Supplementary Information

October 5, 2015

1 Flowchart of FALCON@home

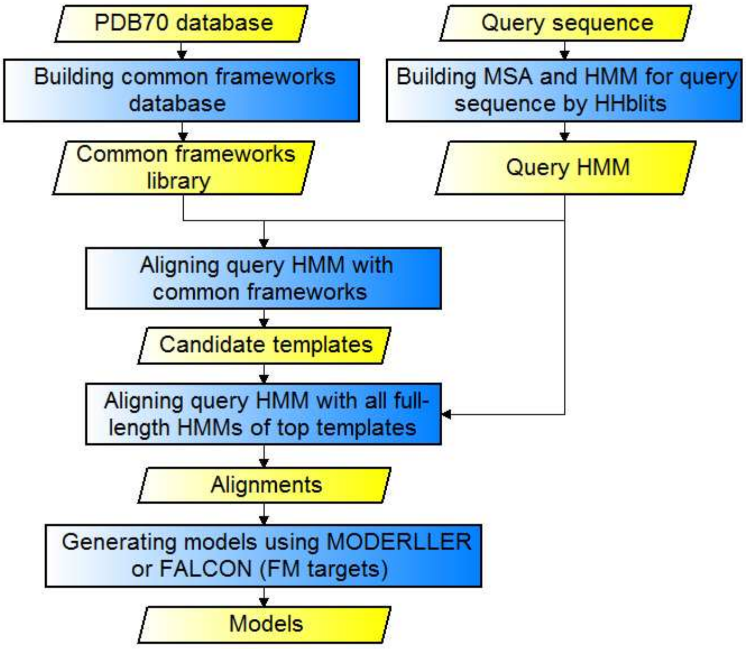
PDB70 database
Query sequence
Building common frameworks database
Building MSA and HMM for query sequence by HHblits
Common frameworks library
Query HMM
Aligning query HMM with common frameworks
Candidate templates
Aligning query HMM with all full-length HMMs of top templates
Alignments
Generating models using MODERLLER or FALCON (FM targets)
Models

Figure S1: Flowchart of FALCON@home protein structure prediction server.

2 TBM module

In the TBM module, we first calculated the common structural frameworks, and then aligned the query protein against the common structural frameworks. The details of TBM module are described as below.

2.1 Identification of common framework shared by homologous proteins

For each template with known structure, all of its homologous proteins were first identified based on sequence and structure similarity. Then, an integer linear program was designed to identify the common framework shared by these homologous proteins. The constraints of the integer linear program guarantee the common framework to be conservative with respect to structure and sequence [1].

Figure S2 shows the common framework identified for protein `1b7y_A` as an example. In general, a common framework consists of a set of dispersed conserved segments. The sequence profiles, profile hidden Markov models (HMM) of these segments, as well as the lengths of the gaps between neighboring segments, are stored for further fold recognition and alignment steps.

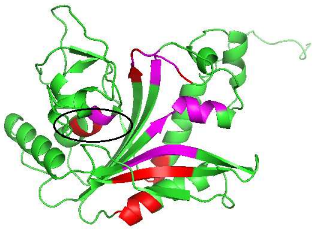

Figure S2: Common framework (in red and purple) shared by protein `1b7y_A` and its homologous proteins.

2.2 Alignment of the query protein against the common frameworks

We aligned a given query protein sequence against the identified common frameworks to avoid vague alignments rooted in the structurally variable segments of templates. Specifically, using profile HMMs as a generative model, the probability of the query protein generated from a certain common framework was first calculated. The probability consists of two parts, i.e., the probability that the conserved segments generate the matched segments in query protein and the probability that gaps in the framework generates unmatched segments in query proteins. The common frameworks with high probability were kept for final model generation.

After recognizing the likely folds by searching against the common frameworks, the full-length alignments were generated via aligning the query sequence against identified templates using TreeThreader [2, 3]. The final structural models were generated by MODELLER [4] and selected according to the dDFIRE [5] energy function.

3 FALCON *ab initio* module

FALCON [6] is an *ab initio* prediction approach that generates models from the very beginning following an iterative strategy. To be specific, FALCON uses *Cosine* model to describe the local bias of torsion angle pair (ϕ, ψ) of each residue. A position specific HMM is used to capture the dependencies among local biases of adjacent residues, based on carefully selected fragments. The Fragment-HMM is used to sample a sequence of torsion angle pairs for the given protein sequence. ROSETTA energy function is used to evaluate the generated decoys, and to direct the sampling process to the better decoys. The generated decoys are fed back to produce more accurate estimations of local structural biases, a more accurate Fragment-HMM and thus, better decoys. This step is executed iteratively to increase the quality of the final decoys, until convergence.

In addition, we have tuned the weight of each ROSETTA energy item when generating the model [7] and ranked the models according to the combined energy scores of dDFIRE [5] and ROSETTA [8].

4 The performance of FALCON@home and HHsearch+Modeller in CASP11 evaluation

We registered a total of four servers in the CASP11 competition. Among these servers, the FALCON_TOPO server is equivalent to FALCON@home (ranked 12th over TBM domains and 16th over FM domains according to the Assessors' formula). Another server, called FALCON_EnvFold, is an enhanced version of FALCON@home—besides the sequence information used in FALCON@home, local structural information is also employed to build query-template alignment in FALCON_EnvFold. In CASP11, FALCON_EnvFold was ranked 9th over TBM category according to GDT_TS measure. Notice that the CASP11 website lists only the overall performance of each participating server, which states that FALCON_TOPO was ranked 12th and HHpredA was ranked 17th over the 81 TBM domains. For the sake of detailed performance comparison, we listed the prediction model quality for each query protein individually. In particular, we first downloaded from CASP11 website the models predicted by FALCON_TOPO and HHpredA; then we run TM-score to calculate GDT_TS of each predicted model. Over the total of 105 TBM and FM domains, FALCON_TOPO showed comparable performance with HHpredA, and outperformed HHpredA when GDT_TS is over 0.6.

The comparison is summarized in Fig. S3, Tables S1 and S2.

FALCON_TOPO server also shows the advantage in remote homologue identification. Take the target T0678 as an example; the challenge was to determine how to align the three N-terminal strands. Using the pre-calculated common frameworks, the FALCON@home successfully identified the most similar template as `4gt6_A` and finally generated a high-quality prediction model with a TM-score [9] of 0.84 to the native structure (Fig. S4).

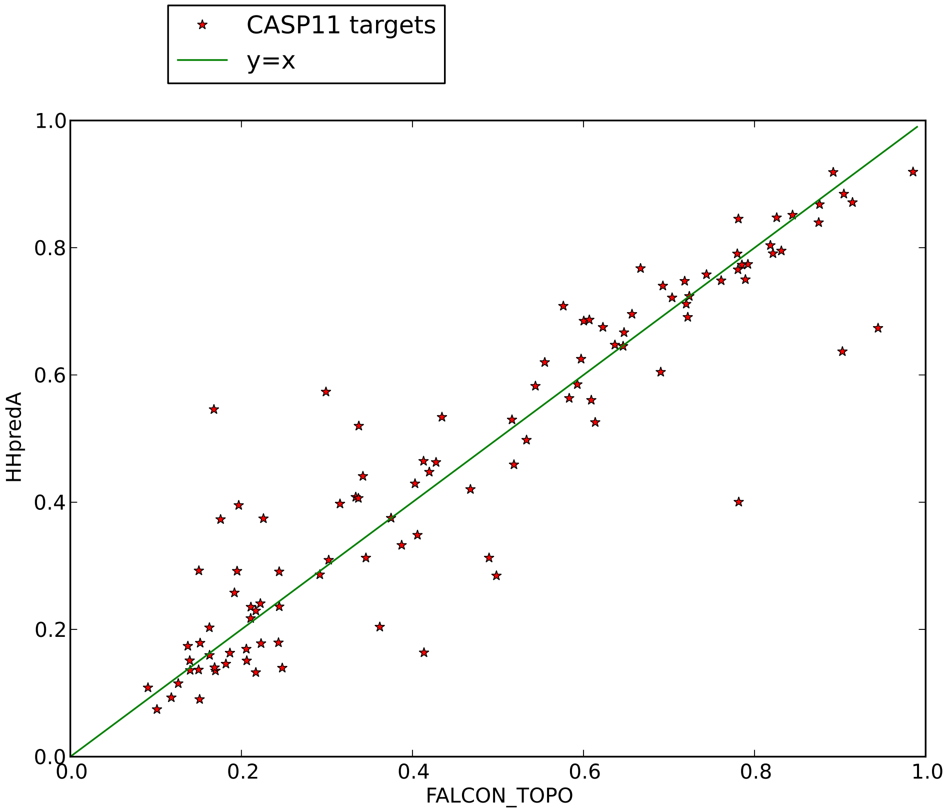

Figure S3: Comparison of FALCON_TOPO with HHpredA using the GDT_TS measure over 105 CASP11 domains. Over the total of 105 CASP 11 domains, FALCON_TOPO showed comparable performance with HHpredA, and outperformed HHpredA when GDT_TS is over 0.6.

Domain	HHpredA	FALCON_TOPO
T0759-D1	0.9191	0.9853
T0759-D2	0.3750	0.3750
T0760-D1	0.7400	0.6928
T0761-D1	0.2358	0.2443
T0761-D2	0.1792	0.2434
T0762-D1	0.8473	0.8259
T0763-D1	0.1365	0.1500
T0764-D1	0.7578	0.7438
T0765-D1	0.4408	0.3421
T0766-D1	0.6736	0.9444
T0767-D1	0.2039	0.3618
T0767-D2	0.1458	0.1819
T0768-D1	0.4003	0.7815
T0769-D1	0.5851	0.5928
T0770-D1	0.6848	0.6003
T0771-D1	0.1738	0.1374
T0772-D1	0.6472	0.6367
T0773-D1	0.6045	0.6903
T0774-D1	0.4474	0.4197
T0776-D1	0.8037	0.8185
T0777-D1	0.0928	0.1181
T0780-D1	0.7237	0.7237
T0780-D2	0.6198	0.5547
T0781-D1	0.1150	0.1262
T0781-D2	0.3729	0.1757
T0782-D1	0.6750	0.6227
T0783-D1	0.7675	0.6667
T0783-D2	0.1635	0.4135
T0784-D1	0.8680	0.8760
T0785-D1	0.1786	0.1518
T0786-D1	0.6452	0.6463
T0789-D1	0.2028	0.1626
T0789-D2	0.1508	0.2063
T0790-D1	0.2352	0.2111
T0790-D2	0.1692	0.2058
T0791-D1	0.1594	0.1628
T0791-D2	0.1630	0.1866
T0792-D1	0.6667	0.6474
T0794-D1	0.6866	0.6068
T0794-D2	0.0901	0.1512
T0796-D1	0.4645	0.4130
T0800-D1	0.4080	0.3337
T0801-D1	0.8514	0.8444
T0803-D1	0.4291	0.4030
T0805-D1	0.7475	0.7183
T0806-D1	0.1084	0.0908
T0807-D1	0.7951	0.8313
T0808-D1	0.5458	0.1679
T0808-D2	0.0743	0.1013
T0810-D1	0.1394	0.2478
T0810-D2	0.6956	0.6567

Table S1: Comparison FALCON_TOPO and HHpredA using the GDT_TS measures over CASP11 targets (Part I, from T0759 to T0810)

Domain	HHpredA	FALCON_TOPO
T0811-D1	0.8845	0.9044
T0812-D1	0.3324	0.3874
T0813-D1	0.7740	0.7922
T0814-D1	0.1350	0.1697
T0814-D2	0.1358	0.1401
T0814-D3	0.4062	0.3368
T0815-D1	0.8396	0.8750
T0816-D1	0.3125	0.3456
T0817-D1	0.8453	0.7811
T0817-D2	0.7905	0.7798
T0818-D1	0.3974	0.3153
T0819-D1	0.7909	0.8215
T0820-D1	0.2861	0.2917
T0820-D2	0.6250	0.5972
T0821-D1	0.5255	0.6137
T0822-D1	0.4627	0.4276
T0823-D1	0.5634	0.5833
T0824-D1	0.2407	0.2222
T0827-D1	0.3951	0.1969
T0827-D2	0.2917	0.1950
T0829-D1	0.4590	0.5187
T0830-D1	0.3483	0.4059
T0830-D2	0.1779	0.2230
T0831-D1	0.3742	0.2258
T0831-D2	0.1510	0.1396
T0832-D1	0.1400	0.1687
T0833-D1	0.7083	0.5764
T0834-D1	0.2576	0.1919
T0834-D2	0.2907	0.2442
T0835-D1	0.4202	0.4678
T0836-D1	0.1324	0.2169
T0837-D1	0.2293	0.2169
T0838-D1	0.5337	0.4345
T0840-D1	0.6368	0.9027
T0840-D2	0.5734	0.2989
T0841-D1	0.8712	0.9145
T0843-D1	0.7730	0.7852
T0845-D1	0.5825	0.5438
T0845-D2	0.4977	0.5334
T0847-D1	0.6908	0.7219
T0848-D1	0.2844	0.4982
T0848-D2	0.2923	0.1503
T0849-D1	0.5604	0.6091
T0851-D1	0.7483	0.7610
T0852-D1	0.7115	0.7201
T0852-D2	0.5198	0.3373
T0853-D1	0.5296	0.5164
T0853-D2	0.3090	0.3021
T0854-D1	0.9186	0.8920
T0854-D2	0.7214	0.7036
T0855-D1	0.2174	0.2109
T0856-D1	0.7500	0.7893
T0857-D1	0.3125	0.4896
T0858-D1	0.7656	0.7806

Table S2: Comparison FALCON_TOPO and HHpredA using the GDT_TS measures over CASP11 targets (Part II, from T0811 to T0858)

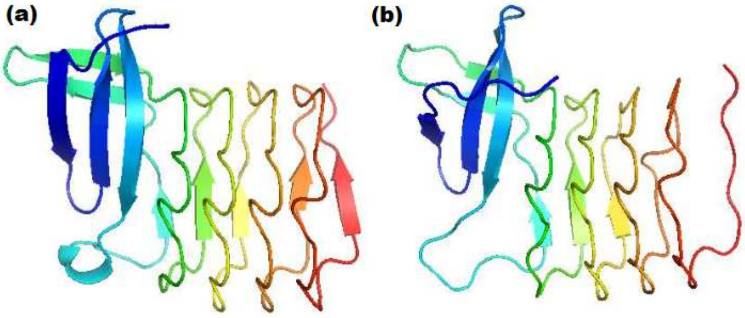

Figure S4: (a) Native structure of TBM protein T0768. (b) Prediction model of T0768 by FALCON_TOPO with a TM-score=0.84 compared with the native structure.

5 Comparision of FALCON@home with HHsearch+Modeller over 1263 PDB70 domains

Besides the CASP11 targets, we also compared FALCON@home against HHsearch+Modeller over a collection of 1263 PDB70 proteins whose native structures were released after the CASP11 evaluation. The 1263 PDB70 proteins were selected from all proteins with newly-released structures by filtering out the proteins that are too short (length < 50) or multiple-domains. The list of the 1263 proteins can be downloaded from the following website: http://protein.ict.ac.cn/FALCON/testset-1263proteins.tgz.

To avoid the overlap between query proteins and the template databases, both FALCON@home and HHsearch were executed over the template databases built before the CASP11 evaluation. Over these proteins, FALCON@home exhibited an average GDT_TS score of 0.68, which is slightly higher than HHsearch+Modeller (0.66). However, FALCON@home is more efficient: it took $\sim$23 hours for FALCON@home to make predictions for the 1263 proteins, while HHsearch+Modeller used $\sim$74 hours.

The comparison is graphically shown in Fig. S5.

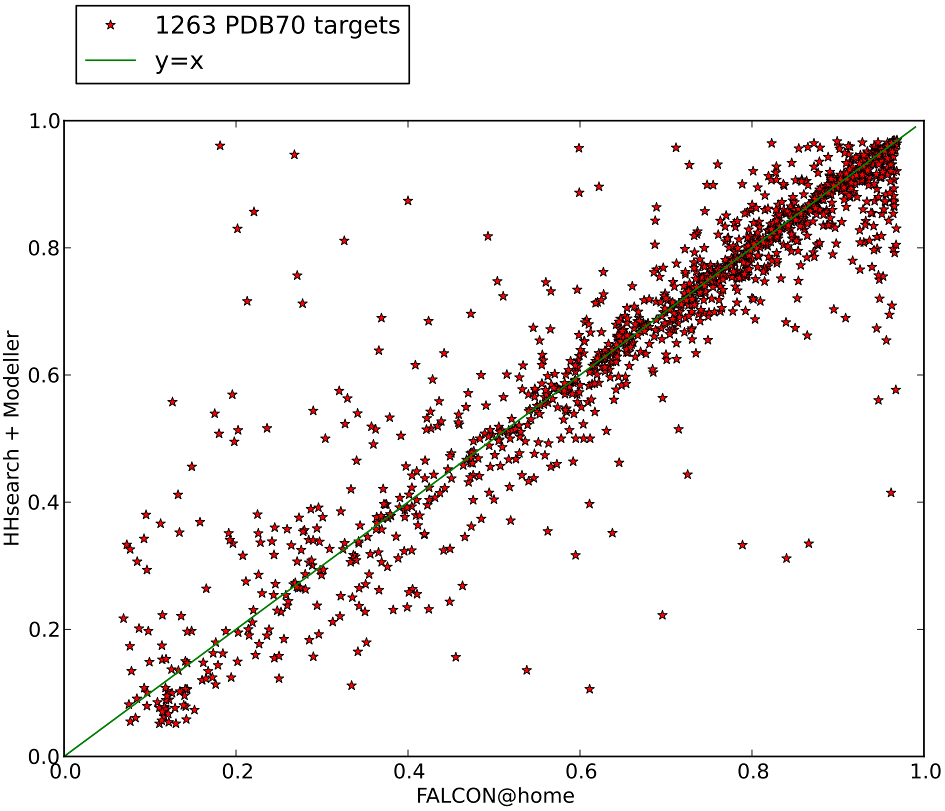

Figure S5: Comparison of FALCON@home with HHsearch+Modeller using the GDT_TS measure over 1263 PDB70 domains with native structures released after the CASP11 evaluation. Over the 1263 PDB70 proteins, FALCON@home exhibited an average GDT_TS score of 0.68, which is slightly higher than HHsearch+Modeller (0.66).

References

[1] Zhu J, Zhang H, Wang C, Ling B, Zheng WM, Bu D. TOPO: Improving remote homologue recognition via identifying common protein structure framework. arXiv preprint arXiv:150703197. 2015;.

[2] Wu W, Chen G, Kan W, Anderson D, Grey F, Li J, et al. Harness public computing resources for protein structure prediction computing. In: The International Symposium on Grids and Clouds (ISGC). vol. 2013; 2013. .

[3] Zhang H, Shao M, Wang C, Zhu J, Zheng WM, Bu D. Improving protein threading accuracy via combining local and global potential using TreeCRF model. arXiv preprint arXiv:150903434. 2015;.

[4] Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen My, et al. Comparative protein structure modeling using Modeller. Current protocols in bioinformatics. 2006;p. 5–6.

[5] Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins: Structure, Function, and Bioinformatics. 2008;72(2):793–803.

[6] Li SC, Bu D, Xu J, Li M. Fragment-HMM: A new approach to protein structure prediction. Protein Science. 2008;17(11):1925–1934.

[7] Wang C, Wei Y, Liu J, Zhang H, Ling B, Li SC, et al. Optimizing weights of protein energy function to improve ab initio protein structure prediction. arXiv preprint arXiv:13126960. 2013;.

[8] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of molecular biology. 1997;268(1):209–225.

[9] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics. 2004;57(4):702–710.