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

Suvobrata Chakravarty, Dario Ghersi and Roberto Sanchez

Electronic Supplementary Material

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

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

D. Ghersi
Current address: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA.
SUPPLEMENTARY INFORMATION

Geometric factor and Alignment factor
As protein structure of one class primarily differs from those of the others in the geometry of the backbone we wanted to see the effect of the alignment error on isolated secondary structural elements. The distance between C$_{\alpha}$ atoms separated by two or three positions in the sequence, varies in different structural elements (Supplementary Figure S2 B). Due to this, the effect of a similar alignment error in the modeled segments of different structural elements may result in numerically different values of C$_{\alpha}$-RMSD. In a simplified example (Supplementary Figure S1), a single shift in the alignment of different structural segments of identical length results in different values of RMSD. An isolated structural segment named segP is superposed onto itself guided by the shifted sequence alignment, to highlight the impact of the same level of alignment error on different backbone architectures. In this simple example, an $\alpha$-helical segment has the smallest C$_{\alpha}$-RMSD (~0.5Å), followed by a $\beta$-strand (~1.5Å), while the remaining segments have C$_{\alpha}$-RMSD of 3Å or more. As expected, when compared to isolated $\alpha$-helix or $\beta$-strand elements, the RMSD of loop regions and segments involving sharp turns, or that change the direction of the polypeptide chain (such as helix-turn-helix or strand-turn-strand patterns) is affected to a larger degree by the alignment error. Even though oversimplified, this example highlights that the impact of the alignment error on model accuracy may depend on the geometry of the polypeptide backbone. We refer to this as the Geometric factor that affects RMSD. Due to the geometric factor, one may expect that the accuracy of comparative models of proteins with different contents/composition of these structural elements will be different, given the same level of alignment accuracy. All-$\alpha$ proteins are composed of helical segments and inter-helical connectivity, whereas all-$\beta$ proteins contain strands and inter-strand connectivity. The $\alpha\beta$ group is a mixture of helical segments, strands and connectivity between them. A number of structural properties such as the number of secondary structural elements per protein, contact order (Plaxco, et al., 1998), average length of secondary structural elements etc, differ substantially between these groups (Supplementary Figure S2). For the same chain
length, all-α proteins have fewer secondary structural elements (helical segments), as the average length of α-helices is longer than that of β-strands and loops. Therefore, they have a smaller number of connectivity elements (Supplementary Figure S2). On the other hand all-β proteins possess a higher number of structural elements (β-strands) and connectivity (Supplementary Figure S2). Hence, due to the structural differences discussed above, one can expect the model accuracy of proteins belonging to different structural classes to be different at the same level of alignment accuracy.

In addition to the geometric factor, other factors may also play a role in influencing the accuracy of models belonging to different structural classes. For example, there may be inherent compositional differences in protein sequences belonging to different structural classes. As noted earlier, even a simple binary pattern (hydrophobic-polar, HP) analysis had showed a stronger sequence-structure correlation for α-helices than for β-strands (West and Hecht, 1995). This association indicates that there might be recognizable differences in amino-acid compositional signatures in the sequences of all-α, all-β and αβ. These differences may in turn influence the quality of alignment in a structural class dependent manner due to differences in the amount of information in the sequences. We refer to this as the alignment-dependent influence on model accuracy or simply Alignment factor. Influence of structural topology on the amount of information in sequence profiles supports this (Koehl and Levitt, 2002). Unlike the alignment factor, the geometric factor is of purely structural origin. Hence, its effect can be viewed as a ‘non-alignment’ influence on model accuracy. The results presented in this study are discussed in the light of these two contrasting factors.

Cα-RMSD between the modeled and experimental structure was taken as the measure of the total error (Eqn.-1). As there was no refinement involved in the modeling procedure, keeping the structural divergence nearly identical, the study was designed to see if additional factor, the Geometric factor (see above) would refine the Eqn.-1 to that shown below:

\[
\text{Total Error} = \text{Structural Difference} + \text{Alignment Error} + \text{Geometric Factor} - \text{Refinement} \quad [2]
\]
We observe that the contribution of the Geometric factor is negligibly small and the alignment factor contributes maximally to the observed difference in the accuracy of proteins belonging to different structural Fold Classes. Hence, Eqn-1 is sufficient to describe modeling error without a need to invoke Eqn.-2.

**Structure derived properties of models**

The effect of alignment error on the accuracy of structure-derived properties depends on the nature of the property. High-resolution features such as distance based properties: inter-atomic, inter-residue distances, residue neighborhood etc. are the most affected whereas low-resolution features such as the electrostatic potential surface and the number of false pockets are affected to a lesser degree by alignment error (Chakravarty, et al., 2005). A plot of the accuracy of residue neighborhood (average of buried and exposed residues) shows that proteins belonging to the all-β and all-α class are affected the most and the least, respectively (Supplementary Figure S3). It is important to note that in the extended conformation such as β-strand, the Cα-Cβ vectors of adjacent residues point in opposite directions; hence, an alignment error such as a shift will change the direction of this vector to a greater degree in a β-strand than in an α-helix, resulting in a lower accuracy of neighborhood in all-β proteins. The higher relative contact order (rCO) (Supplementary Figure S2, C left) for all-β proteins indicates a higher proportion of inter atomic contacts, per unit length of protein, between residues far apart in sequence. Hence, alignment errors in strands will result in more non-native contacts in the modeled segment compared to that of helices, resulting in poorer neighborhood prediction accuracy.
SUPPLEMENTARY METHODS

Dataset
Single domain chains (size: 100-150 residues), as defined by CATH (Orengo, et al., 1997), of high-resolution X-ray structures (2.5Å or better) were selected from the Protein Data Bank (Berman, et al., 2002). Chains belonging to all-α, all-β and αβ classes were sorted based on the first of the four digits of the CATH domain identification number. Members within the same homologous superfamily (same value of the first four CATH levels) were clustered such that alignments with 95% or higher sequence identity with each other and at least 85% of sequence coverage were grouped together. Only the highest resolution member of each group was retained as a representative, resulting in a non-redundant homologous superfamily. The representative chains within the selected homologous superfamily were structurally aligned with each other using CE (Shindyalov and Bourne, 1998), and only pairs of chains with a CE Z-score higher than 4.5 were retained.

Profile Entropy
The Shannon entropy of a profile is estimated by measuring profile column entropy, p×log₂p, where p is the distribution of fraction of 20 amino acid + Gap in a profile-column.

Contact Density and Relative Contact Order
The contact density (CD) was computed as the top eigenvalues of the inter residue contact matrix. A protein with N residues results an N×N symmetric contact matrix with matrix element, eᵢⱼ, representing a ‘binary’ distance measure between residue i and j. eᵢⱼ = 1 when the distance between the Cβ atoms of residue i and j ≤ 6Å, and eᵢⱼ = 0 otherwise. If either i or j is Glycine, the Cα atom is used to compute distance. The eigenvalues of the contact matrix were computed with the 'eigen' function of the statistical package 'R'. Relative contact order rCO was computed using the perl script downloaded from http://depts.washington.edu/bakerpg (Plaxco, et al., 1998).
**Geometric Factor Analysis**

For the superposition experiment of Supplementary Figure S1, a 15-residue segment is chosen that is long enough to capture the essence of a protein structure in terms of the secondary structural elements. For each type of structural segment, pair-wise superposition of 50 different 15-residue long structural segments, were carried out. Cα RMSD was computed by measuring the deviation between the $i$th-Cα in the ‘Red’ structure with the corresponding $i$th-Cα in the ‘Blue’ structure in Supplementary Figure S1.
REFERENCES


SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure S1: The distribution of RMSD of superposition of different structural segments: Each of the different structural segments (15 residues long), in order from top to bottom, loosely defined as single-helix, helix-loop-helix, single-strand, strand-loop-strand and loop is superposed on to itself based on the corresponding shifted alignment. The distribution is generated from pair-wise superposition of 50 different structures (15-residue long) of each of the above segments. The pictured segments are one of the 50 selected cases for each structural segment. Red represents the selected segment and blue, the resultant structure after superposition based on the corresponding shifted alignment. Cα RMSD is computed by measuring the deviation between the \(i\)th-Cα in the ‘Red’ structure with the corresponding \(i\)th-Cα in the ‘Blue’ structure.

Supplementary Figure S2: Comparison of structural properties between all-α, all-β and \(αβ\) proteins: (A) Examples of structures of all-α, all-β and \(αβ\) proteins. (B) The distance distribution of Cα-Cα \((i, i+1; \text{left}), \text{Cα-Cα (}i, i+2; \text{middle}) and \text{Cα-Cα (}i, i+3; \text{right}) in all-α, all-β and \(αβ\) proteins. (C) The distribution of proteins with certain number of secondary structural elements (helix and strand) (left); the distribution of relative contact order (rCO) proteins in the dataset (right). (D) Distribution of the lengths of secondary structural elements.

Supplementary Figure S3: Accuracy of structure derived properties (SDPs): (A) Comparison of accuracy of neighborhood prediction between PRO-models of all-α, \(αβ\), and all-β proteins. (B) Comparison of accuracy of excess surface pockets per protein in PRO-models of all-α, \(αβ\), and all-β proteins. Inset is the representation of a false pocket due to improper modeling of side-chains that is not dependent on topology.

Supplementary Figure S4: Distribution of contact density (CD) values of all-α, all-β and \(αβ\) proteins.
SUPPLEMENTARY FIGURE 1

Chakravarty et al. 2010
SUPPLEMENTARY FIGURE 3

(A) %Correct vs. %Sequence Identity

(B) [Pockets]/Protein vs. %Sequence Identity

Chakravarty et al. 2010