Supporting Information

Prasad et al. 10.1073/pnas.0812152106

SI Text

Data. The methodology described below closely follows the detailed description in our previous publication and we will only summarize here the main points (1). Our database of redundant protein structures is derived from the collection of X-ray structures deposited in the PDB on or before January 2007. Different chains in the same PDB entry (labeled by the same PDB ID) are treated as different entries.

To minimize the effects of low resolution and the changeable technologies of early years of protein crystallography, we used 3 criteria to accept the structures to the analysis. They are the following: (*i*) deposited after 1990; (*ii*) resolution higher than 2.5 Å; and (*iii*) *R*-value lower then 0.25.

Applying these criteria eliminated a large class of lower resolution and higher R factor structures (such as hemagglutinins) that naturally represent higher mobility. As a result, the obtained distribution constitutes only a lower bound for the natural distribution recorded in the PDB.

Finding and Aligning the 100% Identical Structures. We used "SEQRES" records of all of the PDB entries to identify identical proteins. Subsequently we compared the "ATOM" and "SEQRES" records of each PDB entry to check for the presence of atomic coordinates representing a particular sequence. An alignment of sequences parsed from "SEQRES" and "ATOM" tags was performed using a "blast2seq" program in the National Center for Biotechnology Information (NCBI) toolkit (2). We found that in more than half of the structures present in the PDB these 2 differ! We created a database of redundant structures that contains a multiple sequence alignment of sequence of all proteins in each cluster.

Removal Sequence Artifacts (His-tags, Sel-Met, and Other Special Cases). Many proteins are cloned and later crystallized with His-tags. To remove the His-tags, we used the "blastp" program (3) by identifying sequences of the wild type protein. Every sequence in the previously prepared redundant PDB database was checked using Blast against Swiss-Prot (4).

Similarly, N-terminal Met residue has been omitted from our analysis. In these cases where Sel_Met was used for structure phasing and identified in the sequence record, we treated it as Met residues to avoid a mismatch caused by a change of name in the ATOM record.

RMSD Calculations. The RMSD calculations were performed by the algorithm in an integrated environment of BOS (v3.0). We first constructed a complete chain alignment that took into account the incomplete structural representation of individual models.

1. We used the sequence-based structure alignment specified above for identical sequences to form an individual cluster that represented a particular protein. The coordinates of identical residue pairs were used in our RMSD calculations using the implementation of the Kabsh algorithm (5).
2. RMSD calculations were performed for each cluster. Subsequently, the implementation of the clustering procedure of the UPGMA algorithm (unweighted pair group method with arithmetic mean), a bioinformatics method of phylogenetic tree reconstruction, was used. Clustered RMSDs that determined the branches of a conformational states tree were used to determine the number of distinct conformational states. The pseudophylogenetic tree derived from these clusters was used to determine the nodes of conformational speciation or the number of conformational states that correspond to a given RMSD cutoff. The distribution of those states is represented in Fig. 5.
3. Local structural fragment alignment. In this method, for 2 aligned chains (based on sequence) the coordinates of every 25 amino acid pairs were structurally aligned. The RMSD was calculated for each fragment. This calculation resulted in approximately 1 million data points of all pairs of structural fragments.

1. Zhang Y, Stec B, Godzik A (2007) Between order and disorder in protein structures: analysis of ''dual personality'' fragments in proteins. *Structure* 15:1141–1147.
2. Tatusova TA, Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174:247–250.
3. McGinnis S, Madden TL (2004) BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32:W20–W25.
4. Gasteiger E, Jung E, Bairoch A (2001) SWISS-PROT: Connecting biomolecular knowledge via a protein database. *Curr Issues Mol Biol* 3:47–55.
5. Kabsh W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

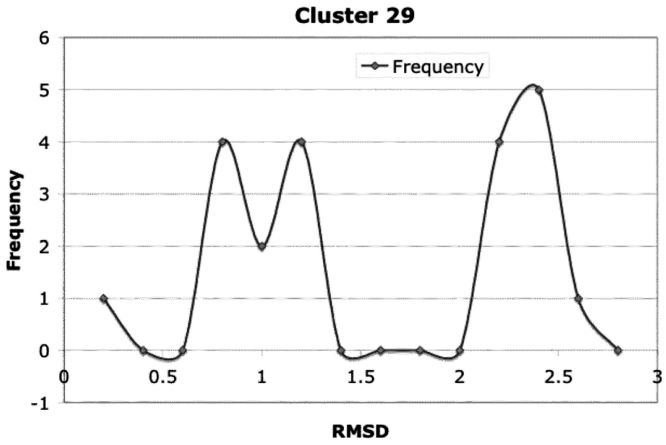

Fig. S1. Example of pair-wise RMSD frequency distribution for cluster 29. Multiple peaks correspond to multiple conformational states in the cluster.

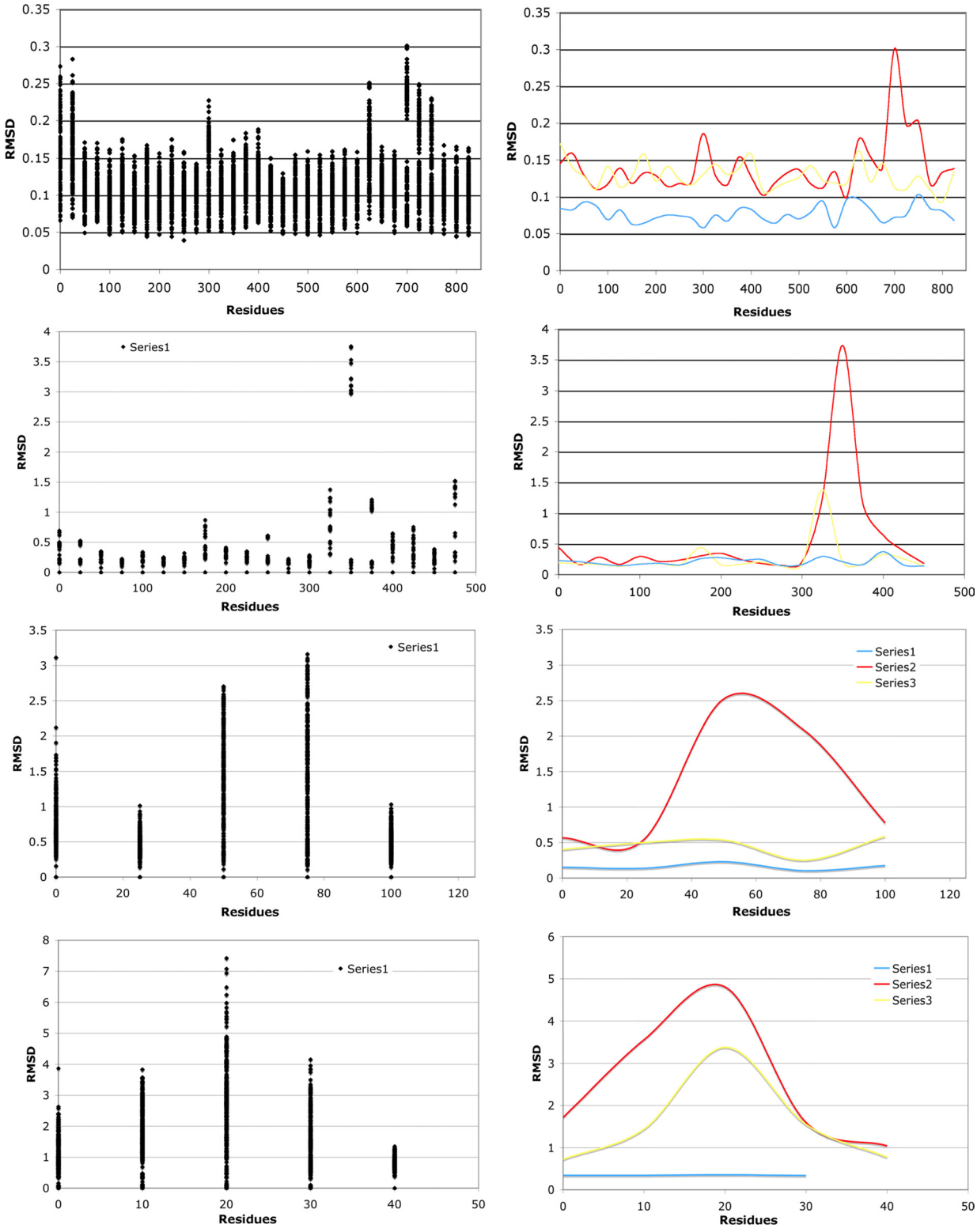

Fig. S2. Examples of distribution of sliding window RMSDs determined on 25 amino acids with progression of 10 amino acids. Panels *A* and *B* represent cluster 48 (transhydroxylase), panels *C* and *D* represent cluster 633 (diphtheria toxin), panels *E* and *F* represent cluster 8,791 (calmodulin), and panels *G* and *H* represent cluster 11,575 (apolipoprotein AII). Panels *A*, *C*, *E*, and *G* show all pair-wise window RMSDs in individual clusters distributed over the sequence. Panels *B*, *D*, *F*, and *H* present examples of 3 conformational states with models with small (yellow), intermediate (red), and large divergence (blue).