Supplemental Information

Improving Protein Structure Prediction Using Multiple Sequence-Based Contact Predictions

Sitao Wu, Andras Szilagyi, and Yang Zhang

Table S1. Cumulative GDT-TS score of the top 20 servers on the 26 FM targets in CASP9 (Data was taken from http://prodata.swmed.edu/CASP9/evaluation/domainscore_sum/human_server-best-Z.html, QUARK from our lab using a different method is not listed here). Related to Figure 6.

Server Name	Cumulative GDT-TS score
ZHANG-SERVER	39.8582
BAKER-ROSETTASERVER	34.3905
MULTICOM-CLUSTER	27.6390
CHUNK-TASSER	27.4773
MULTICOM-REFINE	25.8908
RAPTORX-MSA	25.0199
RAPTORX	24.9208
RAPTORX-BOOST	24.7337
PRO-SP3-TASSER	24.5677
MULTICOM-NOVEL	24.3617
MULTICOM-CONSTRUCT	24.2406
PRDOS2	20.4793
GWS	19.7966
PHYRE2	19.2882
JIANG_ASSEMBLY	19.1613
MUFOLD-MD	18.7891
GSMETASERVER	17.8072
ZHOU-SPARKS-X	16.4345
PCOMB	15.8549
MUFOLD-SERVER	15.0945

Supplemental Experimental Procedures

I-TASSER with template-based contact predictions

I-TASSER uses a statistical energy function derived from structures in the PDB. In addition to the generic energy terms, contact restraints derived from templates are used for protein structure prediction. Contact restraints are of two types: those for C_α atoms and those for side-chain group (SG) centers. Two C_α atoms are considered to be in contact if their distance is <6Å and their sequence separation is ≥6. The definition of side-chain center contacts is quite different. The sequence separation threshold is the same (i.e. 6) but the distance cutoff is defined by

$$cut(A,B) = d(A,B) + 2.5\Delta(A,B)\,. \qquad \text{S1}$$

Here $d(A,B)$ and $\Delta(A,B)$ are the mean and the standard deviation of the distance between the side-chain centers of amino acids A and B that are in contact, as calculated from 6,379 non-homologous PDB structures. Two residues are considered to be in contact if at least one pair of their heavy atoms is closer than 4.5Å. The derived values of $d(A,B)$ and $\Delta(A,B)$ for all amino acid pairs as listed at our website at http://zhanglab.ccmb.med.umich.edu/LOMETS/sidechain_contact.txt.

From the multiple sequence alignments generated by LOMETS, we generate template-based contact predictions by selecting residue pairs that are in contact in multiple templates, the equivalence of the residues being defined by the multiple alignment. The confidence score is simply the number of templates where the contact occurs divided by the total number of templates in the multiple alignments. For any given conformation generated during the I-TASSER simulation, and for a residue pair (i, j) predicted to be in contact, we calculate the C_α distance (d_{ca}^{ij}) and the side-chain center distance (d_{sg}^{ij}). The contact energy used in I-TASSER is then defined as

$$E_{contact\_temp} = w_1 \sum_{(i,j)\in,list_1,\, j\geq i+6} f(d_{ca}^{ij}) + w_2 \sum_{(i,j)\in,list_2,\, j\geq i+6} g(d_{sg}^{ij})\,. \qquad \text{S2}$$

The first term in $E_{contact\_temp}$ stands for the C_α atom contact energy. The summation runs for the list $list_1$ of predicted C_α contacts having a confidence score ≥0.1. The function $f(.)$ assigns a rewarding energy to the pair (i, j) if their distance x is shorter than a cutoff:

$$f(x) = \begin{cases} score_{ij}, if\ x < d_0 \\ 0, \qquad if\ x \geq d_0 \end{cases}. \qquad \text{S3}$$

Here, d_0 is the distance cutoff, which is 6 Å for C_α atoms. [When this function is used in another context (see Eq. S7) for side-chain center contacts, the distance cutoff depends on the involved amino acids: d_0=$cut(A,B)$.] The energy contribution of the contact $score_{ij}$ depends on the confidence of the predicted contacts, with more confident contacts having a more favorable contribution:

$$score_{ij} = \begin{cases} -1-(conf_{ij}-cut_0)^4, if\ conf_{ij} > cut_0 \\ -1+(conf_{ij}-cut_0)^2, if\ conf_{ij} \leq cut_0 \end{cases}, \qquad \text{S4}$$

where $conf_{ij}$ is the confidence score of the predicted contact (i, j). The parameter cut_0 determines how confident a contact must be to have a score of −1, and is set by target type ("easy/medium/hard"). The definition of easy/medium/hard targets is based on the significance of LOMETS templates.

The second term in Eq. S2 stands for the side-chain center contact energy. The summation runs for the list $list_2$ of predicted side-chain center contacts having a confidence score ≥0.1, and the function $g(.)$ assigns an energy penalty to the pair (i, j) if their distance x is longer than a cutoff (threshold):

$$g(x) = \begin{cases} 0, & \text{if } x < cut(AA_i, AA_j) \\ -score_{ij}, & \text{if } x \geq cut(AA_i, AA_j) \end{cases}. \qquad \text{S5}$$

Here AA_i is the amino acid type of residue i, $cut(.,.)$ is as defined in Eq. S1, and $score_{ij}$ is as defined in Eq. S4. To balance the energy contributions of C_α atom and side-chain center contacts, the weights w_1 and w_2 are set by target type; a higher relative weight is assigned to the C_α contacts for hard targets while the weights are equal for easy targets.

I-TASSER with combined sequence- and template-based contact predictions

The SVMSEQ program was originally developed to predict C_α atom contacts with a distance cutoff of 8Å. To increase the diversity of sequence-based contact predictions, we extended the original SVMSEQ and constructed 9 different versions of SVMSEQ predictors (based on the same training proteins) to predict contacts for C_α, C_β atoms and side-chain centers with distance cutoffs 6Å, 7Å and 8Å for C_α and C_β, and 0.8**cut*(*A*,*B*), *cut*(*A*,*B*) and 1.2**cut*(*A*,*B*) for side-chain centers. Thus, the 9 different SVMSEQ predictors include 3 predictors for C_α contacts, 3 for C_β contacts, and 3 for side-chain center contacts. We have designed two ways to combine the sequence-based contact predictions with the template-based ones and to include them into I-TASSER's energy function.

The first solution retains the functional form of the contact restraints shown in Eq. S2 but replaces the purely template-based contacts with a consensus of sequence- and template-based contacts. The two new sets of contacts (one for C_α atoms and one for side-chain centers) are generated by the following procedure. First, we take a weighted average of the confidence scores of the 9 sequence-based and 1 template-based contacts and form a new, consensus confidence score:

$$\text{conf}(i, j) = \sum_{n=1}^{10} w_n \text{conf}_n(i, j), \qquad \text{S6}$$

where conf(i, j) is the consensus C_α (or side-chain center) contact confidence score for residues i and j, $\text{conf}_n(i, j)$ is the contact confidence score for the nth individual predictor (9 predictors are sequence-based, and 1 predictor is template-based), w_n is the weighting factor for the nth predictor. (The weight for the template-based contacts was set to 0.6 while those for sequence-based ones were all set to 0.1, after trying a few combinations.) Then, the contacts with a consensus confidence score <0.1 are discarded, and thus a new set of contacts are generated and then used in I-TASSER in the usual way, via Eq. S2.

The other solution to utilize the sequence-based contact predictions in I-TASSER is to keep the energy contribution of the template-based contacts as described by Eq. S2, and add a further energy term that introduces the sequence-based contact predictions. Here, we simply use one energy term at the right side of Eq. S7 for sequence-based predicted contacts:

$$E_{contact\_consensus} = E_{contact\_temp} + \sum_{k=1}^{9} w_k \sum_{(i,j)\in list_k, j\geq i+6} f(d_k^{ij}), \qquad \text{S7}$$

where $f(.)$ is the energy function defined in Eq. S3, which rewards residue pairs that satisfy the predicted contact constraints; k stands for the kth sequence-based contact predictor; d_k^{ij} is the (C_α, or C_β or side-chain center) distance between residues i and j for the kth predictor, $list_k$ is the list of contacts predicted contacts by the kth predictor, w_k is the weighting factor for the kth predictor. Here, w_k was set for the same value for each k.

Testing both methods on an independent training set of hard, medium, and easy targets, we found that the first solution works better (i.e. yields more accurate models) for hard targets, and the second option works better for medium and easy targets. This is expected, as explained in the Discussion section.

Both methods have a number of tunable parameters: thresholds for the confidence scores, weighting factors, as well as the cut_0 parameter (Eq. S4) in the energy function. The relationship between the confidence threshold and contact accuracy was derived by performing prediction on a set of 124 proteins including hard, medium, and easy targets. This set was independently constructed and is non-homologous to either the training set of SVMSEQ or the hard, medium and easy test sets used in the paper. The cut_0 values were derived for all 9 variants of SVMSEQ, and for all three target types (27 values), and were used as confidence thresholds to select the sequence-based contacts to be included in the consensus method by Eq. S6 and the lists in Eq. S7. The weights in Eqs. S2, S6, and S7 were adjusted by trying ~10 combinations with the 124-protein set, optimizing the average TM-scores of the first models generated by I-TASSER. All the parameters can be found at: http://zhanglab.ccmb.med.umich.edu/contact/parameter.pdf.

Modeling results on the 26 Free Modeling targets in CASP9

Table S1 lists the cumulative GDT-TS score of Zhang-Server (which used SVMSEQ contact predictions in I-TASSER simulations), together with that by 19 best ranking server groups from other groups, on the 26 CASP9 FM targets/domains. The 26 defined hard targets are T0529_1, T0531_1, T0534_1, T0534_2, T0537_1, T0537_2, T0544_1, T0547_3, T0547_4, T0550_1, T0550_2, T0553_1, T0553_2, T0561_1, T0571_1, T0571_2, T0578_1, T0581_1, T0604_1, T0604_3, T0608_1, T0616_1, T0618_1, T0621_1, T0624_1, and T0629_2. The domain definition and the experimental structures with reordered residues are available at http://predictioncenter.org/casp9/domain_definitions.cgi. The GDT-TS scores was taken from the assessor's website at http://prodata.swmed.edu/CASP9/evaluation/domainscore_sum/human_server-best-Z.

html. The QUARK method developed in our lab is not listed in the table, which took a complete different approach of *ab initio* folding from I-TASSER and will be discussed elsewhere (Xu and Zhang, 2010a, b).

Supplemental References

Xu, D., and Zhang, Y. (2010a). QUARK ab initio protein structure prediction I: Method developments. Submitted.

Xu, D., and Zhang, Y. (2010b). QUARK ab initio protein structure prediction II: results of benchmark and blind tests. Submitted.