Supplementary Methods

1. Threading program PPAS

PPAS is a sequence profile–profile alignment approach assisted with secondary structure matches. The scoring function of PPAS is defined by

$$\text{Score}_{\text{PPAS}}(i,j) = \sum_{k=1}^{20} F_q(i,k) L_t(j,k) + c_1 \delta\left[S_q(i), S_t(j)\right] + shift \quad (1)$$

where $F_q(i, k)$ represents the frequency of the kth amino acid at the ith position of the MSAs obtained by PSI-BLAST[1] search through the NCBI non-redundant sequence database with 3 iterations and E-value cutoff =0.001. The Henikoff weights[2] are exploited to reduce the redundancy of aligned multiple sequences in the frequency profile calculation. $L_t(j, k)$ denotes the log-odd profile of the template for the kth amino acid at the jth position which was pre-calculated for each template by the PSI-BLAST search. $S_q(i)$ stands for the secondary structure at the ith position of the target sequence and $S_t(j)$ represents the secondary structure at the jth position of the template. $\delta[S_q(i), S_t(j)]$ is the Kronecker delta function with value=1 if $S_q(i)=S_t(j)$, or =0 otherwise. The secondary structure for the query is predicted by PSSpred[3], while that for the template is assigned by STRIDE[4], which contains three states: alpha-helix (H), beta-strand (E), and coil (C). The *shift* parameter is introduced to avoid the alignment of unrelated residues in the local regions.

The Needleman-Wunsch dynamic programming algorithm[5] is used to search for the best possible match between the query and template sequences while ending gap penalty is neglected. Parameters of c_1 (=0.65), *shift* (=-0.96), gap opening (*go*=-7.0) and gap extension (*ge*=-0.54) penalties were optimized using a set of 300 non-redundant training proteins by maximizing the average TM-score of the threading models[6].

2. Threading program Env-PPAS

The alignment scoring function of Env-PPAS is similar to PPAS but with a new environment potential added:

$$\text{Score}_{\text{Env-PPAS}}(i,j) = \text{Score}_{\text{PPAS}}(i,j) + c_2 \text{Envior}(j, AA_q(i)) \quad (2)$$

where $\text{Envior}(j, AA_q(i))$ denotes the fitness score of the ith residue of the query, $AA_q(i)$, in the structural environment of the jth residue of the template structure. This potential describes the residue-specific propensity to specific backbone torsion angle, solvent accessibility, secondary structure, and side-chain orientation specific contacts[7, 8], which were derived from the statistics of 5,606 non-redundant PDB structure based on the quasi-chemical approximation[9]. To increase the sensitivity of the algorithm to random alignments, the alignment score is renormalized against the reversed sequences:

$$\frac{\text{Score}_{\text{Env-PPAS}} - \text{RScore}_{\text{Env-PPAS}}}{L} \quad (3)$$

$\text{RScore}_{\text{Env-PPAS}}$ is the raw alignment score of Env-PPAS on an artificial sequence that constitutes the reversal of the query sequence, and L denotes the length of the query sequence. c_2 (*=0.45*) is decided based on our training dataset. The Smith-Waterman local dynamic programming algorithm[10] is used to identify the best alignment between the query and the template.

3. Threading program wPPAS

wPPAS is an extension of PPAS by introducing a new weighted sequence profile that is constructed from a different weighting scheme[11, 12], i.e.

$$\text{Score}_{\text{wPPAS}}(i,j) = \text{Score}_{\text{PPAS}}(i,j) + c_3(\text{Score}_{\text{fPPA}}(i,j) + fshift) \quad (4)$$

where the scoring function for $\text{Score}_{\text{fPPA}}(i,j)$ is defined by

$$\text{Score}_{\text{fPPA}}(i,j) = \sum_{k=1}^{20}\sum_{m=1}^{20} FP_q(i,k) B(k,m) FP_t(j,m) \quad (5)$$

In Eq. (5), $FP_q(i, k)$ represents the frequency profile of the kth amino acid at the ith position of the query sequence; $FP_t(j, m)$ is the frequency profile of the mth amino acid at the jth position of the template sequence; $B(k,m)$ is the BLOSUM62 scoring matrix for aligning the kth amino of the query with the mth amino acid of the template. The constant *fshift* is introduced to balance the new sequence profile term, with the values of c_3=1.0 and *fshift*=-0.4 decided in the training set of proteins.

4. Threading program dPPAS and dPPAS2

dPPAS aligns the query sequence to template structure using an additional depth-based scoring function of

$$\begin{aligned}\text{Score}_{\text{dPPAS}}(i,j) = &\sum_{k=1}^{20} F_q(i,k)\left[L_{str}(j,k) + L_t(j,k)\right] \\ &+ c_4 \delta\left[S_q(i), S_t(j)\right] + dshift\end{aligned} \quad (6)$$

where $F_q(i, k)$, $L_t(j,k)$ and $\delta[S_q(i), S_t(j)]$ are defined in Eq. (1). $L_{str}(j, k)$ stands for structure profile reflecting the environment and depth of the neighboring residue fragments in the template structure[6].

To calculate the structural profile, each template structure is split into small fragments with nine residues, which are used as seed fragments and compared by gapless threading with nine-residue fragments from a set of non-redundant PDB proteins selected by PISCES[13], where the similarity is measured by RMSD and the fragment depth similarity in the structures. These fragments collected from the database are used to calculate the position specific frequency profile $L_{str}(j, k)$ for the templates. The Smith-Waterman local alignment algorithm[10] is used to identify the best match between the target and template sequences. The parameters of dPPAS are determined (c_4=6.5, *dshift*=-0.96, *go*=-7.0 and *ge* =-0.54) by a grid search using our training dataset based on the TM-score of the threading models[6]. dPPAS2 uses the same scoring function and alignment algorithm as that of dPPAS but with a stronger weight on the sequence and depth profile terms.

5. Threading program wdPPAS

wdPPAS is an extension of dPPAS and wPPAS with the score function defined by

$$\text{Score}_{\text{wdPPAS}}(i,j) = \text{Score}_{\text{dPPAS}}(i,j) + c_3(\text{Score}_{\text{fPPA}}(i,j) + fshift) \quad (7)$$

The new weighting profile and parameters are the same as in Eqs. (4) and (5), with the alignment searched by the Needleman-Wunsch global dynamic programming algorithm.

6. Threading program MUSTER

MUSTER (multi-source threader) is a threading algorithm combining sequence profiles with different resources of structure predictions[6]:

$$\begin{aligned}\text{Score}_{\text{MUSTER}}(i,j) = &\sum_{k=1}^{20}[Pc_q(i,k) + Pd_q(i,k)]L_t(j,k) + w_1\delta[s_q(i), s_t(j)] \\ &+ w_2\sum_{k=1}^{20} Ps_t(j,k)L_q(i,k) + w_3[1 - 2|SA_q(i) - SA_t(j)|] \\ &+ w_4[1 - 2|\phi_q(i) - \phi_t(j)|] + w_5[1 - 2|\varphi_q(i) - \varphi_t(j)|] \\ &+ w_6 M[AA_q(i), AA_t(j)] + wshift\end{aligned} \quad (8)$$

Here, $Pc_q(i, k)$ and $Pd_q(i, k)$ represent two profiles collected from close (E-value <0.001) and distant (E-value <1.0) homology sequences, respectively. The second and third terms in Eq. (8) are for secondary structure and structural profile matches, similar to that used for dPPAS in Eq. (6). The fourth, fifth and sixth terms count for the match of solvent accessibility, backbone torsion angles between query and templates. The seventh term is a generic hydrophobic scoring matrix. A systematic lattice-based search, which was designed to maximize the average TM-score[14] of the 300 non-

redundant training proteins, resulted in the optimized parameter sets: w_1=0.65, w_2=1.10, w_3=4.49, w_4=2.01, w_5=0.59, w_6=0.20, *wshift*=1.00, g_o=6.99, and g_e=0.54.

7. Threading program wMUSTER

Compared to MUSTER, two new weighted terms are added in the wMUSTER scoring function:

$$\begin{aligned}\text{Score}_{\text{wMUSTER}}(i,j) = \text{Score}_{\text{MUSTER}}(i,j) + w_7[\text{Score}_{\text{fPPA}}(i,j) + fshift] \\ + w_8\{\delta_2[S_q(i), S_{t7}(j), conf] + sshift\}\end{aligned} \quad (9)$$

Here, PSSpred[3] was retrained with discrete scoring matrixes based on 1,192 non-redundant PDB proteins that were collected by PISCES[13]. For each query residue, wMUSTER considers three secondary structure states (i.e. H, E and C) and ten confidence scores (i.e. [0, 1, …, 9]) as given by the retrained PSSpred. Each residue in the template structure is assigned with one of the seven secondary structure states by STRIDE[4], i.e. H: alpha helix, G: 3-10 helix, I: PI-helix, E: extended conformation, B: isolated bridge, T: turn and C: coil. There are thus in total 21 probability scores for different pairs of PSSpred predictions and STRIDE assignments, each with ten confidence levels. We calculate 3 × 7 substitution matrices, one for each level of confidence value by

$$\delta_2[S_q(i), S_{t7}(j), conf] = log\frac{P[S_q(i), S_{t7}(j), conf]}{P[S_q(i), conf]P[S_{t7}(j)]} \quad (10)$$

where $P[S_q(i), conf]$ is the probability of the predicted secondary structure type i ($i \in$ [H, E, C]) occurred at the confidence score *conf*; $P[S_{t7}(j)]$ is the probability of occurrence of the actual secondary structure type j ($j \in$ [H, E, C, G, B, S, T]). We compared the predicted and actual secondary structure for 1,192 proteins, and calculated the joint probability $P[S_q(i), S_{t7}(j), conf]$ for the predicted secondary structure type i at the confidence score *conf* and the actual secondary structure type j. In this way, we calculated the log-odds scoring function (3×7 substitution matrix) for each confidence value. The constant *sshift* in Eq. (9) is introduced to balance the new term $\delta_2[S_q(i), S_{t7}(j), conf]$. The parameters associated with the new terms in Eq. (9) are: w_7=0.5, *fshift*=-0.5, w_8=0.5 and *sshift*=-0.8.

8. Threading template combination and target classification by LOMETS

The threading templates and alignments are combined by LOMETS[15]. For each template from program P, a Z-score is calculated to evaluate the significance of the alignment compared to other templates:

$$Z_i(P) = \frac{E_i(P) - \langle E \rangle_P}{\sqrt{\langle E^2 \rangle_P - \langle E \rangle_P^2}} \quad (11)$$

where $E_i(P)$ is the score of the alignment by P and $\langle \ldots \rangle_P$ indicates the average over all template alignments by the threading program P. A Z-score cutoff, $Z_0(P)$, is assigned to each program to distinguish if the template is good or bad, based on the large-scale benchmark test[16], i.e. Z_0(PPAS)=7.0, Z_0(Env-PPAS)=8.0, Z_0(wPPAS)=7.0, Z_0(dPPAS)=9.3, Z_0(dPPAS2)=10.5, Z_0(wdPPAS)=9.3, Z_0(MUSTER)=5.8, and Z_0(wMUSTER)=5.8.

If there are more than eight templates with $Z_i(P) > Z_0(P)$, i.e. the average number of good templates is >1 per threading program, the target is defined as an Easy target which usually corresponds to a homologous protein. If there is no template with $Z_i(P) > Z_0(P)$, it is defined as a Hard target which usually corresponds to a non-homologous protein. The others are Medium targets.

An 'init.dat' file is generated which collects the top 20 templates from each of the threading programs. The templates in 'init.dat' are listed in the order of good alignments followed by the bad alignments, enumerated from high to low confident programs, where the confidences of the threading programs are ranked by the average TM-score of the alignments in large-scale benchmark tests. This file will be used by I-TASSER to generate spatial restraints and for constructing initial conformations that the assembly simulations start from.

9. Conformational representation of protein structure in I-TASSER

Each residue in I-TASSER is represented by its Cα atom and side-chain center of mass. Following the query-to-template alignments from LOMETS, the query sequence is split into threading aligned- and unaligned-regions. The conformations in the threading aligned regions are excised from the continuous pieces with length >5 residues of the template structures. The local structures of this portion of sequence are kept frozen during the simulation and move on an off-lattice system; this is designed to enhance the fidelity of the high-resolution structure available from the template alignment (Figure A).

The conformational simulations of the threading-unaligned regions are conducted on a lattice-based system with a grid-scale of 0.81 Å (white color regions in Figure A); this is designed to reduce the conformational search entropy. To further speed up the conformational search, we defined 312 Cα-Cα on-lattice bond-vectors with the length ranging from 3.25 to 4.35 Å[7], where all the allowable 2-bond combinations are pre-calculated by excluding any 2-bond vectors with a bond-angle below 65° or above 165°, which have never been seen in the PDB structures. This filter can considerably reduce the on-lattice searching space. The on-lattice bond-vectors have an average length 3.81 Å that is consistent with the standard Cα-Cα bond length seen in the PDB structures; but the individual bond length may vary between 3.25 and 4.25 Å for the purpose of increasing the conformational flexibility of the on-lattice movements.

The topology of the I-TASSER models are decided by the relative orientation of the individual threading fragments, where the unaligned on-lattice regions are built from scratch which serve as linkage of the threading fragments.

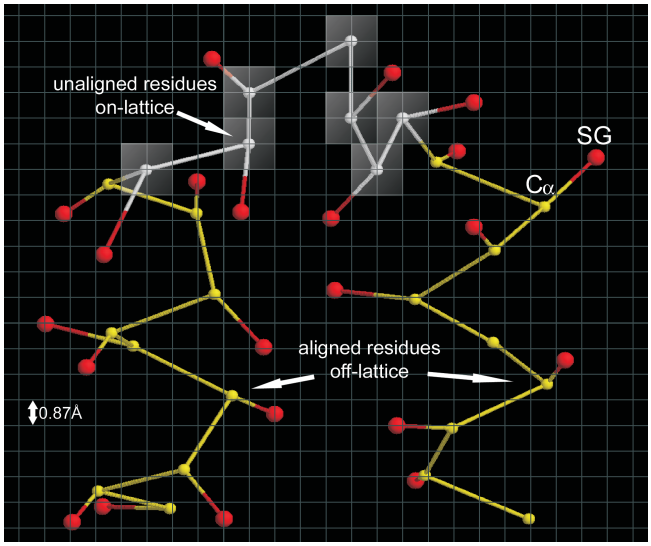

Figure A. The on-and-off lattice system for I-TASSER model representation. Each residue is represented by its Cα atom and side-chain center of mass (SG, red color). The Cα traces in threading unaligned regions are on lattice (white balls); the Cα traces in the aligned regions are excised from template structures and moved continuously (yellow balls).

11. I-TASSER movements and Monte Carlo search engine

The structure reassembly in I-TASSER is conducted by the replica-exchange Monte Carlo (REMC) simulations[17, 18]. There are two types of the conformational movements in the I-TASSER simulations (Figure B).

The first type of on-lattice movements consist of six sets of local moves: (A1) 2-bond vector walk; (A2) 3-bond vector walk; (A3) 4-bond vector walk; (A4) 5-bond vector walk; (A5) 6-bond vector walk;

(A6) N- or C-terminal random walk (Figure B). To speed up the simulations, all the 2-bond and 3-bond conformations for any given distance vectors spanning the moving windows are pre-calculated, so that the A1 and A2 movements can be quickly conducted by enumerating the look-up table of movements without calculating the specific residue coordinates. The A3-A5 movements can also be quickly conducted by the combination of A1 and A2 movements, i.e. by enumerating the neighboring 2- or 3-bond vectors sequentially along the moving window.

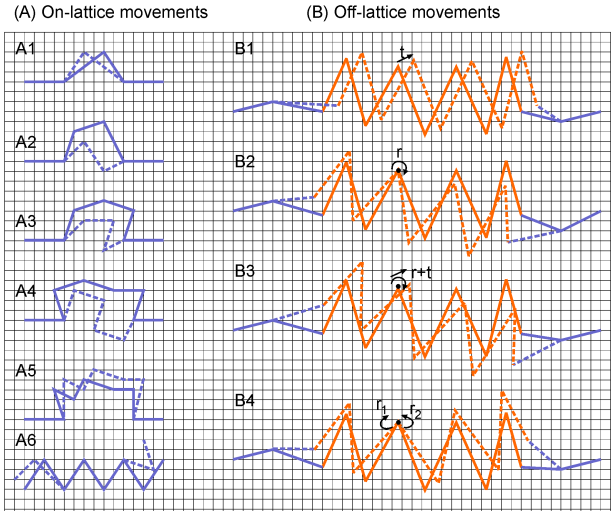

Figure B. Conformational movements in the I-TASSER Monte Carlo simulations. Solid and dashed lines are Cα traces before and after movements. Blue lines represent the threading unaligned regions with moves on the lattice system and orange ones are from the aligned regions with moves off-lattice. **Left Panel**: (A1) two bond vector random walk; (A2) three bond vector walk; (A3) four bond vector walk; (A4) five bond vector walk; (A5) six bond vector translation; (A6) N- and C-terminal random walk. **Right Panel**: (B1) fragment translation; (B2) fragment rotation around a random residue; (B3) fragment rotation and translation; (B4) small deformation of a fragment. After each fragment movement, the neighboring vectors at the two ends are regenerated to keep the chain connection.

The second type of off-lattice movements consist of four sets of global moves of a randomly selected continuous fragment: (B1) a translation of the fragment; (B2) a rotation of the fragment around a randomly selected residue; (B3) a combination of rotation and translation of the fragment; (B4) a small deformation by independently rotating N- and C-terminals of the fragment. After each off-lattice movement of the fragment, the neighboring lattice vectors at the two ends are regenerated to keep the chain connection (Figure B).

Following the standard REMC protocol[17], there are N_{rep} replicas of the simulations which are implemented in parallel, with the temperature of the ith replica being

$$T_i = T_{\min}\left(\frac{T_{\max}}{T_{\min}}\right)^{(i-1)/(N_{rep}-1)} \quad (12)$$

where $T_{\min}$ and $T_{\max}$ are the temperatures of the first and the last replicas, respectively. N_{rep} ranges from 40 to 80, $T_{\min}$ from 1.6 to 1.98 k_B^{-1}, and $T_{\max}$ from 66 to 106 k_B^{-1}, depending on the protein size with larger proteins having more replicas and higher temperatures. These parameter settings can result in an acceptance rate of ~3% for the lowest-temperature replica and ~65% for the highest-temperature replica for different sizes of proteins.

After every 200*L conformational movements, a global swap movement between each pair of neighboring replicas is attempted following the standard Metropolis criterion with a probability of ~exp($\Delta\beta\Delta E$), where the temperature distribution in Eq. (12) with the above temperature and N_{rep} parameter setting resulted in an approximate 40% of acceptance rate for the swap movements between all neighboring replicas.

12. I-TASSER force field construction

The simulations of both unaligned- and aligned-regions in the I-TASSER simulations are governed by the same unified knowledge-based force field, which is built on the reduced model with each residue specified by the Cα atom and the side-chain center of mass. The I-TASSER force field consists of following three major components.

(1) ***Generic statistical potentials***. The generic statistical potentials are derived from the structural regularities observed in the PDB library:

$$E_{stat} = \sum_{i<j}^{L}\left[E_{contact}(A_i,A_j,g_{ij},\theta_{ij}) + E_{pair}(g_{ij},\theta_{ij}) + E_{excl}(A_i,A_j,d_{ij},g_{ij})\right] + \sum_{i=1}^{L}\left[E_{hydro}(A_i,h_i) + E_{env}(A_i,n_{ip},n_{ia},n_{it}) + \sum_{j=2}^{5}\left\{E_{corr}(A_i,A_{i+j},d_{i,i+j},c_{i,i+j}) + E_{\sec}(d_{i,i+5},\vec{r}_{i,i+j})\right\}\right] \quad (13)$$

where L is the query length; d_{ij} and g_{ij} are the Cα and side-chain center distances between the ith and the jth residues, respectively, with A_i and A_j being the amino acid identity of the residues. θ_{ij} denotes the orientation of side-chain vectors of the two residues in contact (i and j) with the values categorized into three groups (parallel, antiparallel and perpendicular).

$E_{contact}(A_i, A_j, g_{ij}, \theta_{ij})$ is the generic orientation-specific contact potential derived from 6,500 non-redundant high-resolution PDB structures. $E_{pair}(g_{ij}, \theta_{ij})$ was derived in the same way as $E_{contact}$ but the counts of contacts are weighted by the sum of the BLOSUM mutation score between the residue pairs of the query and the PDB structures over a window of ±5 neighboring residues. This potential is query-sequence specific but an alignment between the query and the PDB structure is not needed since we counted all the contact pairs in the PDB structures that have the same amino acid identity (A_i, A_j) to the query. $E_{excl}(A_i, A_j, d_{ij}, g_{ij})$ denotes the soft-core excluded volume potential between the Cαs and the side-chain centers based on a $1/d_{ij}$ (or $1/g_{ij}$) dependence.

$E_{hydro}(A_i, h_i)$ is the hydrophobic interactions based on a combined score of the generic Kyte-Doolittle hydrophilic matrix[19] and the sequence-dependent solvent accessibility prediction by neural-network training, with h_i being the depth of the ith residue in the structure.

$E_{env}(A_i, n_{ip}, n_{ia}, n_{it})$ is the orientation-specific contact profile, where n_{ip}, n_{ia} and n_{it} are the numbers of residues in contact with the ith residue, with the side-chain vector being parallel, antiparallel or perpendicular to that of the ith residue, respectively. $E_{corr}(A_i, A_j, d_{i,i+j}, c_{i,i+j})$ is the short-range Cα distance correlation between the ith and the (i+j)th residues where $c_{i,i+j}$ denotes the chirality of the local structure. $E_{\sec}(d_{i,i+5}, \vec{r}_{i,i+j})$ is the secondary structure propensity that is specified by $d_{i,i+5}$, i.e. the local structure will be regulated towards alpha-helix (or beta-strand) when $d_{i,i+5}$ is below 7.5 Å (or above 11 Å), where the secondary structure regularity is represented by the Cα distance and the relative orientation of the neighboring Cα bond-vectors ($\vec{r}_{i,i+j}$) based on the parameters of the standard alpha-helix and beta-strand structures.

Many terms in Eq. (13) have been described in Refs.[7, 20, 21] The parameters used in the terms are downloadable at http://zhanglab.ccmb.med.umich.edu/potential/.

(2) ***Hydrogen-bonding networks***. The hydrogen-bonds (H-bonds) in I-TASSER are specified by the backbone geometry following the PSSpred secondary structure predictions[3]. The occurrence of H-bonds in the simulations is evaluated by the contact order (CO, residue distance along sequence) and the relative orientation and

distance of the donor and receptor residues (Figure C). For instance, a hydrogen bond in an alpha-helix occurs only when CO=3 and the products of the bisector and normal vectors of the donor and receptor residues (*bb* and *cc*) are above certain thresholds and the Cα distance (*r*) is below a cutoff; while CO is requested to be >4 or >20 for hydrogen-bonded residues in anti-parallel or parallel beta-sheets, respectively, given that the same H-bond geometry cutoffs are satisfied.

The hydrogen-bonding energy is calculated by

$$E_{HB} = \begin{cases} \sum -\dfrac{w_{HB}(1-|cc-cc_0|)(1-|bb-bb_0|)}{(1+|bri-br_0|)(1+|brj-br_0|)}, & \text{alpha-helix} \\ \sum -\dfrac{w_{HB}|bb|cc}{(1+bri/2)(1+brj/2)}, & \text{beta-sheet} \end{cases} \quad (14)$$

which is set to penalize the structural deviation of the candidate H-bonds from the standard H-bonding geometry. w_{HB}=1 if both donor and receptor residues are predicted as alpha-helix or beta-stands; otherwise w_{HB}=0.5. The cutoff parameters for standard hydrogen bonds (cc_0, bb_0, br_0) were calculated from an average of 500 high-resolution PDB structures with the secondary structure elements assigned by STRIDE[4].

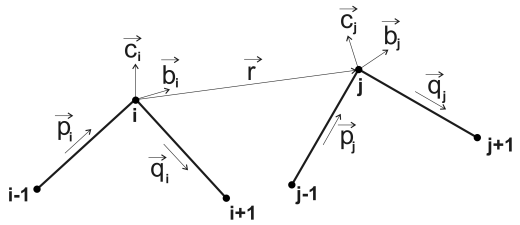

Figure C. Definition of H-bond geometry between residues *i* and *j*. $\vec{p}$ and $\vec{q}$ are Cα-Cα vectors, $\vec{c}=(\vec{p}-\vec{q})/|\vec{p}-\vec{q}|$ is the unit bisector vector and $\vec{b}=\vec{c}\times\vec{q}/|\vec{c}\times\vec{q}|$ is the unit normal vector. $\vec{r}$ is a vector pointing from donor Cα to receptor Cα atom. Accordingly, we define $cc=\vec{c}_i\cdot\vec{c}_j$, $bb=\vec{b}_i\cdot\vec{b}_j$, $bb=\vec{b}_i\cdot\vec{b}_j$, $pp=\vec{p}_i\cdot\vec{p}_j$, $qq=\vec{q}_i\cdot\vec{q}_j$, $bri=|l\vec{b}_i-\vec{r}|$, $brj=|l\vec{b}_j-\vec{r}|$, where l=5 Å for alpha-helix and 4.6 Å for beta-sheets.

(3) *Threading template-based restraints*. The threading based spatial restraints contain Cα distance maps and side-chain contacts which are collected from the 'init.dat' file by taking the consensus of the top N_{temp} threading templates from LOMETS[15], where N_{temp}=20 for the Easy targets, 30 for the Medium targets, and 50 for the Hard targets following LOMETS target definition. In the second round of iterative simulations, the external restraints also include, in addition to the threading restraints, the restraints from the I-TASSER cluster centroids and the restraints from the PDB templates detected by the structure alignment program TM-align[22] that uses the cluster structures as the probe[23].

Four types of restraints are implemented in I-TASSER by

$$E_{rest} = \begin{cases} \sum_{i<j} |d_{ij}-d_{ij}^p|, \text{ short-range C}\alpha \text{ distance with } |i-j| \le 6 \\ \sum_{i>j} -1/|d_{ij}-d_{ij}^p|, \text{ long-range C}\alpha \text{ distance with } |i-j|>6 \\ \sum_{i>j} -w(conf_{ij})\theta(6.5-d_{ij}), \text{ C}\alpha \text{ contact restraints} \\ \sum_{i>j} w(conf_{ij})\theta(g_{ij}-g_0^{A_iA_j}), \text{ side-chain contact restraints} \end{cases} \quad (15)$$

where d_{ij} and g_{ij} are the Cα and side-chain center of mass distances between the *i*th and the *j*th residues; d_{ij}^p is the predicted distance between the two Cα atoms; $g_0^{A_iA_j}$ is the amino acid specific distance cutoff for side-chain contact; $\theta(x)$ is a step function which equals to 1 when $x\ge0$ or 0 when $x<0$; $w(conf_{ij})=1+|conf_{ij}-conf_{cut}|^4$ when the confidence score of the restraint ($conf_{ij}$) is higher than the threshold ($conf_{cut}$), and otherwise $w(conf_{ij})=1-|conf_{ij}-conf_{cut}|^2$.

The I-TASSER energy terms are combined by a linear regression, where the weighting parameters of different terms are optimized on a set of 100×60,000 decoys by maximizing the correlation between the total energy and the TM-score of decoys to the native states[7].

13. Structural model selection and atomic-level refinement

The most frequently occurring conformations in the I-TASSER structure assembly simulations are selected by the SPICKER clustering program[24]. These conformations correspond to the models of the lowest free-energy state in the Monte Carlo simulations, since the number of decoys at each conformational cluster n_c is proportional to the partition function Z_c, i.e. $n_c \sim Z_c = \int e^{-\beta E}\,dE$. Thus, the logarithm of normalized cluster size is related to the free energy of the simulation, i.e. $F=-k_BT\log Z \sim \log(n_c/n_{tot})$, where n_{tot} is the total number of decoys submitted for clustering.

The centroid model in each SPICKER cluster is generated by averaging the coordinates of all the clustered conformations. Since the centroid models often contain steric clashes, a second round of assembly simulation is conducted by I-TASSER to remove the local clashes and to further refine the global topology. In the second round of reassembly simulations, spatial restraints are added from structure templates detected by searching the PDB library with TM-align for structures that are similar to the cluster centroids from the first round of simulation.

The final atomic model is constructed by ModRefiner[25] starting from the low-energy conformations selected from the second round simulation trajectories. In ModRefiner, the backbone structure is first built from the Cα-traces. The side-chain atoms are then constructed from a rotamer library with the full-atomic conformation refined by energy minimizations based on a composite physics- and knowledge-based force field[25].

14. Global quality estimation of I-TASSER structure predictions

Not all the structure predictions are accurate. An estimation of the modeling accuracy is thus essential to decide how the users should utilize the models in their own research. The accuracy of the I-TASSER structure models is estimated through the calculation of the confidence score (or C-score) of the structure assembly simulations:

$$\text{C-score} = \ln\left(\frac{M}{M_{tot}}\cdot\frac{1}{\langle RMSD\rangle}\cdot\frac{1}{8}\sum_{i=1}^{8}\frac{Z(i)}{Z_0(i)}\right) \quad (16)$$

where M is the multiplicity of structures in the SPICKER cluster; M_{tot} is the total number of the I-TASSER structure decoys used in the clustering; $\langle RMSD\rangle$ is the average RMSD of the decoys to the cluster centroid. These terms correspond to the degree of convergence of the structure assembly simulations. $Z(i)$ and $Z_0(i)$ are the highest Z-score of the templates by the *i*th LOMETS threading program and the corresponding Z-score cutoff for distinguishing between good and bad templates, as defined in Section 8. The normalized Z-scores measure the significance of threading alignments and correlate with the quality of the LOMETS templates.

The large-scale benchmark tests[26] show that there is a strong correlation between C-score and the accuracy of the I-TASSER models. The correlation coefficient between C-score and the actual TM-score is 0.91 and that between C-score and RMSD is 0.75. These data allow a quantitative estimation of the TM-score and RMSD of the first predicted model related to the native state:

$$\begin{cases} \text{TM-score} = 0.0006C^2 + 0.13C + 0.71 \\ \text{RMSD} = 0.09(C-\ln L)^2 - 1.14(C-\ln L)] - 3.17 \end{cases} \quad (17)$$

where C is the C-score and L is the length of the target sequence. In a large-scale benchmark test on 500 non-redundant proteins, the average errors of quality estimation using Eq. (17) are 0.008 and 2.0 Å for TM-score and RMSD, respectively.

15. Residue-level local quality estimation of I-TASSER structure predictions

An algorithm (ResQ) was developed for estimating the residue-level local quality of the I-TASSER predictions on the basis of the variations of modeling simulations and the uncertainty of homologous alignments[27]. To train the method, we first conducted I-TASSER based structure predictions for 1,270 non-redundant single-domain proteins from the PDB, which are randomly split into two sets of training and test proteins. Support vector regressions (SVRs) were used to train residue-specific distance error of the I-TASSER models, in comparison with the native structure, on the following five features.

(1) Structural variation of assembly simulations. The structural variation of *j*th residue in the REMC simulations is defined by the average and standard deviations:

$$\begin{cases} \mu_j = \frac{1}{N}\sum_{i=1}^{N} d_{ij} \\ \nu_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_{ij}-\mu_j\right)^2} \end{cases} \quad (18)$$

where N is the number of decoys in the SPICKER cluster; d_{ij} is the distance for the *j*th residue between the *i*th decoy structure and the centroid structure model after the TM-score superposition[14]. In general, residues with a higher variation have the larger errors relative to the native, and vise versa.

(2) Consistency between final model and sequence-based feature predictions. The secondary structure (SS) and solvent accessibility (SA) of the target sequence are predicted by PSSpred[3] and SOLVE (Zhang et al, unpublished) programs, which are compared with the actual SS and SA of the 3D structural models that are assigned by the STRIDE program[4]. The residues with inconsistent SS and SA between model and prediction have usually larger error.

(3) Threading alignment coverage. The alignment coverage of a residue is defined as the number of threading templates that have the query residue aligned divided by the total number of templates by LOMETS. The residues with a higher threading coverage indicate more constraints on them during simulations and presumably have a higher modeling accuracy.

(4) Structural variation of templates from LOMETS threading. Considering N_j templates which are aligned to the *j*th residue on the query sequence by LOMETS[15], the structural variation of the LOMETS threading templates is defined as:

$$\lambda_j = \frac{1}{N_j}\sum_{n=1}^{N_j} d_n(j) \quad (19)$$

where $d_n(j)$ is the distance between the *j*th residue on the model and the residue on the *n*th template that is aligned to the *j*th residue in query. The distance is calculated after superposing the template structures on the query model by TM-score rotation matrix[14]. In case that N_j is zero, the value of λ_j is set to be 10 Å.

(5) Structural variation of templates from TM-align structure alignments. The query structure model is threaded against a non-redundant PDB library by TM-align[22] to find templates that share similar topology. The structural variation of the TM-align templates is defined in the same way as Eq. (19), but with N_j and $d_n(j)$ defined by the TM-align templates.

Starting from the I-TASSER model and the simulation decoys, ResQ generates distance error estimation for each residue. A benchmark test on the 506 non-redundant proteins that have a I-TASSER C-score >-1.5 showed that the residue-level accuracy can be estimated with an average error <1.5 Å in comparison with the X-ray crystallography data[27].

16. B-factor estimation

B-factor is associated the inherent thermal mobility of local atoms and residues, which is essential for proteins to fold and function in the physiological environment. The B-factor prediction was trained by SVRs on the template-based assignments and the PSI-BLAST profiles.

(1) Template-based assignments. I-TASSER assigns B-factor of each query residue on the basis of the experimental B-factor values of the top homologous and analogous templates that are extracted from the original PDB entries by LOMETS and TM-align, i.e.

$$b_q(j) = \frac{1}{n_j}\sum_{i=1}^{n_j} b_t(i,j) \quad (20)$$

where n_j is the number of the templates that have a residue aligned on the query residue *j*, and $b_t(i, j)$ is the normalized B-factor value of the residue from the *i*th template that is aligned on *j* by LOMETS and TM-align.

(2) Sequence profiles. The target sequence is searched by PSI-BLAST[1] through the NCBI non-redundant sequence database to retrieve homologous sequences, which are represented in the form of a position-specific scoring matrix (PSSM). For each residue, a sliding window with the size of 9 residues is used to extract profile features from the PSSM after converting its elements *x* in the range of (0, 1) by 1/[1+exp(-x)]. The secondary structure and solvent accessibility, which are both derived from the PSI-BLAST sequence profiles, are also used as features for B-factor prediction. The hypothesis of using the sequence profile is that the more conserved residues are often structurally more stable and therefore have a lower B-factor, and vice versa.

The benchmark test on the 635 non-redundant proteins showed the estimated B-factor by ResQ has a Pearson's correlation coefficient 0.60 with the X-ray crystallography data[27].

17. BioLiP: a composite database for structure-based protein function annotations

The PDB library[28] is the central source for protein structure and function analysis studies. But many proteins in the PDB contain redundant entries, mis-ordered residues and mis-annotated functions. Moreover, many protein structures were solved using artificial molecules as additives to facilitate structure determination. These errors and artificial additives prevent the library from being a reliable resource for precise structure-based function analyses, which requires the development of cleaned protein libraries with the biological functions carefully validated. We proposed a hierarchical procedure consisting of multiple computer filtering and manual literature validation for assessing the biological relevance of co-structured ligands in the PDB. A comprehensive function database, BioLiP[29], was constructed from known structure/function databases and literature in PubMed, with each entry containing annotations on ligand-binding sites, binding affinity, and catalytic sites. BioLiP is updated weekly and freely available to the community at http://zhanglab.ccmb.med.umich.edu/BioLiP. The current BioLiP release (Oct 31, 2014) contains 298,556 ligand-binding entries from 64,287 unique proteins, with 37,223 entries for DNA/RNA-protein, 13,498 for peptide-protein, 84,248 for metal ion-protein, and 163,587 for small molecule-protein interactions. There are in total 23,492 entries with experimental binding affinity data collected from public databases[30-32] and manual literature survey[29].

Most recently, BioLiP was extended to integrate sequence-based function annotations from UniProtKB[33], Enzyme Commission (EC)[34] and Gene Ontology (GO)[35], where the EC and GO information from the sequence databases was mapped to the BioLiP structure entries through stringent sequence alignments assisted by manual validation[36]. It involves 75,462 protein chains with 513 unique first 3-digit and 2,413 unique 4-digit enzyme commission (EC) numbers, and 37,178 chains with known catalytic residues derived from Catalytic Site Atlas (CSA)[37]. It also contains 119,004 chains/domains associated with 8,315 unique gene ontology (GO) terms. These data provide a fundamental recourse for structure-based function annotation in the I-TASSER Suite.

18. COFACTOR for ligand-binding site (LBS) predictions

COFACTOR predicts LBS by mapping the LBSs obtained from known structures of ligand-protein complexes in the BioLiP library[29] that have similar structures to the query structure models. The structural similarity is detected by combining global structure alignment and local geometry refinement[38, 39].

For the global structure alignment, TM-align[22] is used to identify templates that have a similar fold to the query, i.e., with a TM-score >0.5[14, 40]. For the local geometry alignment, an iterative Needleman-Wunsch dynamic programming is used to align the template binding site regions with the query fragments, which are extracted from the query structure based on MSA conservations. The local structural and sequence similarity (L_{sim}) between query and template proteins is evaluated by

$$L_{sim} = \frac{1}{N_t}\sum_{i=1}^{i=N_{ali}} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2} + \frac{1}{N_t}\sum_{i=1}^{i=N_{ali}} M_{ii} \quad (21)$$

where N_t represents the total number of residues within a sphere centered at the known binding sites, N_{ali} is the number of query-template aligned residue pairs within the sphere, d_i is the Cα distance between ith aligned residue pair, and d_0 is the distance cutoff chosen to be 3.0 Å, and M_{ii} is the normalized BLOSUM62 substitution scores between the ith pair of residues. For each binding pocket on the template, this procedure is implemented for all the conserved query motifs; and the one with the highest L_{sim} is recorded. The template ligands are superimposed onto the query structure following the rotation and translation matrix from the local structure alignment. Finally, the predicted ligand conformations from all templates are clustered based on the spatial proximity with a distance cutoff 8 Å. If a binding pocket binds with multiple ligands (e.g. an ATP binding pocket may also bind MG, PO_4^{3-} and ADP), ligands within the same pocket are clustered further based on their chemical similarity (Tanimoto coefficient cutoff =0.7) using the average linkage clustering procedure to rank the predicted binding sites.

From each cluster, the protein-ligand complex with the highest ligand-binding confidence score (CS_c) is eventually selected as the functional site predictions for the query protein, i.e.

$$CS_c = \frac{2}{1+e^{-\left(\frac{N}{N_{tot}}\times\left(0.25\,L_{sim}+TM-score+2.5ID_{Str}+\frac{2}{1+\langle D\rangle}\right)\right)}} - 1 \quad (22)$$

where N is the number of template ligands in the cluster and N_{tot} is the total number of predicted ligands using the templates. ID_{str} is the sequence identity between the query and the template in the structurally aligned region. $\langle D\rangle$ is the average distance of the predicted ligand to all other predicted ligands in the same cluster.

19. TM-SITE for ligand-binding site predictions

TM-SITE is another structure-based function annotation algorithm extended from COFACTOR for mapping the LBS of known proteins by structural similarities[41]. As opposed to COFACTOR that combines global and local structure alignments, TM-SITE detects function homologies by considering the structure similarity of a subsequence from the first binding residue to the last binding residue (called SSFL) on the query and template proteins.

To identify the functional homology, TM-SITE uses TM-align[22] to align the query structure with the template SSFLs which are precalculated in the BioLiP[29]. The match between each pair of query and template SSFL structures is evaluated by a composite scoring function which counts for the global and local, and structural and sequence similarities:

$$q_{str} = \frac{2}{1+e^{-\left[L_c\left(0.4L_g+0.3L_s+0.2JSD\right)+TM\right]^2}} - 1 \quad (23)$$

where L_c is the fraction of template binding residues that are aligned to the query structure by TM-align. $L_g = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+(d_i/d_0)^2}$ accounts for the local structure similarity between the binding pockets of query and template proteins, where n is the number of the aligned residue pairs associated with the binding pockets of the template and d_i is the distance of the ith residue pair. $L_s = \frac{1}{n}\sum_{i=1}^{n} B(R_i^q, R_i^t)$ measures the evolutionary relationship between the aligned binding residue pairs where $B(R_i^q, R_i^t)$ is the normalized BLOSUM62 score for the ith aligned residue pair. JSD is an evolutionary conservation index defined as the average Jensen–Shannon divergence score over the predicted binding residues, which is calculated from multiple sequence alignments. $TM = 2TM_t * TM_q/(TM_t + TM_q)$ is the harmonic average of the two TM-scores returned by TM-align when aligning the query and template SSFLs, where TM_t and TM_q are TM-scores normalized by query and template lengths, respectively.

To select the LBS residues from multiple SSFL templates, all ligands bound with the proteins in the putative template pool are projected to the query structure based on the corresponding alignments from TM-align. These ligands are clustered based on the spatial distance between their geometric centers, where an average linkage clustering algorithm is conducted with a distance cutoff 4 Å. For each cluster, a set of consensus binding residues, which usually correspond to one binding pocket, are deduced from all the ligands in the cluster based on the maximum voting. The residues receiving >25% votes are considered as the final predicted LBS residues in the binding pocket. A confidence score of the predicted binding residues, associated with specific ligand-binding clusters, is defined by

$$CS_t = \frac{2}{1+e^{-\left[\frac{m}{M}q_{str}^{max}+0.2\ln(1+\sqrt[3]{m})+0.2JSD_{Ta}\right]}} - 1 \quad (24)$$

where m is the number of template ligands in the cluster and M is the total number of templates selected. q_{str}^{max} is the maximum of q_{str} score from the templates in the cluster as calculated by Eq. (23). JSD_{Ta} is the average JSD score for the predicted LBS residues from the ligand cluster. In the final TM-SITE predictions, the binding pockets are selected and ranked based on the CS_t score, with the binding residues sorted by the number of votes in each binding pocket.

20. S-SITE for ligand-binding site predictions

S-SITE is the third template-based method in the I-TASSER package, which detects protein function templates and the ligand binding sites using binding-site specific, sequence profile-profile comparisons. To obtain the profile of the query protein, PSI-BLAST is used to thread the query sequence through the NCBI NR sequence database to construct multiple sequence alignments, similar to the procedure used in PSSpred and LOMETS. A position-specific frequency matrix (PSFM) is then computed from the multiple sequence alignments. The template profiles are represented by the position-specific scoring matrices (PSSM) and pre-constructed by the PSI-BLAST searches for all proteins in the BioLiP library. To detect homologous templates from BioLiP, the query profile PSFM is compared with the template profile PSSMs in the library using the Needleman-Wunsch dynamic programming algorithm[5], with the score to align the ith residue in the query to the jth residue in template defined as

$$S_{i,j} = \sum_{k=1}^{20} F_q(i,k)L_t(j,k) + \delta\left[S_q(i), S_t(j)\right] + 2b_j^t B\left(R_i^q, R_j^t\right) \quad (25)$$

The first two terms in Eq. (25) describe the profile-profile alignment and secondary structure matches, similar to Eq. (1). The third term describes the evolutionary conservation of binding pocket residues, where $b_j^t = 1$ if the jth residue is at the binding site in the template, or $b_j^t = 1$ otherwise; $B(R_i^q, R_j^t)$ is the normalized BLOSUM62

similarity score for residues R_i^q in query and R_j^t in template with value in [0,1]. The quality of a template match in S-SITE is estimated by

$$q_{seq} = \frac{2}{1+e^{-(0.5A_S+0.5L_c L_s+0.2JSD)}} - 1 \quad (26)$$

where $A_S = \frac{1}{L}\sum_{i=1}^{L_{ali}} S_{i,i}$ is the profile-alignment score normalized by the query sequence length L. L_c, L_s and JSD are similar to that defined in Eq. (23) but with the alignments generated from the ligand-binding specific profile-profile comparisons. All proteins in BioLiP with a q_{seq} score above 0.5 are selected as putative templates. If the number of putative templates is below 10, the top 10 templates with the highest q_{seq} score will be returned for the next step of LBS selection analysis. The residues on the query, which are aligned with the binding residues on the templates following the sequence profile-profile alignments, are assigned as putative binding residues in the S-SITE prediction. Since the binding sites of different templates will match with different query residues, a consensus-voting scheme is applied to select the most consensus binding residues. The residues receiving >25% votes are considered as the final binding residues by S-SITE.

A confidence score CS_s is defined for the binding residues in S-SITE:

$$CS_s = \frac{2}{1+e^{-\left[q_{seq}^{\max}+0.1\ln(1+\sqrt[3]{N})+0.2JSD_{Sa}\right]}} - 1 \quad (27)$$

where $q_{seq}^{\max}$ is the maximum value of q_{seq} among all the putative templates, N is the number of the selected templates, JSD_{Sa} is the average Jensen–Shannon divergence score of all the predicted LBS residues. Here, the confidence score CS_s is in a similar format as that of TM-SITE (CS_t) but no clustering is conducted in S-SITE since the templates detected from the sequence profile comparison are converged in most cases.

21. COACH for consensus ligand-binding site prediction

COACH is a consensus approach to LBS prediction that combines the prediction results of algorithms from COFACTOR, TM-SITE and S-SITE using a linear Support Vector Machine (SVM) (Figure D). To generate a LBS prediction, the query sequence along with the I-TASSER structure model are provided as input and fed into the individual programs. The probability of a residue to be a LBS residue is calculated from individual methods, which are used as the input feature vectors of the SVM for the residue. The probability for COFACTOR is taken from the default confidence score (CS_c in Eq. 22). For TM-SITE and S-SITE, the probabilities are computed as the confidence score of the ligand cluster or templates (i.e. CS_t and CS_s in Eqs. 24 and 27) multiplied by the ratio of votes on the residue. Finally, all feature vectors are fed into SVM to make the consensus prediction, with classifiers trained on 400 non-redundant training proteins collected by PISCES[13].

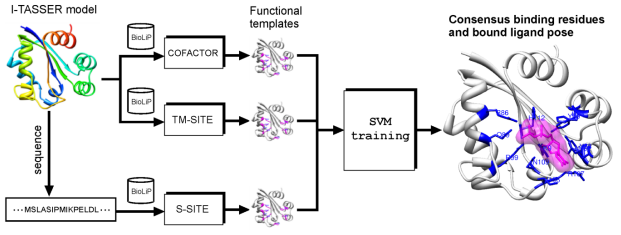

Figure D. Flowchart of COACH for ligand binding site prediction. The final results are combined from the complementary predictions by COFACTOR, TM-SITE and S-SITE using support vector machine (SVM).

The linear kernel in SVM-light was used with the optimal value of the cost parameter C selected based on an exhaustive grid search. To avoid over-training, a 10-fold cross-validation procedure was applied in COACH, i.e. the training set was randomly divided into 10 subsets of equal size, where 9 subsets were used to train the SVM and the remaining subset was used as validation to calculate the average Matthews's correlation coefficient (MCC) of the LBS predictions. For each parameter C in the grid space, such random sample division was repeated 10 times and an overall MCC was calculated as the mean of the 10 MCCs. The parameter C with the highest overall MCC was finally selected for SVM training in COACH.

22. Structure-based enzyme function predictions

Enzymes are proteins that catalyze chemical reactions in physiological processes. Based on the reactions they catalyze, enzymes are categorized into hierarchical families using a numerical classification scheme known as Enzyme Commission (EC) number. To predict the EC number from amino acid sequence, our method involves structural matching of the predicted protein structures to global and local 3D enzyme templates. The confidence score of the EC predictions, the EC-score, is defined as

$$\text{EC-score} = \begin{cases} Cs\left[TM + \left(\frac{1}{1+RMSD_{ali}}Cov\right)\right] + 2ID_{ali}Cov + \frac{L_{sim}}{2}, & \text{if } L_{\text{sim}} > 1.1 \\ Cs\left[TM + \left(\frac{1}{1+RMSD_{ali}}Cov\right)\right] + 3ID_{ali}Cov, & \text{if } L_{\text{sim}} < 1.1 \end{cases} \quad (28)$$

where Cs is the C-score of I-TASSER prediction as defined in Eq. (16). TM, $RMSD_{ali}$, Cov, and ID_{ali} are the TM-score, RMSD, alignment coverage, and sequence identity returned by TM-align[22] when comparing the target I-TASSER model with the template structures in BioLiP[29]. L_{sim} is the local active site match score between the query and template binding pockets as calculated in Eq. (21).

To generate initial binding pockets for query, the known catalytic/active residues of the template are scanned through the query sequence. The residues, whose amino acid type is similar to that of the catalytic/active residues on templates, are marked as potential active site locations in the query. The binding pocket sphere of the template is defined as the sphere centered at the geometric center of the template's active site and containing <30 residues (or with a radius <20 Å). The optimized alignment of the binding pockets is conducted by a heuristic iteration procedure similar to that used in COFACTOR[39] and TM-align[22].

23. Structure-based gene ontology predictions

Gene ontology (GO)[42] is a widely used vocabulary for describing three different taxonomies or "aspects" of gene functions: molecular function, biological process, and cellular component. Each GO aspect is represented as a structured, directed acyclic graph, where the nodes in the graph represent a GO term and describe a component of gene product function, while the edges between the nodes are equivalent to the relationships ('*is-a*' or '*part-of*') between the GO terms. The GO terms are held in a form of functional hierarchy, where more general functions are present on the top while more specific functions are further down the graph.

To identify the GO terms and the function components of a query protein, we use a structure-based approach similar to that used for the LBS and EC predictions. Since multiple templates can be identified by TM-align structure comparison, the confidence of the GO predictions, the GO-score, is defined as a normalized sum of the sequence and structural similarity score (SaS) of all templates that have the same GO term:

$$\begin{cases} \text{GO-score}(\lambda) = \frac{1}{N}\sum_{\lambda=1}^{N_\lambda} \text{SaS}(\lambda) \\ \text{SaS}(\lambda) = Cs_\lambda\left(TM_\lambda + \frac{1}{1+RMSD_{ali,\lambda}}Cov_\lambda\right) + 3ID_{ali,\lambda}Cov_\lambda \end{cases} \quad (29)$$

where λ represents a given gene ontology term. Cs_λ, TM_λ, $RMSD_\lambda$, Cov_λ, $ID_{ali,\lambda}$ are similar to the terms defined in Eq. (28) but with a specific GO label λ. N_λ is the number of templates which are associated with the GO term λ, and N is the total number of templates

selected for generating the consensus GO predictions. When multiple homologous templates are available, we only consider the templates with a SaS score >1. For those query proteins with less than 10 templates of SaS score >1, the top 10 templates are selected for generating the consensus prediction regardless of the SaS score. Once a GO term is identified with the GO score higher than the GO score cutoff, all its ancestor terms in the directed acyclic graph (DAG) are eliminated to avoid redundancy. The recent large-scale benchmark tests on 700 non-redundant proteins[36, 38] showed that the accuracy of the EC and GO predictions based on the I-TASSER models is significantly higher than that based on sequence-based (PSI-BLAST[1]) and threading-based (MUSTER[6] and HHpred[43]) approaches.

24. Installation and implementation of the I-TASSER Suite

(1) I-TASSER *installation*. The I-TASSER Suite was developed on the Linux operating system, with the core programs written in FORTRAN, C/C++, and Java languages. The package can be obtained at http://zhanglab.ccmb.med.umich.edu/I-TASSER/download/. Weekly updated structure and function libraries are available at http://zhanglab.ccmb.med.umich.edu/library, where a Perl script 'download_lib.pl' is provided in the package for automated library download. The package includes a program ('update_IT_lib.tar.bz2') which users can use to create template files on their own from any protein structures. Figure E presents a screenshot that illustrates the structure of the I-TASSER package after installation.

(2) *I-TASSER implementation*. To interpret the implementation of I-TASSER Suite, let us assume that the package, the template library, the java executable program, and the input files are located at $pkgdir, $libdir, $java/bin/java, and $datadir, respectively. These directories can be installed at any folders in user's computer.

The main script to run I-TASSER is located at '$pkgdir/I-TASSERmod/runI-TASSER.pl'. Running '$pkgdir/I-TASSERmod/runI-TASSER.pl' without argument can print out a brief help information of the various options. One example command for running I-TASSER is

$pkgdir/runI-TASSER.pl -pkgdir $pkgdir -libdir $libdir –runstyle serial -LBS true -EC true -GO true -seqname example -datadir $datadir –outdir $outdir -java_home $java -light true -hours 6

In this example, I-TASSER is running in the *serial* mode as assigned by option '*–runstyle serial*' with the query protein named '*example*'. The modeling will be conducted for both 3D structure prediction and biological function annotations including ligand binding site (LBS), Enzyme Commission number (EC), and Gene Ontology terms (GO). The maximum running time for each I-TASSER simulation job is set as 6 hours as assigned by the options '*-light true*' and '*-hours 6*'. Users can also run I-TASSER in the *parallel* mode which is assigned by option '*–runstyle parallel*', where the threading and the I-TASSER simulation jobs will be submitted to a cluster of computer nodes by the job scheduler using the '*qsub*' command (assuming the PBS job scheduler system is installed). The *parallel* mode can significantly accelerate the I-TASSER modeling.

(3) *Input data preparation*. The only input information required to run the I-TASSER Suite is the amino acid sequence of the target protein, which must be saved in a file named '*seq.fasta*' in the folder $datadir. Figure F shows an example of the input. In addition to the sequence information, users can assign additional restraints and templates to guide I-TASSER modeling, using one of the four options: *-restraint1, -restraint2, -restraint3, -restraint4*. Users must first create all the restraint files in $datadir and specify the file names according to the options (see http://zhanglab.ccmb.med.umich.edu/I-TASSER/download/README.txt).

(4) *Interpreting structure prediction results*. The output results of the I-TASSER structure modeling are located in the folder $outdir that was specified by the option '*–outdir*'. If no output directory is specified, the default folder $outdir=$datadir. The structural modeling outputs include threading templates, simulation trajectory files, up to five structure models and the confidence score estimations.

Figure G shows a screenshot of the output results from the I-TASSER based structural modeling, where Supplementary Table 1 summarizes the annotations of the output files of I-TASSER structure predictions.

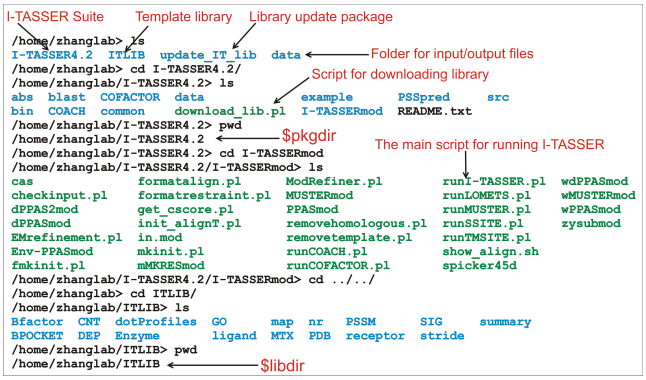

Figure E. A screenshot illustrating the file structure of the I-TASSER Suite installation.

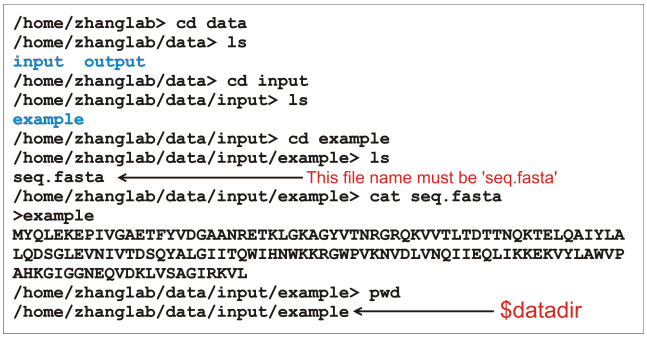

Figure F. A screenshot illustrating input files for running the I-TASSER Suite.

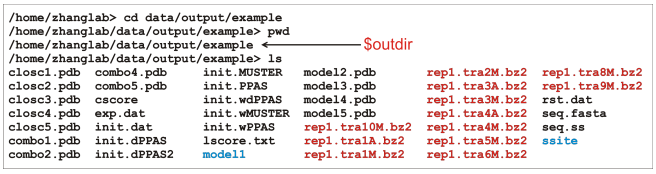

Figure G. A screenshot illustrating output results by I-TASSER structure prediction pipeline.

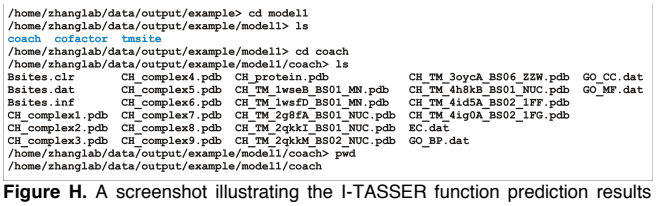

Figure H. A screenshot illustrating the I-TASSER function prediction results from COACH.

(5) *Interpreting function prediction results*. The results of the structure-based function predictions by COACH are located at $outdir/model1/coach, where the file items are explained in Figure H and Supplementary Table 2. In addition to the consensus COACH results, the function annotations by individual predictors of TM-SITE, S-SITE and COFACTOR are listed in $outdir/model1/tmsite, $outdir/ssite and $outdir/model1/cofactor/, respectively.

Supplementary Table 1. Summary of structure modeling results by the I-TASSER Suite.

File Name	Interpretation	Note
model1-5.pdb	All-atomic 3D structure models predicted by I-TASSER	There may be < 5 models, if the number of SPICKER clusters is < 5. The models usually have a good quality in such case.
combo1-5.pdb	Cluster centroid models	The cluster centroids often have clashes
closc1-5.pdb	Decoy structure closest to the cluster centroid	Cα trace models
cscore	Confidence score of global fold	Quality estimation of global fold of the models.
lscore.txt	Confidence score of local structures	Estimation of residue-level quality of the models and B-factor of the protein.
seq.ss	Predicted secondary structure	Secondary structure predicted by *PSSpred*
exp.dat	Predicted solvent accessibility	Predicted by the program *solve* at different cutoffs.
init.dat	Threading templates used by I-TASSER	The head of this file indicates the type of the target: easy, medium, or hard. Templates are ranked according to Z-score and the consensus of templates.
init.*	Templates generated by 8 individual LOMETS threading programs	The templates are generated by default using the 8 in-house threading programs. Users are encouraged to add external threading programs to increase complementarity. To add a new threading program, users need to prepare a new mod script that can be adapted from MUSTERmod and add the program name to the array @TT in 'runI-TASSER.pl'.
rst.dat	Summary of SPICKER clustering	Clustering is done by the program SPICKER
rep*.bz2	Structural decoys from the low-temperature replicas in I-TASSER simulations.	Users can automatically remove these files by specifying the option *–traj false.*

Supplementary Table 2. Summary of the structure-based function prediction by COACH.

File Name	Interpretation	Note
Bsites.dat	Ligand-binding site predictions	Columns in the file are: C-score, relative cluster size, product of top templates z-score, binding residues
Bsites.clr	Clustering of all predicted binding sites	Each cluster leads to one line in 'Bsites.dat'. Each line in one cluster corresponds to one template-based prediction from COFACTOR (COF), TM-SITE (TMS) or S-SITE (SST).
Bsites.inf	Summary of clustering results in 'Bsites.clr'	For each site (cluster), there are three lines: **Line1**: site #, c-score of coach prediction, cluster size **Line2**: algorithm, PDB ID, ligand ID, center of binding site, c-score of the algorithm's prediction, binding residues from single template **Line 3**: Statistics of ligands in the cluster
CH_complex*.pdb	Predicted protein-ligand complex structures	For each cluster, all putative bound ligands are put in the same file and separated by "TER" and/or "END".
CH_*_*_*.pdb	Representative single complex structure for each site/cluster	The file name for a cluster can be obtained by combing 'CH' with the first three columns of **line 2** in 'Bsites.inf'.
EC.dat	Predicted EC number and active residues	The columns are: PDB_ID, TM-score, RMSD, Sequence identity, Coverage, Confidence score, EC number, and Active site residues
GO_MF.dat	GO terms in 'molecular function'	The columns are: GO terms, Confidence score, Name of GO terms.
GO_BP.dat	GO terms in 'biological process'	
GO_CC.dat	GO terms in 'cellular component'	

Supplementary Table 3. Summary of I-TASSER modeling on the six examples in Figure 1.

Target[a]	L[b]	Structure prediction results		Structure-based function prediction		
		C-score/TM_{est}/$RMSD_{est}$(Å)[c]	TM/RMSD(Å)[d]	L-RMSD(Å)[e]	Top GO term[f]	EC-score[g]
R0006	169	-0.75/0.61±0.14/6.6±3.9	0.62/4.6	-	0044238	-
R0007	161	-0.67/0.63±0.13/6.3±3.8	0.62/4.7	-	0044421	-
T0652	138	1.00/0.85±0.08/2.7±2.0	0.87/2.2	1.25	0005515	-
C0081	241	1.75/0.96±0.05/2.3±1.8	0.96/1.7	0.52	0005886	-
C0050	139	0.55/0.79±0.09/3.5±2.4	0.88/1.5	1.52	0005737	2.3.1.57 (2.3.1.-)
C0046	132	1.43/0.91±0.06/1.8±1.5	0.94/1.4	0.75 [0.33]	0003676	3.1.31.1 (3.1.31.1)

[a]Target ID in CASP10 and CAMEO experiments. R0006, R0007 and T0652 are from CASP10 and the rest are from CAMEO experiment.
[b]Length of query sequence
[c]C-score of the I-TASSER modeling and the estimated TM-score and RMSD of the first models
[d]Actual TM-score and RMSD of the first I-TASSER model compared to the experimental structure
[e]Ligand RMSD of the first I-TASSER model to the native with value in brackets being calcium ion RMSD
[f]Top Gene Ontology (GO) term of the I-TASSER prediction
[g]Predicted Enzyme Commission (EC) number with values in parentheses being the experimental EC numbers

REFERENCES

1. Altschul, S.F. et al. *Nucleic acids research* **25**, 3389-3402 (1997).
2. Henikoff, S. & Henikoff, J.G. *J Mol Biol* **243**, 574-578 (1994).
3. Zhang, Y. http://zhanglab.ccmb.med.umich.edu/PSSpred/ (2012).
4. Frishman, D. & Argos, P. *Proteins* **23**, 566-579 (1995).
5. Needleman, S.B. & Wunsch, C.D. *J Mol Biol* **48**, 443-453 (1970).
6. Wu, S. & Zhang, Y. *Proteins* **72**, 547-556 (2008).
7. Zhang, Y., Kolinski, A. & Skolnick, J. *Biophys. J.* **85**, 1145-1164 (2003).
8. Bowie, J.U., Luthy, R. & Eisenberg, D. *Science* **253**, 164-170 (1991).
9. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. *Protein Science* **6**, 676-688 (1997).
10. Smith, T.F. & Waterman, M.S. *J. Mol. Biol.* **147**, 195-197 (1981).
11. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. *Protein Science* **9**, 232-241 (2000).

12. Edgar, R.C. & Sjolander, K. *Bioinformatics* **20**, 1301-1308 (2004).
13. Wang, G. & Dunbrack, R.L., Jr. *Bioinformatics* **19**, 1589-1591 (2003).
14. Zhang, Y. & Skolnick, J. *Proteins* **57**, 702-710 (2004).
15. Wu, S.T. & Zhang, Y. *Nucl. Acids. Res.* **35**, 3375-3382 (2007).
16. Yan, R., Xu, D., Yang, J., Walker, S. & Zhang, Y. *Sci Rep* **3**, 2619 (2013).
17. Swendsen, R.H. & Wang, J.S. *Physical Review Letters* **57**, 2607-2609 (1986).
18. Zhang, Y., Kihara, D. & Skolnick, J. *Proteins* **48**, 192-201 (2002).
19. Kyte, J. & Doolittle, R.F. *J. Mol. Biol.* **157** (1982).
20. Wu, S., Skolnick, J. & Zhang, Y. *BMC Biol* **5**, 17 (2007).
21. Zhang, Y. & Skolnick, J. *Proc. Natl. Acad. Sci. USA* **101**, 7594-7599 (2004).
22. Zhang, Y. & Skolnick, J. *Nucleic. Acids Res.* **33**, 2302-2309 (2005).
23. Wu, S.T., Skolnick, J. & Zhang, Y. *BMC Biology* **5**, 17 (2007).
24. Zhang, Y. & Skolnick, J. *J Comput Chem* **25**, 865-871 (2004).
25. Xu, D. & Zhang, Y. *Biophys J* **101**, 2525-2534 (2011).
26. Zhang, Y. *BMC Bioinformatics* **9**, 40 (2008).
27. Yang, J. & Zhang, Y., Submitted (2014).
28. Berman, H.M. et al. *Nucleic acids research* **28**, 235-242 (2000).
29. Yang, J., Roy, A. & Zhang, Y. *Nucleic acids research* **41**, D1096-1103 (2013).
30. Benson, M.L. et al. *Nucleic Acids Res* **36**, D674-678 (2008).
31. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. *J Chem Inf Model* **49**, 1079-1093 (2009).
32. Liu, T., Lin, Y., Wen, X., Jorissen, R.N. & Gilson, M.K. *Nucleic Acids Res* **35**, D198-201 (2007).
33. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. *Methods Mol Biol* **406**, 89-112 (2007).
34. Bairoch, A. *Nucleic Acids Res* **28**, 304-305 (2000).
35. The Gene Ontology Consortium *Nucleic Acids Res* **40**, D559-564 (2012).
36. Roy, A., Mukherjee, S., Hefty, P.S. & Zhang, Y., Submitted (2014).
37. Furnham, N. et al. *Nucleic acids research* **42**, D485-489 (2014).
38. Roy, A., Yang, J. & Zhang, Y. *Nucleic acids research* **40**, W471-477 (2012).
39. Roy, A. & Zhang, Y. *Structure* **20**, 987-997 (2012).
40. Xu, J. & Zhang, Y. *Bioinformatics* **26**, 889-895 (2010).
41. Yang, J., Roy, A. & Zhang, Y. *Bioinformatics* **29**, 2588-2595 (2013).
42. Ashburner, M. et al. *Nat Genet* **25**, 25-29 (2000).
43. Soding, J., Biegert, A. & Lupas, A.N. *Nucleic. Acids Res.* **33**, W244-248 (2005).