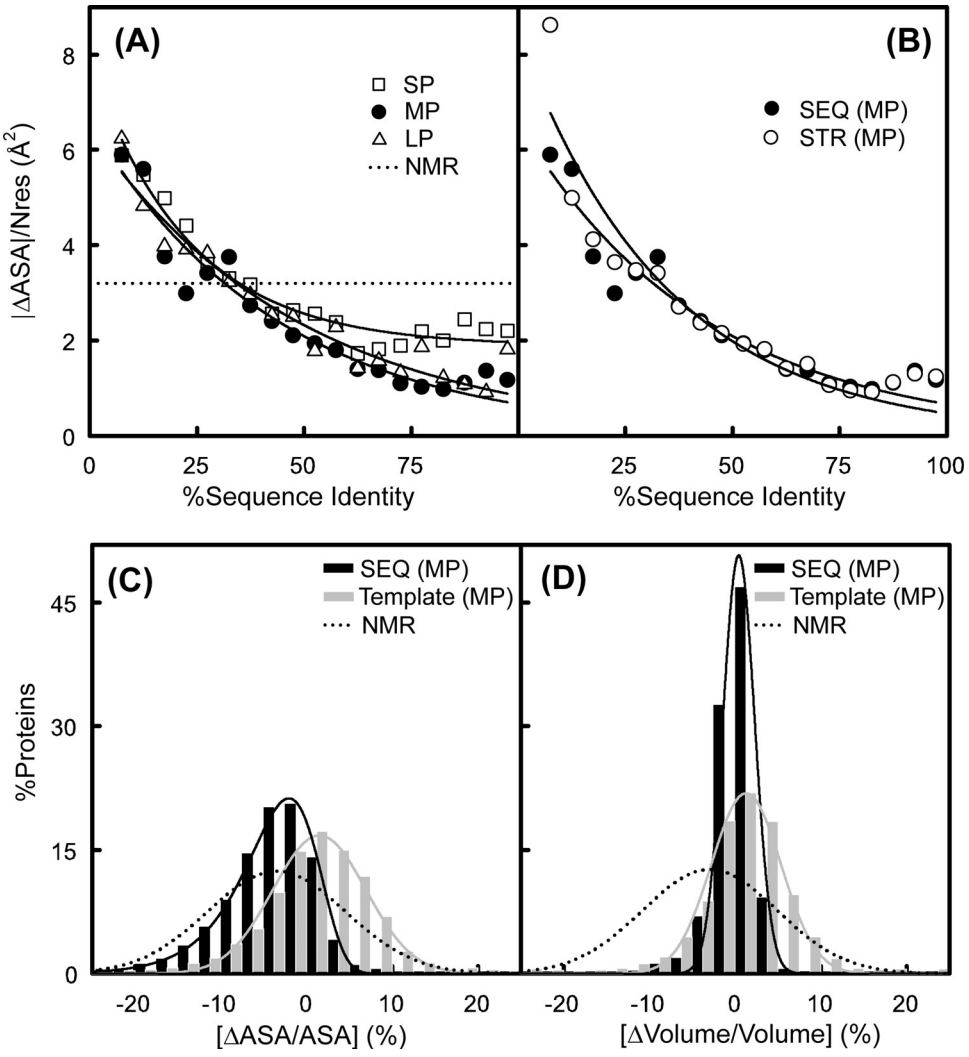


**Figure 5.** Surface and volume. ASA and volume of models are compared with the corresponding target experimental structures. Change in ASA per residue, $|ASA_E - ASA_M| / N_{res}$, as a function of template:target sequence identity: (**A**) Comparison among SEQ models of SP, MP and LP, the dotted line corresponds to the NMR/X-ray difference; (**B**) Comparison between SEQ and STR models of MP. (**C** and **D**) Distribution of relative change in ASA (C) and volume (D) for model–target comparison (black bars), template–target comparison (gray bars) and NMR/X-ray comparison (dotted line). Relative ASA change is $(ASA_E - ASA_M)/ASA_E$. Relative volume change is $(Volume_E - Volume_M)/Volume_E$.

Binding Protein family showed large variation even between very close homologs (data not shown). However, this difference is not due to changes in substrate specificity or due to presence or absence of substrate, 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. Even the side chains of the same protein can show large variation in conformation when crystallized in different space groups due to differences in crystal packing (38). Hence, for this study only the identification and location of pockets were dealt with and comparison of pocket size was avoided. Program PASS (29) was used for detection and identification of pockets (see Methods). 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 Methods). The pocket analysis was carried out only for the set of medium-sized models. One-third of the pockets of medium-sized models had 10 or more boundary residues and the remaining two-thirds had fewer than 10 boundary residues. We looked at large pockets (with 10 or more boundary residues) because the largest pocket of a protein is most often the biological active site (54). The volumes of these pockets range between 100 and 800 $Å^3$ (Figure 7A, inset). The number of large pockets per protein is not only much higher in models but also dependent on template:target sequence identity (Figure 7A). Alignment errors are not the cause of the increased number of pockets since STR models are as accurate as SEQ models in this case (Figure 7A). Hence, this effect is probably due to the lack of treatment of insertions and incorrect modeling of non-conserved side chains. Side chains adopt the conformation of the corresponding template residue in case

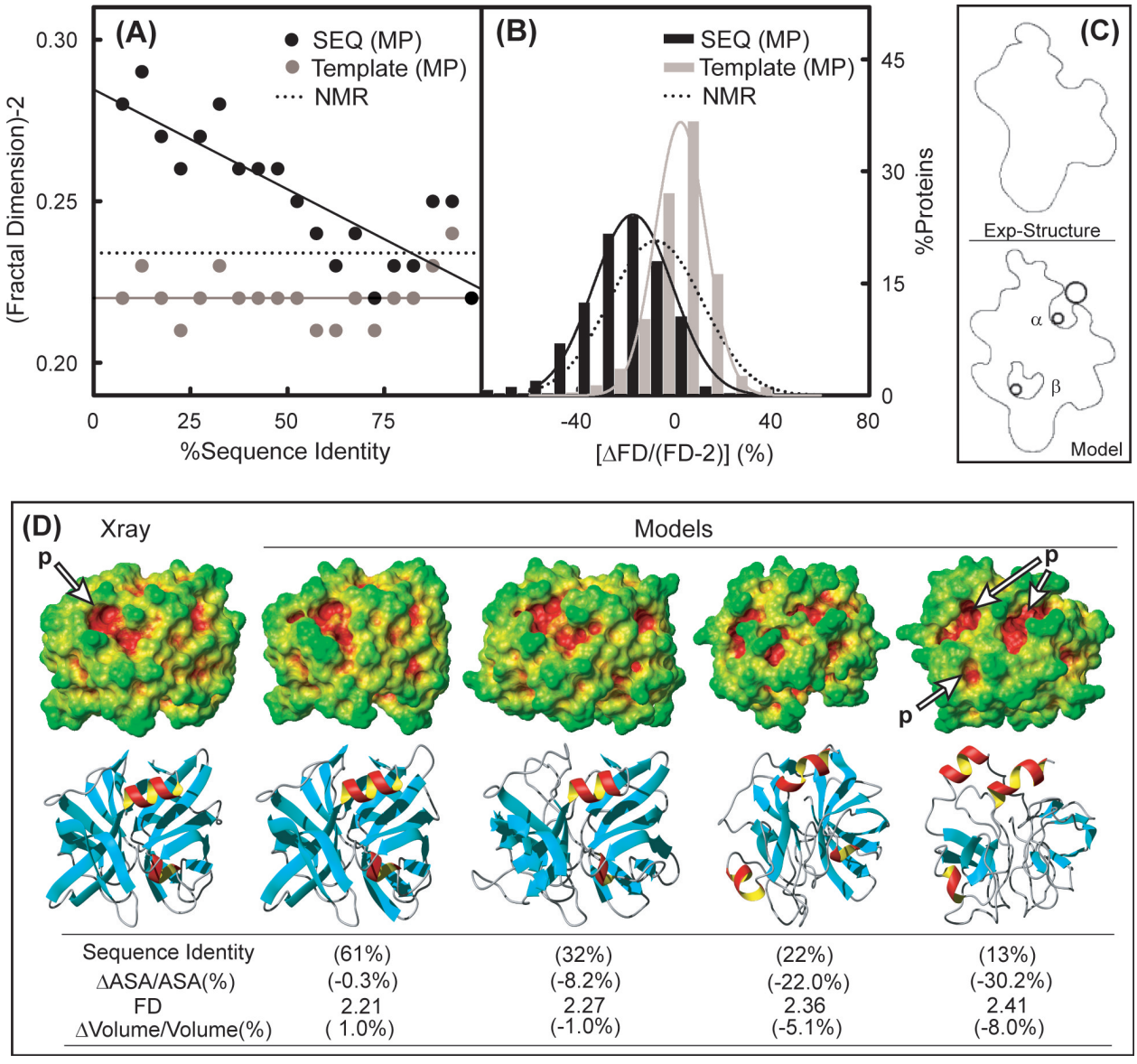


**Figure 6.** FD of models is compared with the corresponding target experimental structures: (**A**) FD of model surface as a function of template:target sequence identity (closed circles), FD of template surface (gray circles) and FD of NMR structures (dotted line). (**B**) Distribution of the relative change in FD, $(FD_E - FD_M)/(FD_E - 2)$, for model–target comparison (black bars), template–target comparison (gray bars) and NMR/X-ray comparison (dotted line). (**C**) A schematic diagram representing the rugged surface of the model (bottom) and a relatively smoother surface of the experimental structure (top). The dark circles represent the probes used to trace the surface and α and β are surface pocket and internal cavity, respectively. Interior cavity and pockets become inaccessible as probe size increases, hence, the rate of change of surface area with an increase in probe size will be higher for a rugged protein surface, accounting for the higher FD of models (see text). (**D**) The surface (top) and backbone (bottom) of the X-ray structure (1sgp) and models of *S.griseus* protease B. Surface is colored by solvent accessibility using MOLMOL (37). Pockets are indicated by arrows in the X-ray structure and the lowest sequence identity model. The sequence identity, ΔASA/ASA, FD and Δ*V*/*V* of the models are indicated below.

of identical (or very similar) residues as is observed between residues of homologous proteins (55,56), but otherwise adopt conformations defined by MODELLER restraints on the $\chi_{1-4}$ dihedral angles (18). This, together with lack of solvent treatment on the side chains, seems to give rise to pocket artifacts. The increase in the number of pockets adds to the ruggedness of the protein surface contributing to the higher FD and ASA observed in models (previous sections). SAFD (26) was measured for the PASS-pockets of models and X-ray structures. We found that the SAFD of pockets in models is also larger than that of pockets in X-ray structures and shows a strong correlation with template:target sequence identity (data not shown). This implies that the larger FD observed in models is not only due to the larger number of pockets but also due to the increased roughness of pockets themselves.

The accuracy of detection of a pocket is the ratio between the number of identical pockets (see Methods) and the total number of pockets in a model. Although alignment errors did not have a clear effect on the number of pockets (Figure 7A), it appears to have some effect on the number of identical pockets, because STR models are slightly more accurate than SEQ models in terms of the fraction of correct pockets (Figure 7B).

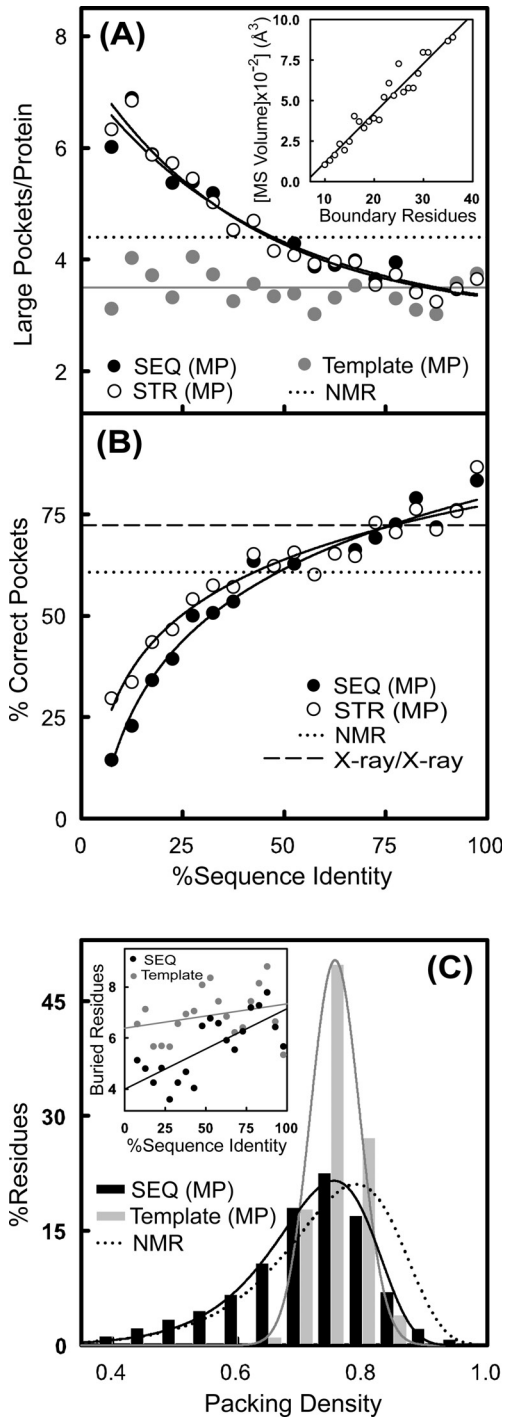


**Figure 7.** Surface pockets. Accuracy of number and identification of surface pockets is shown as a function of sequence identity. (**A**) Number of large pockets (pockets with 10 or more boundary residues, see text) per protein is compared among SEQ model (closed circles), STR model (open circles), template (gray circles) and NMR structures (dotted line). The inset shows the relationship between the number of boundary residues of a pocket and the pocket volume. (**B**) Comparison of the accuracy of pocket detection among SEQ models (closed circles), STR models (open circles) and NMR structures (dotted line). The dashed line shows the difference between independent X-ray structures of the same protein. (**C**) Distribution of residue packing density: SEQ models (closed bars), templates (gray bars) and NMR structures (dotted line). Inset figure shows the average number of buried residues (accessibility = 0.0) of templates (gray circles) and models (closed circles) of MP as a function of sequence identity.

## Packing density and cavity

Although on average the protein interior is well packed, there are variations in efficiency of packing (27). Low-packing density in general is due to the presence of cavities. On the other hand even high-resolution structures have atomic clashes (57). Here, the quality of packing in models is compared with that in experimental structures. The packing density of buried residues (see Methods) of the template (X-ray structure) shows a narrow distribution with an average value of ~0.75 (Figure 7C). It is to be noted that the number of buried residues for medium-sized models ranged from 4 to 7 for low- and high-sequence identity, respectively; whereas the corresponding X-ray templates had more or less a constant number of 7 buried residues (Figure 7C, inset). Since it is easier to pack fewer objects, packing a smaller number of buried residues may give a false impression of tighter packing in models. In fact, we observed that the packing density of the buried residues in models varies from 0.85 (low-sequence identity models) to 0.75 (high-sequence identity model) (data not shown). To have a comparison based on equal number of residues, the packing density of model residues corresponding to the buried residues of the template was determined, irrespective of the exposure state of those model residues. Since the model is based on the template, a buried residue in the template will generally correspond to a buried or a nearly buried residue in the model (12). The distribution of packing density of models is much broader compared to that of X-ray structures and extended towards lower values (Figure 7C) indicating that the overall packing of the model regions corresponding to the hydrophobic core of template structures are loosely packed, although some instances of over packed sites are also observed. The loosely packed residues in models are due to the presence of pockets and cavities. Low-packing density may also contribute to the higher FD observed in models (above, Figure 6C). The distribution of packing density of models is comparable in shape to that of NMR structures, with the NMR structures showing slightly higher packing values on average (Figure 7C, dotted line). Using a different method, an earlier packing analysis (58) on 70 proteins for which both X-ray and NMR structure were available, had showed that X-ray structures had packing values lying in a narrow range, whereas the NMR structures showed a much larger scatter in packing density values, consistent with the present observation. The inaccuracy in packing density we observe in models is similar to that seen in NMR structures. Hence, interiors of models are qualitatively similar to NMR structures, but the surface of models is more rugged.

## Salt-bridges

As a practical extension of the inter-residue distance, the accuracy with which the presence of salt-bridges can be determined from comparative models was analyzed. Salt-bridges are close-range electrostatic interactions formed by pairs of oppositely charged residues. Salt-bridges and ion pairs are predominantly found at surface exposed sites (35,59). Because the accuracy of inter-residue distance at exposed sites is lower than at buried sites (Figure 2A), the accuracy of salt-bridge identification is expected to be low. The number of salt-bridges per model increases with template:target sequence identity (Figure 8A). At very high-sequence identity most

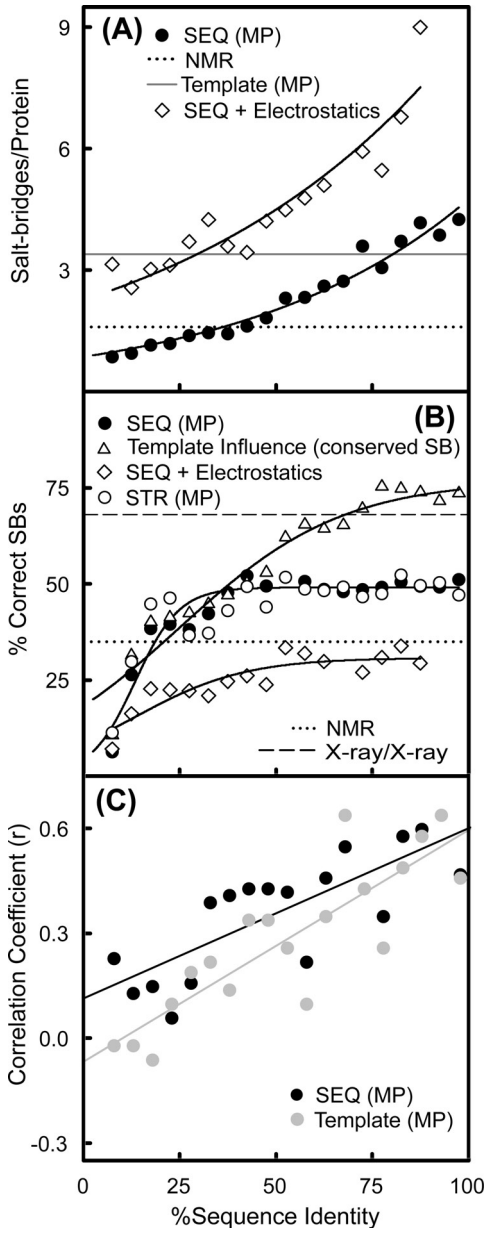


**Figure 8.** Salt-bridge. Accuracy of salt-bridge identification as a function of template:target sequence identity is shown. (**A**) Number of salt-bridges per protein for models calculated with (diamonds) and without (circles) electrostatic interactions (see text). (**B**) Comparison of accuracy between SEQ models (open circles), STR models (closed circles), SEQ models with electrostatics interaction (open diamonds) and the number of conserved template salt-bridges is shown (open triangles). (**C**) Linear correlation coefficient between the number of salt-bridges in the target experimental structure and SEQ models (closed circles) or templates (gray circles).

side chains are conserved and adopt the conformation of the equivalent template residue, thus maintaining the number of salt-bridges, but below 70% sequence identity the average number in models is lower than in their templates (Figure 8A). To determine whether the loss of salt-bridges was due to the absence of an electrostatic interaction term in the default objective function used for model building (see Methods), a new set of models was built after switching on the electrostatic term in the MODELLER objective function (18). In this case, the number of salt-bridges increases by almost a constant factor over the whole-sequence identity range (Figure 8A), even surpassing the number observed for the experimental structure. This suggests that the use of the electrostatic term is generating many false salt-bridges in the models. The accuracy measured as the fraction of identical salt-bridges in the two structures, is quite low compared to other features discussed so far (Figure 8B). The lower accuracy is not only due to surface exposure but also due to the fact that salt-bridges involve interaction between long side chains (Arg, Lys, Glu) with more torsional degrees of freedom (53). It is to be noted that even for X-ray/X-ray pairs (see Methods), the accuracy is only 67%. Above 25% sequence identity the accuracy is constant and is unaffected by alignment error (Figure 8B). Even though the number of salt-bridges increases when the electrostatic interactions are turned on, the accuracy is even lower indicating that incorrect ion-pairs have been forced to interact. The simple treatment of electrostatics in MODELLER (18) is probably not sufficient to form correct salt-bridges in the models. The corresponding comparison for NMR/X-ray pairs shows a large difference between estimates of salt-bridges in NMR and X-ray structures (Figure 8B), this is in agreement with previously published estimates based on smaller sets of structures (60). The most frequent atoms among those showing a large difference in area between the corresponding NMR/X-ray pairs are the terminal side-chain atoms of Arg, Lys, Glu, indicating that many side-chain-mediated ionic interactions in an X-ray structure are absent in the corresponding NMR structure. Salt-bridges in the model that are also present in the template are defined as conserved. The proportion of the conserved salt-bridges is higher than that of the accurately predicted ones indicating that the template influences the side chains of the target sequence to adopt conformation like itself favoring salt-bridge formation even though these are not detected in the experimental structure of the target (Figure 8B). This also explains the dependence of the number of salt-bridges in the model on their number in the template. In spite of this, the number of salt-bridges in the models is slightly better correlated with that in their experimental target structures rather than the template (Figure 8C), thus being another example of added-value in comparative models (12). An earlier large-scale genome analysis had used models instead of alignments to show that thermophilic proteins have more salt-bridges than mesophilic homologs with an underlying assumption that models would be more informative than alignments in this respect (7). This study confirms the earlier assumption.

## Electrostatic potential

Calculations of electrostatic potential in protein structures are frequently used to identify regions of positive or negative charge that may represent binding pockets or active sites (61,62). We calculated the accuracy of the electrostatic potential by comparing the 3D grid resulting from the electrostatic potential calculation of models with the grid obtained from the experimental target structure (see Methods). The electrostatic potential similarity is measured by the correlation of potential values between equivalent positions in the pair of grids. Figure 9A shows the correlation coefficient obtained when comparing the electrostatic potential of SEQ models (closed circles), STR models (open circles) and NMR structures

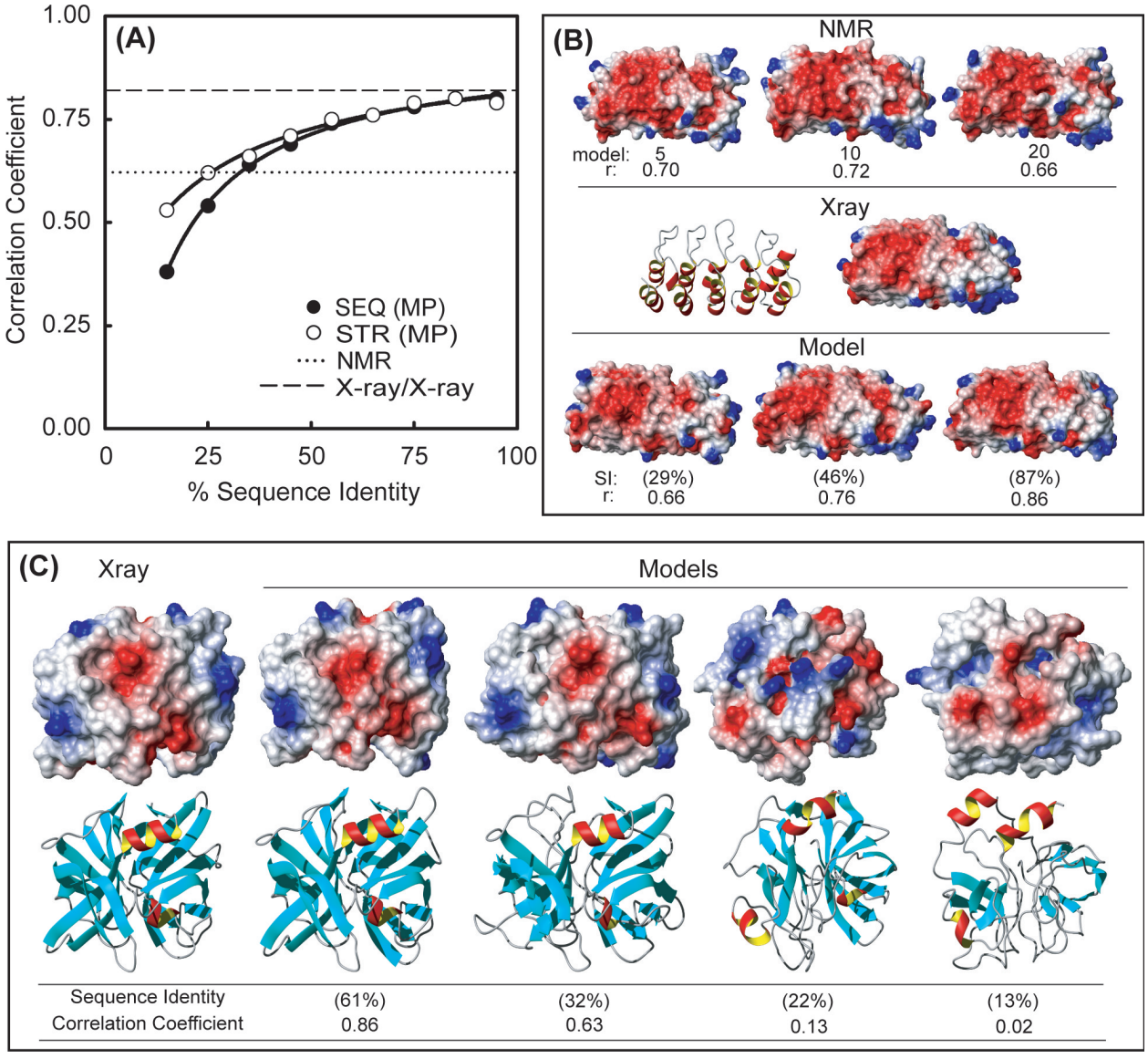


**Figure 9.** Electrostatic potential. (**A**) Accuracy of electrostatic potential as a function of template:target sequence identity. The accuracy is measured by the rank correlation coefficient between the values of the electrostatic potential of the target and models (SEQ, closed circles and STR, open circles). (**B**) Electrostatic potential colored surfaces of the NMR structure (1ap7, top), X-ray structure (1blx, middle) and SEQ models (bottom) of cell cycle inhibitor $p19^{INK4d}$. Surfaces of the 5th, 10th and 20th models of the NMR structure file are shown. The sequence identities of the SEQ models are indicated in parentheses. (**C**) Surface colored by electrostatic potential (top) and backbone (bottom) of the X-ray structure (1sgp) and models of *S.griseus* protease B. The figure was created using program MOLMOL (37).

(dotted line) with their corresponding X-ray structures. Only medium-sized proteins were analyzed here. As expected, the electrostatic potential accuracy drops with decreasing template:target sequence identity, with alignment errors starting to affect the accuracy below 50% sequence identity. The NMR/X-ray pair difference is comparable to models based on 30% sequence identity. Along with the conformational variation among NMR ensembles, the larger volume of NMR structures (Figure 5D) also affects the grid comparison resulting in a poor correlation. X-ray/X-ray pair correlation is 0.83. As an illustration the electrostatic potential colored surface of the X-ray structure, NMR structure and models of the mouse cell cycle inhibitor $p19^{INK4d}$ is shown in Figure 9B.

## DISCUSSION

Overall structural similarity, as measured by RMSD or number of equivalent atoms (and combinations of these measures), has been primarily and routinely used for comparing experimental protein structures and in the assessment of comparative models (1,8–10,41). Most studies consider lowering RMSD as a significant step towards quality improvement in structure prediction. It is obvious that improving the overall accuracy of models, no matter how it is measured, is useful. However, since models are not perfect the question remains as to what impact the errors have on structure-derived properties of practical interest, such as exposure state or electrostatics. By directly analyzing the accuracy of several structure-derived

properties of practical interest, as a function of template:target similarity, we find that all properties do not behave in the same way. In all cases the accuracy drops as a function of template: target sequence similarity, mainly by the contribution of two factors: (i) the relationship between the divergence of sequence and structure in proteins (43) and (ii) the increasing number of errors in the template:target alignment as the sequence similarity decreases. However, the exact shape of this relationship is not the same for different structure-derived properties; and the impact of alignment errors and the protein size dependence also varies from one property to another, thus rendering the overall accuracy measure an ineffective tool to predict the accuracy of specific structure-derived properties.

Comparison of the accuracy of structure-derived properties of models with the difference observed between NMR and X-ray structures for the same proteins shows that in the 25–40% sequence identity range accuracies of comparative models reach the same value as NMR/X-ray differences. The exact boundary depends on the particular feature being measured, e.g. salt-bridge and pocket accuracy. This suggests that depending on the feature being analyzed many comparative models may provide information that is comparable in accuracy to that derived from structures determined by NMR spectroscopy. It is important to note that only the magnitude of the difference is similar in these cases, but the nature of the difference is not necessarily the same for comparative models and NMR structures. Because X-ray structures were used as the accuracy reference in this work, it may seem that the average accuracy of models that use NMR structures as templates, even at high-sequence identity, would at most be as high as the accuracy of models based on X-ray templates in the 30–40% sequence identity range. This would only be true if the aim of comparative modeling was to produce protein structure models that are as close as possible to the high-resolution X-ray structure of the protein in its crystalline environment. However, the statement could also be reversed and if the aim of comparative modeling is to produce models that represent the solution structure of proteins as described by high-resolution NMR structures then X-ray structures would be the poorer templates. In any modeling case, a decision has to be made as to what environment the protein is being modeled in. This decision should direct both our choice of template (NMR or X-ray) and our interpretation of the model accuracy. This is particularly relevant when analyzing surface-related properties that show larger differences between NMR and X-ray structures and are more influenced by the environment in which the structure is determined. In spite of these environmental effects and the dynamic nature of protein surfaces, a clear dependence on template:target sequence similarity was observed for the accuracy of all surface-related properties. This indicates that the conservation of surface features between homologous proteins provides information about the surface-related properties that is detectable even when considering the above mentioned effects.

The models used here have not only been built using a single program but also have not been refined further [e.g. loop modeling (63) was not used]. In the absence of refinement, comparative models follow their templates very closely and their accuracy is mainly determined by the template:target similarity and alignment accuracy. Under these conditions, the differences between alternative model-building methods are negligible (39,64). Hence, the set of models used here represents the simplest type of comparative model, providing a baseline 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 (1,65). It is expected that the use of multiple templates, which allows comparative modeling to select the best parts from different structures to build the target model (64), would provide an improvement in accuracy. Anecdotal evidence indicates that this is the case (64), but no systematic study has been performed. Refinement of models, in the form of loop modeling (63), may also provide an improvement over the simple models presented here. How much loop modeling affects the accuracy of different structure-derived properties 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 simple models used here.

A general picture that emerges from our analysis is that the surface of models is very different from the surface of natural proteins in terms of the number of pockets it contains and its general ruggedness, as measured by its FD. Because FD of the models linearly increases with diminishing template:target identity (Figure 6A), even cases with very few or no insertions have elevated FD. This indicates that incorrectly modeled loops are not the only reason, and probably not even the main contributor, for increased FD. A more general refinement of the surface of models is needed. Hence, one interesting approach would be to perform molecular dynamics (MD) simulations with explicit (or implicit) solvent and see whether there is an improvement in the surface properties of models. MD-based refinement of models is non-trivial and several studies have been published (66,67). Unfortunately, these studies have concentrated on evaluating the improvement of the overall accuracy of models and not the potential improvement of the model's surface, which is where these simulations could potentially have the greatest impact. It would therefore be informative if future studies of model accuracy and refinement included explicit measures of SDP accuracy, similar to those described here.

## ACKNOWLEDGEMENTS

We thank Carlos Madrid for general assistance with hardware and software; Dr Patrice Koehl for help with the ProShape suite of programs; 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 NIH. Funding to pay the Open Access publication charges for this article was provided by Mount Sinai School of Medicine Startup Funds.

## REFERENCES

1. Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
2. Xu,L.Z., Sanchez,R., Sali,A. and Heintz,N. (1996) Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.*, **271**, 24711–24719.
3. Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.

4. Wallace,A.C., Borkakoti,N. and Thornton,J.M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*., **6**, 2308–2323.
5. Liu,T., Rojas,A., Ye,Y. and 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.
6. Maggio,E.T. and Ramnarayan,K. (2001) Recent developments in computational proteomics. *Trends Biotechnol*., **19**, 266–272.
7. 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.
8. Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53**, 352–368.
9. 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*, **53**(Suppl. 6), 503–516.
10. Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (2003) Assessment of progress over the CASP experiments. *Proteins*, **53**(Suppl. 6), 585–595.
11. Sanchez,R. and Sali,A. (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods Mol. Biol*., **143**, 97–129.
12. Chakravarty,S. and Sanchez,R. (2004) Systematic analysis of added-value in simple comparative models of protein structure. *Structure (Cambridge)*, **12**, 1461–1470.
13. Andrade,M.A., O'Donoghue,S.I. and Rost,B. (1998) Adaptation of protein surfaces to subcellular location. *J. Mol. Biol*., **276**, 517–525.
14. 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.
15. 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.
16. 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.
17. 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.
18. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.
19. Doreleijers,J.F., Raves,M.L., Rullmann,T. and Kaptein,R. (1999) Completeness of NOEs in protein structure: a statistical analysis of NMR. *J. Biomol. NMR*, **14**, 123–132.
20. Laskowski,R.A., Rullmannn,J.A., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
21. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*., **55**, 379–400.
22. Hubbard,S.J. and Thornton,J.M. (1993) NACCESS: Computer Program. Computer Program, Department of Biochemistry and Molecular Biology, University College London.
23. 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.
24. Sanner,M.F., Olson,A.J. and Spehner,J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.
25. Lewis,M. and Rees,D.C. (1985) Fractal surfaces of proteins. *Science*, **230**, 1163–1165.
26. Pettit,F.K. and Bowie,J.U. (1999) Protein surface roughness and small molecular binding sites. *J. Mol. Biol*., **285**, 1377–1382.
27. Richards,F.M. (1977) Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng*., **6**, 151–176.
28. Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol*., **105**, 1–12.
29. 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.
30. Richards,F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol*., **82**, 1–14.
31. Creighton,T. (1993) *Proteins: Structures and Molecular Properties*. W. H. Freeman and Co., NY.
32. Voronoi,G. (1907) Nouveles applications des parameters continus a la theorie des formes quadratiques. *J. Reine Angewandte Mathematik*, **133**, 97–178.
33. Harpaz,Y., Gerstein,M. and Chothia,C. (1994) Volume changes on protein folding. *Structure*, **2**, 641–649.
34. Edelsbrunner,H. and Koehl,P. (2003) The weighted-volume derivative of a space-filling diagram. *Proc. Natl Acad. Sci. USA*, **100**, 2203–2208.
35. Kumar,S. and Nussinov,R. (1999) Salt bridge stability in monomeric proteins. *J. Mol. Biol*., **293**, 1241–1255.
36. 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.
37. Koradi,R., Billeter,M. and Wuthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph*., **14**, 51–55.
38. Jacobson,M.P., Friesner,R.A., Xiang,Z. and Honig,B. (2002) On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol*., **320**, 597–608.
39. 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.
40. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
41. 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.
42. Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Rost,B. and Sali,A. (2002) Reliability of assessment of protein structure prediction methods. *Structure (Cambridge)*, **10**, 435–440.
43. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J*., **5**, 823–826.
44. Chung,S.Y. and Subbiah,S. (1999) Validation of NMR side-chain conformations by packing calculations. *Proteins*, **35**, 184–194.
45. Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol*., **213**, 859–883.
46. 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.
47. Ponstingl,H., Henrick,K. and Thornton,J.M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.
48. Janin,J. (1997) Specific versus non-specific contacts in protein crystals. *Nature Struct. Biol*., **4**, 973–974.
49. Boniface,J.J., Reich,Z., Lyons,D.S. and Davis,M.M. (1999) Thermodynamics of T cell receptor binding to peptide-MHC: evidence for a general mechanism of molecular scanning. *Proc. Natl Acad. Sci. USA*, **96**, 11446–11451.
50. Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
51. Timchenko,A.A., Galzitskaya,O.V. and Serdyuk,I.N. (1997) Roughness of the globular protein surface: analysis of high resolution X-ray data. *Proteins*, **28**, 194–201.
52. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*., **12**, 2001–2014.
53. Bower,M.J., Cohen,F.E. and Dunbrack,R.L.Jr (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol*., **267**, 1268–1282.
54. 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.
55. Flores,T.P., Orengo,C.A., Moss,D.S. and Thornton,J.M. (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci*., **2**, 1811–1826.
56. Sali,A. and Overington,J.P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci*., **3**, 1582–1596.

57. Word,J.M., Lovell,S.C., LaBean,T.H., Taylor,H.C., Zalis,M.E., Presley,B.K., Richardson,J.S. and Richardson,D.C. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.*, **285**, 1711–1733.
58. Ratnaparkhi,G.S., Ramachandran,S., Udgaonkar,J.B. and Varadarajan,R. (1998) Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry*, **37**, 6958–6966.
59. Barlow,D.J. and Thornton,J.M. (1983) Ion-pairs in proteins. *J. Mol. Biol.*, **168**, 867–885.
60. Kumar,S. and Nussinov,R. (2001) Fluctuations in ion pairs and their stabilities in proteins. *Proteins*, **43**, 433–454.
61. Honig,B. and Nicholls,A. (1995) Classical electrostatics in biology and chemistry. *Science*, **268**, 1144–1149.
62. 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.
63. Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.
64. Sanchez,R. and Sali,A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **Suppl. 1**, 50–58.
65. Peitsch,M.C., Schwede,T. and Guex,N. (2000) Automated protein modelling–the proteome in 3D. *Pharmacogenomics*, **1**, 257–266.
66. Fan,H. and Mark,A.E. (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci.*, **13**, 211–220.
67. Flohil,J.A., Vriend,G. and Berendsen,H.J. (2002) Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins*, **48**, 593–604.