Supporting Information

Nugent and Jones 10.1073/pnas.1120036109

SI Methods.

Calculation of Predicted Secondary Structure Regular Expression Strings. Part of the input to FRAGFOLD and FILM3 is a string that allows fragments to be filtered based on predicted secondary structure. The string is essentially a simple regular expression that permits HELIX, COIL, or STRAND flags to be set. For example HELIX|COIL indicates that HELIX or COIL would be permitted at that position.

Rather than using PSIPRED alone to calculate this regular expression string, as is the case for FRAGFOLD, in FILM3, we also made use of MEMSAT-SVM predictions of transmembrane helices. Predictions were combined using a simple consensus scheme with scoring thresholds for the two methods optimized using 99 TMPs (Table S6) of known structure that had insufficient homologous sequences available to be used as prediction targets. We further checked that these proteins had no detectable sequence homology to the targets (E-value < 0.001) or were members of the same OPM (1) superfamily. Raw residue preference scores for each method were used to determine the ensemble with strong transmembrane helix predictions overriding PSIPRED predictions. Where MEMSAT-SVM did not predict helix, the ensemble was constructed using HELIX, COIL, or HELIX|COIL depending on PSIPRED confidence; whereas, STRAND was only used in rare cases where PSIPRED confidence was high. Additionally, a small amount of coil was enforced in the center of predicted transmembrane loops if it did not already exist in the ensemble.

The complete simple decision tree rule-set is as follows:

IF MEMSAT-SVM predicts HELIX with raw score > 1.15, but PSIPRED predicts COIL, the ensemble prediction will be HELIX.

IF MEMSAT-SVM predicts HELIX but PSIPRED predicts STRAND, the ensemble prediction will be HELIX.

IF MEMSAT-SVM predicts a signal peptide but PSIPRED predicts a STRAND, the ensemble prediction will be COIL.

IF MEMSAT-SVM does not predict HELIX:

- **IF PSIPRED predicts HELIX:**
  - **IF the PSIPRED HELIX score > 0.87, the ensemble prediction will be HELIX**
    - **ELSE prediction will be HELIX|COIL**
- **IF PSIPRED predicts a STRAND:**
  - **IF PSIPRED STRAND score > 0.96, the ensemble prediction will be STRAND**
    - **ELSE prediction will be HELIX|COIL|STRAND**
- **IF PSIPRED predicts a coil:**
  - **IF the PSIPRED helix score > 0.72, the ensemble prediction will be**
    - **ELSE prediction will be HELIX|COIL**

1. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: Orientations of proteins in membranes database. *Bioinformatics* 22:623–625.

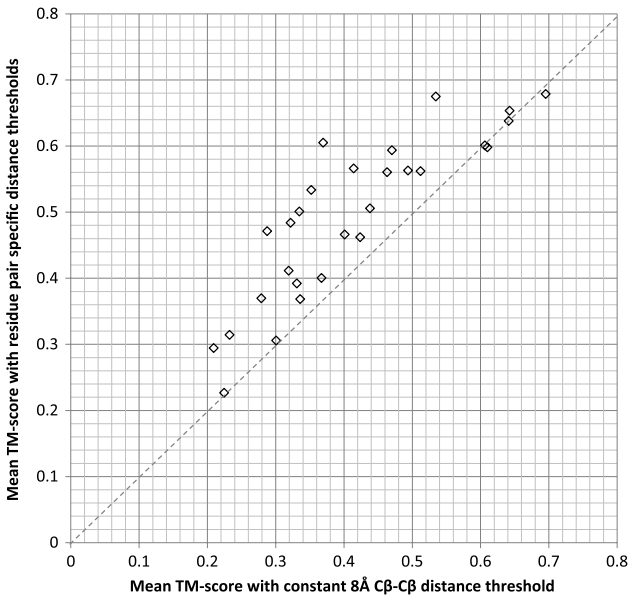

Fig. S1. Mean TM-score for all 200 models for each target using residue pair specific distance thresholds plotted against mean TM-score for all 200 models for each target using the standard constant Cβ-Cβ distance threshold of 8 Å.

Other Supporting Information Files

Table S1. Table of alignment summary statistics for the 28 target proteins. For each target, the chain length and number of aligned sequences is given, as before. Alongside these values are the minimum and mean percentage amino acid sequence identities between each sequence and the target, along with a count of the number of sequences which have ≥30% sequence identity to the target protein sequence (N_{30}).
Table S1 (DOCX)

Table S2. Complete list of 244 high-resolution protein structures from which fragments were selected during the initial fragment assembly stage. In each case, the PDB code, chain identifier, chain length, and brief description is given.
Table S2 (DOCX)

Table S3. List of 99 TMP chains of known structure used to calculate thresholds for the secondary structure decision tree.
Table S3 (DOCX)

Table S4. Maximal Cβ-Cβ distance thresholds (Å) for all amino acid pairs. Where in sufficient pairs were observed a default maximum value of 10.0 Å over all pairs is given.
Table S4 (DOCX)

Table S5. Satisfaction of contacts predicted using PSICOV in native structures based on the table of maximum Cβ-Cβ thresholds. Columns 2–5 show the absolute number of predicted contacts (column 2) followed by contact precision at three ranges of sequence separation (columns 3–5) where PSICOV PPV for these predictions is in the range $0.5 \leq PPV \leq 0.75$. Columns 6–9 show the absolute number of predicted contacts (column 6) followed by contact precision at three ranges of sequence separation (columns 7–9) where PSICOV PPV for these predictions is >0.75.
Table S5 (DOCX)

Table S6. Satisfaction of contacts predicted using PSICOV in refined models based on the table of maximum Cβ-Cβ thresholds. Columns 2–5 show the absolute number of predicted contacts (column 2) followed by contact precision at three ranges of sequence separation (columns 3–5) where PSICOV PPV for these predictions is in the range $0.5 \leq PPV \leq 0.75$. Columns 6–9 show the absolute number of predicted contacts (column 6) followed by contact precision at three ranges of sequence separation (columns 7–9) where PSICOV PPV for these predictions is >0.75.
Table S6 (DOCX)