EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction

Robert Sheridan[1,*], Robert J. Fieldhouse[1,2*], Sikander Hayat[1,2], Yichao Sun[1], Yevgeniy Antipin[1], Li Yang[2], Thomas Hopf[2,3], Debora S. Marks[2,§], Chris Sander[1,§]

1 Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA
2 Department of Systems Biology, Harvard Medical School, Boston, MA, USA
3 Department for Bioinformatics and Computational Biology, Technical University of Munich, Germany
* Joint first authors. § Joint senior authors.

ABSTRACT

Recently developed maximum entropy methods infer evolutionary constraints on protein function and structure from the millions of protein sequences available in genomic databases. The EVfold web server (at EVfold.org) makes these methods available to predict functional and structural interactions in proteins. The key algorithmic development has been to disentangle direct and indirect residue-residue correlations in large multiple sequence alignments and derive direct residue-residue evolutionary couplings (EVcouplings or ECs). For proteins of unknown structure, distance constraints obtained from evolutionarily couplings between residue pairs are used to *de novo* predict all-atom 3D structures, often to good accuracy. Given sufficient sequence information in a protein family, this is a major advance toward solving the problem of computing the native 3D fold of proteins from sequence information alone.

Availability: EVfold server at http://evfold.org/

Contact: evfoldtest@gmail.com

Abbreviations: DI: direct information; EC: evolutionary coupling; EV: evolutionary; MSA: multiple sequence alignment; PLM: pseudo-likelihood maximization; PPV: positive predictive value (number of true positives divided by the sum of true and false positives); TM-score: template modeling score

1 INTRODUCTION

Evolution of species is now revealed at the molecular level through genomic sequencing. Extensive output from high-throughput sequencing has made it increasingly productive to mine this data for interactions affecting protein structure and function. Knowledge of which protein residues are involved in functionally important interactions and how these are arranged in space benefits research in many areas of biology. A long-standing goal in computational biology has been to predict 3D structure from amino acid sequence alone. As the Critical Assessment of Techniques for Protein Structure Prediction (CASP) has shown, accurate structure prediction remains a largely unsolved challenge, especially in the absence of a homologous template [1]. In addition, predicting the identity and role of functional residues in incompletely characterized proteins is another persistent challenge [2]. However, recent developments have led to a breakthrough in computational methods employing co-evolution [3–6] for both structure and function prediction. The use of evolutionary couplings (ECs) between residues to accurately predict all-atom protein structures was, to our knowledge, first demonstrated in the fall of 2010 and published in 2011 [7].

The EVcouplings method uses a global probability model for an isostructural set of protein sequences in the form of an exponential model, which has a pseudo-energy expression up to second order (residue pair interactions) as the exponent. The parameters in the model are extracted (not fit in the sense of machine learning) from co-variation counts for all pairs of residue positions in a multiple sequence alignment (MSA) of evolutionarily related protein sequences. The inferred ECs disambiguate direct from indirect correlations. Importantly, ECs between residue pairs are often involved in key functional and structural interactions that in general cannot be detected using single-column conservation within an MSA. ECs often occur between residues in structural contact, enabling their use in *de novo* predictions of protein structure. Complementary to the EVold server, which commenced public operation in April 2013, there are now several additional online resources for protein contact prediction [8–10]. Development of methods and applications is very active in the field of evolutionary couplings, including focused efforts on beta-barrel membrane proteins [11], protein complexes [12, 13] and hybrid methods for structure determination, such as combining sparse NMR data with residue-residue ECs (EC-NMR) [14].

2 EVOLUTIONARY COUPLINGS AND PROTEIN 3D STRUCTURES

2.1 Capabilities: structures and functional interactions

For a target protein sequence, embedded in a protein family alignment, the EVfold server derives ECs, which reflect structural and functional constraints. From the residue pair constraints, the server can *de novo* model 3D structures from sequence information alone (Figure 1), without the need for the 3D structure of homologous proteins or protein fragments, as in template model building or model building by homology.

Beyond their use in computing 3D structures, the inferred ECs can be used to identify functionally constrained residue interactions indicative of active sites, protein-protein interfaces and other functional sites, and can be used to guide or interpret experiments that measure the phenotypic consequences of residue substitutions. The server can handle 3D structure prediction for globular and helical transmembrane proteins, with server capability for beta-barrel membrane proteins and for protein complexes technically feasible [11, 12], but not yet implemented.

2.2 Input: sequences

Minimal server input is a specific protein (database ID or amino acid sequence) and a sequence range (domain) within the protein. Computing ECs for a target protein sequence requires a MSA for a set of proteins plausibly isostructural with the target protein, typically paralogs from the same organism and homologs from other species. The MSA can be provided by the user or is retrieved from the Pfam domain database [15], or generated using software that searches for homologs in protein sequence databases, such as HHblits [16] or jackhmmer [17]. The depth (number and diversity of sequences) and breadth (coverage of the target protein domain by the aligned residues) of the protein family alignment has to be sufficient to allow reliable extraction of ECs. Typically, we suggest at least 5L sequences in the MSA, where L is the number of residues of the target protein domain and 75% breadth of coverage. For example, we suggest a 200 residue target protein have at least 1000 sequences in the MSA. For unknown structures, the server automatically predicts secondary structure using PSIPRED [18] and alpha-helical transmembrane topology using MEMSAT-SVM [19] and uses these as input to the 3D structure generation. Users can override the secondary structure predictions using experimental data or other predictions.

If an experimental structure or 3D structure model of the protein is known, this can be entered via a Protein Databank (PDB) identifier [20]. In the EVFold mode, the server assesses the accuracy of structure prediction from sequence alone by comparing the EVfold predicted structure to the known structure using standard 3D superimposition methods, such as the one in the PyMOL molecular graphics software. In the EVcouplings mode, the server can map ECs onto the known structure for functional interpretation, without 3D structure prediction.

2.3 Algorithm: maximum-entropy model for protein sequences

The algorithm uses a by now well-established [3,7,21–24] maximum-entropy probability model to identify ECs in the protein family. It uses a subset of the ECs to impose residue pair distance constraints in the molecular modelling software CNS [25] to generate 3D structures using distance geometry projection followed by (moderately long, as of March 2015) simulated annealing by molecular dynamics.

The server has a choice of two levels of approximation for inferring the parameters of the global, maximum-entropy probability model: “DI”, a mean-field based coupling analysis with parameters obtained by inversion of the covariance matrix and direct information scoring using an analogue of mutual information with direct information marginal probabilities [7,21]; and “PLM”, a pseudo-likelihood maximization approximation with corrected norm scoring [23]. DI, also called mean field DCA, tends to be the faster but less accurate method compared to PLM. For a review of methods see [24]. We recommend the slower but more accurate PLM as default.

2.4 Output: evolutionary couplings and/or predicted 3D structures

The server provides a list of evolutionarily coupled residue pairs ranked by EC score, 2D maps of predicted contacts and/or a set of predicted all-atom 3D structures. High-ranking ECs represent strong evolutionary constraints and tend to reflect residue proximity in the folded protein structure and/or functionally important interactions for any protein in the family. The ECs are visualized as a 2D contact map and also mapped onto a known or predicted 3D structure of the target protein, for example as lines connecting two residues. The total strength of ECs affecting a single particular residue is visualized as a residue property in 3D, with the aggregate EC strength (summed over all partner residues) as residue atomic sphere color (Figure 2) or residue thickness in “sausage” mode.

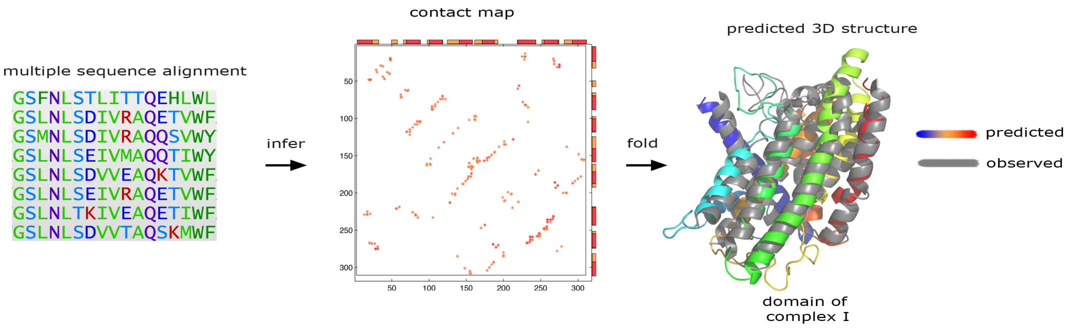


Figure 1: **EVfold server process: from amino acid sequences to all-atom 3D structures (option EVfold).** User supplies a sequence of interest (the target sequence) in the context of a large multiple sequence alignment (left) that provides residue-residue covariation information. Evolutionary constraints (ECs) are computed using DI or PLM (see Method) and the top ECs are predictive of residue-residue contacts (middle). The predicted 3D fold of the human subunit one of Complex I (right, gene name: MT-ND1, Uniprot: NU1M_HUMAN) [26] agrees well with the subsequently published crystal structure of the *Thermus thermophilus* homolog (Uniprot: NQO8_THET8, PDB: 4HE8) [27].

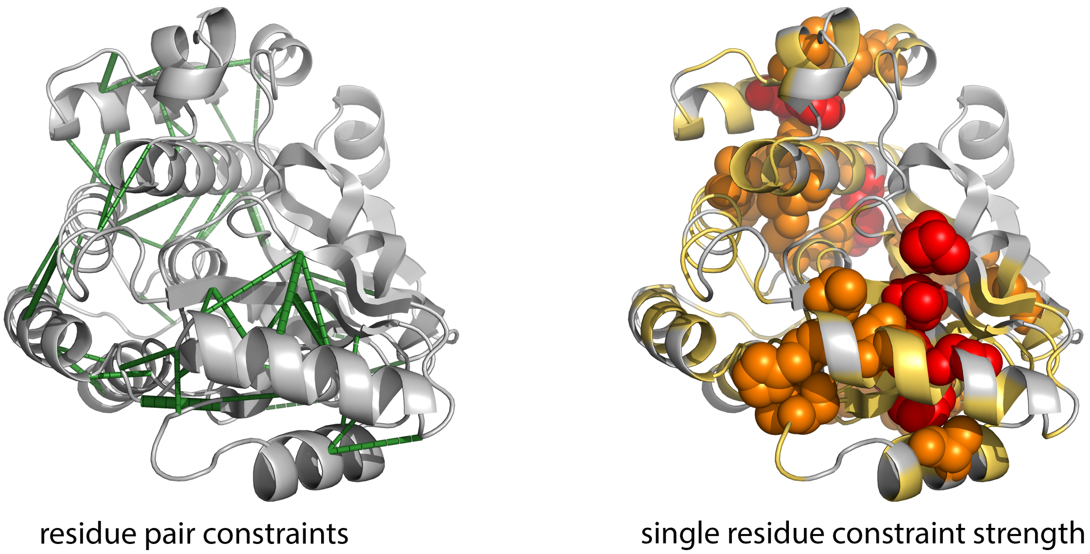


Figure 2: **Evolutionary couplings visualized on known structures or predicted structures are informative about essential residue-residue interactions (option EVcouplings for known structures, EVfold for unknown structures).** Many strong residue pair couplings (left, green lines) in a bacterial protein transacylase (Uniprot: FABD_ECOLI; PDB: 1MLA) link residues that are in contact in 3D. Residues with strong single residue constraint strength (ECs summed over all coupled residue partners) may point to interaction sites that implement evolutionary requirements (the most strongly coupled residues as red spheres, followed by orange spheres for medium strength, and yellow ribbon for low strength).

To reflect modelling uncertainty for a given set of distance constraints and stochastic decisions in the simulated annealing protocol, the server typically computes several hundred models for a given target protein, e.g, ∼2L models for a protein of length L. The models are ranked by a scoring function that estimates the likelihood of being a correct and typical protein structure, using criteria based on distributions of dihedral angles, torsion angles, solvent accessibility, constraint satisfaction, and agreement with predicted secondary structure. The model scoring function is under development and improvements are expected in a future version of the server.

2.5 Current status of prediction accuracy

For a blinded benchmark set (testing 3D prediction on proteins of known 3D structure without using any information from the known structure) good 3D models (TM-score $\geq$ 0.5 [28] can be predicted, in the current implementation, in about half of all cases (Table 1). For this test, we used a structurally representative set of domains of known structure (using the CATH database as a guide [29]) and, following sequence searches and MSA construction, used a stringent cutoff on the minimal depth and breadth of the multiple sequence alignment. In particular, we required that the effective number of non-redundant sequences (Meff) normalized by protein length (L) is greater than 4.0 (depth) and that the coverage of the protein domain being modeled by non-gappy alignment columns exceeds 75 percent (breadth). In the reduced set of 63 protein domains (out of 140), which exceeded these thresholds, 38 (60%) had very good prediction accuracy (TM-score $>=$ 0.5), while 25 (40%) had accuracy below the customary threshold (TM-score $<$ 0.5). Deeper alignments (more sequences available in the family) tend to lead to better prediction accuracy. As a rule of thumb, proteins above these thresholds yield good predictions in about half the cases (Figure 3, Table 1).

While the TM-score as a quantitative measure of 3D structure prediction accuracy is useful, and often used, it does not fully reflect the essence of the success of the EVfold method, which in many cases correctly predicts the topographical arrangement, i.e., the relative spatial arrangement of secondary structure elements. In some cases predicted structures with TM-score $<$ 0.5 have perfect prediction of topography, while for TM-score $>$ 0.5 there can be topographical errors. We therefore provide, for intuitive inspection by the reader, a number of explicit examples of predicted structures in the 'well folded' (Figures 4-6) and 'less well or badly folded' (Figures 7-9) categories, both as 2D contact maps as well as 3D structural superimpositions on the known structure. Coordinates for these structures are on the EVfold.org web site. While a purist comparison of prediction accuracy would only assess the top-ranked predicted structure, we typically compute prediction error in the blinded test for the actually best structure in the top ten ranked ones, on the grounds that serious exploration of predicted structure, for functional interpretation for example, can afford scanning through ten predicted structures. Prediction accuracy for the single top-ranked structure are provided for completeness.

2.6 Factors affecting prediction accuracy and future improvements

The most important factor affecting 3D structure prediction accuracy is the depth and diversity of sequence information in the family containing the target protein domain. Inspection of about 50 cases, anecdotally, suggests additional factors related to lower prediction success,

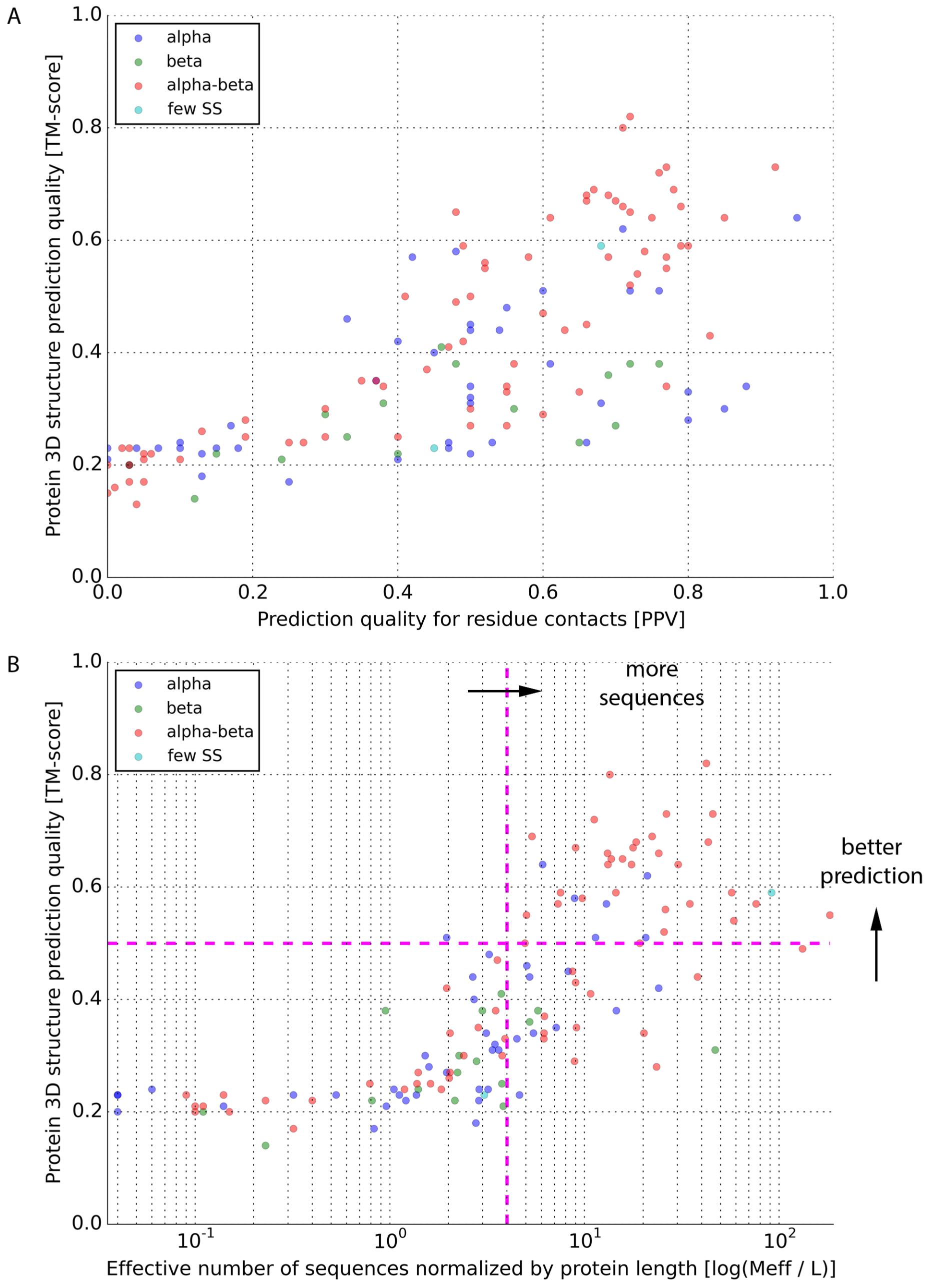


Figure 3: **EVfold benchmark on 140 proteins of known structure indicates requirements for the amount of sequence information for good predictions.** Folding results from proteins (dots) with diverse CATH topologies. (A) 3D structure prediction quality as measured by TM-score as a function of the prediction quality for residue contacts as measured by PPV. (B) 3D structure prediction quality as measured by TM-score as a function of the effective number of sequences normalized by protein length. Proteins in the upper right quadrant ('well folded' in Figures 4-6) were well predicted. Proteins in the lower right quadrant were less well predicted ('less well or badly folded' in Figures 7-9) although they passed the threshold for sufficient sequence information. Runs were done at E-value 10^{-4}, 5 jackhmmer iterations, and filtering of alignment columns and rows if gap content exceeded 30 percent. PPV is the positive predictive value (number of true positives divided by the sum of true and false positives). TM-score is the template modeling score (>0.5 considered good). Results are for the best model within top 10 ranked. Secondary structure protein fold types are alpha helical (blue), beta-strand (green), alpha-beta mix (red), few secondary structures ('few SS', teal).

Table 1: **Performance of EVfold on a diverse set of proteins.**

Uniprot ID	PDB ID	Start	End	PPV in number of contacts used	Max TM-score in top 1 ranked	Max TM-score in top 10 ranked	Max TM-score in all computed	Meff / L	Note
MOBA_ECOLI	1e5k	1	194	0.52	0.53	0.55	0.60	182	
BLAC_STAAU	1alq	24	244	0.48	0.36	0.49	0.57	132	T
BPT1_BOVIN	1aal	36	93	0.68	0.44	0.59	0.64	91	
EX3_ECOLI	1ako	1	268	0.58	0.52	0.57	0.62	76	
Q9WZB9_THEMA	1nf2	1	268	0.73	0.54	0.54	0.60	59	
ADK_HUMAN	1bx4	22	362	0.49	0.49	0.59	0.64	57	
CYP1_BRUMA	1a33	1	177	0.74	0.65	0.69	0.69	54	
PTHP_ECOLI	1cm2	1	85	0.95	0.71	0.72	0.75	46	
YBAK_HAEIN	1dbu	1	158	0.69	0.54	0.68	0.68	43	
FABD_ECOLI	1mla	1	309	0.72	0.78	0.82	0.83	42	
RUBR_CLOPA	1b13	1	54	0.73	0.33	0.33	0.37	39	A,L
TOLA_PSEAE	1lr0	226	347	0.55	0.31	0.44	0.50	38	H
UBC9_HUMAN	1a3s	1	158	0.77	0.57	0.57	0.57	35	
DEF_ECOLI	1bs4	2	169	0.85	0.62	0.64	0.70	30	
RS15_GEOSE	1a32	2	89	0.47	0.51	0.57	0.57	29	
RNPA_BACSU	1a6f	2	116	0.77	0.62	0.73	0.73	26	
CPD_ARATH	1fsi	1	181	0.52	0.51	0.56	0.57	26	
Y828_PYRHO	1v30	1	116	0.72	0.45	0.57	0.57	26	
RL22_THETH	1bxe	1	113	0.79	0.59	0.66	0.67	24	
GNLY_HUMAN	1l9l	63	136	0.40	0.40	0.42	0.42	24	L
Q8U3S5_PYRFU	1vk1	1	242	0.19	0.16	0.28	0.31	23	A
CCPR_YEAST	1a2f	71	361	0.71	0.60	0.62	0.63	21	
Q926X2_LISIN	2icg	1	159	0.77	0.19	0.34	0.46	20	A
LOLA_ECOLI	1iwl	22	203	0.74	0.41	0.41	0.43	19	B
RL13_PYRHO	1j3a	1	142	0.50	0.47	0.50	0.53	19	
KINH_HUMAN	1bg2	1	325	0.66	0.59	0.68	0.71	18	
RL17_THET8	1gd8	1	118	0.70	0.63	0.67	0.70	18	
Q97S59_STRPN	1g2r	1	97	0.75	0.64	0.64	0.64	17	
PSTS_ECOLI	1a40	26	346	0.72	0.54	0.65	0.67	16	
DMA_BPT4	1q0s	1	259	0.61	0.26	0.38	0.53	15	R
ACTP_ACACA	1ahq	2	138	0.80	0.59	0.59	0.64	15	
RL4_THEMA	1dmg	2	226	0.48	0.63	0.65	0.67	14	
Y1468_THEMA	1mgp	1	288	0.71	0.73	0.80	0.81	14	
YEDK_ECOLI	2icu	1	222	0.61	0.60	0.64	0.66	13	
UNG_HUMAN	1akz	94	313	0.71	0.27	0.66	0.71	13	
YEBC_ECOLI	1kon	1	246	0.72	0.46	0.51	0.52	11	
Q9K0A8_NEIMB	1rv9	1	259	0.76	0.56	0.72	0.72	11	
Y1033_PYRHO	1wmm	1	145	0.47	0.24	0.41	0.49	11	A,H
CAH2_HUMAN	12ca	1	260	0.74	0.56	0.58	0.59	10	
O27021_METTH	1ihn	1	111	0.37	0.24	0.35	0.40	9	A
PTN1_HUMAN	1a5y	1	330	0.66	0.49	0.67	0.68	9	
PTH2_HUMAN	1q7s	63	179	0.83	0.43	0.43	0.47	9	H
DMSD_SALTY	1s9u	1	204	0.72	0.55	0.56	0.60	9	
RNMC_MOMCH	1bk7	1	191	0.60	0.23	0.29	0.37	9	
Y467_VIBCH	2aj2	1	187	0.66	0.32	0.45	0.46	9	H
ARIS_PENRO	1dgp	40	339	0.44	0.19	0.47	0.47	8	R
PEBP1_BOVIN	1a44	2	186	0.79	0.57	0.59	0.61	8	
O29167_ARCFU	2isb	1	180	0.69	0.57	0.57	0.59	7	
Q0PBQ7_CAMJE	1vqr	1	285	0.37	0.28	0.35	0.48	7	R
NPT1_YEAST	1vlp	1	429	0.44	0.16	0.37	0.58	6	R
FRDA_HUMAN	1ekg	86	210	0.55	0.27	0.34	0.46	6	A,H,L
UBIC_ECOLI	1fw9	2	165	0.65	0.33	0.33	0.35	6	H
CYCP_RHOPA	1a7v	22	146	0.95	0.59	0.64	0.64	6	
ANM3_RAT	1f3l	208	528	0.78	0.29	0.37	0.41	6	H
FETP_PSEAE	1t07	1	90	0.50	0.24	0.34	0.37	5	L
LPXC_AQUAE	1p42	2	271	0.78	0.69	0.69	0.70	5	
LYSC_CHICK	132l	19	147	0.54	0.32	0.44	0.46	5	A
OLPA_TOBAC	1aun	22	229	0.69	0.23	0.36	0.44	5	H
PHLC_BACCE	1ah7	39	283	0.33	0.35	0.46	0.53	5	L
NAT_MYCSM	1gx3	1	275	0.77	0.48	0.55	0.57	5	
ARCH_THEMA	1j5u	1	124	0.41	0.44	0.50	0.52	5	
IAAT_ELECO	1b1u	1	122	0.15	0.19	0.23	0.24	5	A
HMOX1_RAT	1dve	1	267	0.80	0.28	0.38	0.50	5	L

Meff - effective number of non-redundant sequences. L - protein length. Our initial dataset included 140 proteins, which were filtered such that Meff normalized by length is greater than 4 and the coverage (alignment columns used) within the region of the protein being modeled is greater than 75 percent. After filtering, 63 cases remained.
A - Alignment issue
R - Ranking issue (best structure missed in top 10)
T - TM-score just below cutoff of 0.5 for good structures
H - Hydrogen-bond formation suboptimal (refinement issue)
L - Structurally important ligand not used in folding procedure
B - Unusual beta-barrel (unoptimized)

such as unusual amino acid composition or disulfide patterns, apparently erratic sections of the MSA, or apparently discontinuous subfamily structure. In some cases, predicted structures in very good agreement with experimental structures were at a low rank in the set of all computed structures (generally $\sim$2L), indicating less than fully adequate criteria in the ranking score. In some cases of correctly predicted topography of structure segments, well-placed beta-strands did not have the expected pattern of well-formed hydrogen bonds. Therefore, three important areas of desired improvement are: improved alignment procedure (cutoffs, filters, uneven sequence weights); improved refinement of the 3D coordinates of the protein models using more elaborate simulated annealing by molecular dynamics; and improved ranking criteria for structures in the set of predicted structures for a target protein.

3 RECOMMENDATIONS

Before submitting a protein for 3D structure prediction, it is useful to first check for protein domains in the target sequence, via the Pfam database for example, and consider submitting each domain separately. For each domain, one should check whether a 3D structure or that of a homologous 'template' is available, which would directly lead to a 3D structure model. Such models are often already deposited in the very useful database of protein models, www.proteinmodelportal.org, or can be generated using existing tools such as HHpred [30]. For any prediction or EC analysis run, one can check the alignment quality by inspection, as this is the single most important factor affecting predicition quality. We suggest a minimum number of $\sim$5L diverse sequences (depth) and coverage of the target domain over at least $\sim$0.75L (breadth). Once a contact map (top ECs) is generated, visual inspection can provide useful clues as to likely prediction success, looking for 'protein-like' structured patterns in the 2D predicted contact map. Caution is needed for homo-multimers and alternative conformations, as inferred ECs from the target sequence may reflect contacts between monomers of a homo-oligomer or contacts in alternative conformations, such as closed and open forms of a channel. The server returns a ranked set of 3D models for the target protein, typically several hundred; it is reasonable to inspect the top ranked model and perhaps about a dozen top additional ones. In our view, both the ranking score as well as the 3D structure refinement process can be substantially improved (work in progress). In summary, currently about one in two proteins with alignment quality above the recommended thresholds lead to excellent predictions. The number of proteins accessible to the method is rising rapidly as genome sequencing accelerates.

4 ACKNOWLEDGEMENTS

Author contributions

RS and LY developed the backend; YS developed the front end; YA, RS developed MSA viewer; TH, DSM developed the transmembrane pipeline; RJF, CS, SH collected and analyzed the protein datasets; RJF, CS, RS wrote the text; all authors edited the text; CS, DSM conceived of and designed the project.

Funding Support

We thank Richard Stein for helpful discussions and feedback. This work was funded by the US National Institutes of Health (R01GM106303 to CS at MSKCC and DSM at HMS) and the Canadian Natural Sciences and Engineering Research Council (fellowship to RJF).

References

[1] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round X. Proteins. 2013 Oct 29;.

[2] Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nature methods. 2013 Mar;10(3):221–227.

[3] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nature biotechnology. 2012 Nov;30(11):1072–1080.

[4] Taylor WR, Hamilton RS, Sadowski MI. Prediction of contacts from correlated sequence substitutions. Current opinion in structural biology. 2013 Jun;23(3):473–479.

[5] de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nature reviews Genetics. 2013 Apr;14(4):249–261.

[6] Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences. 2012 Jun 26;109(26):10340–10345.

[7] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PloS one. 2011;6(12):e28766.

[8] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences. 2013 Sep 24;110(39):15674–15679.

[9] Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics (Oxford, England). 2014 Nov 26;.

[10] Skwark MJ, Raimondi D, Michel M, Elofsson A. Improved contact predictions using the recognition of protein like contact patterns. PLoS computational biology. 2014 Nov 6;10(11):e1003889.

[11] Hayat S, Sander C, Marks DS, Elofsson A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. Proceedings of the National Academy of Sciences. 2015;p. 201419956.

[12] Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014;3.

[13] Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014 May 1;3:e02030.

[14] Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. Protein structure determination by combining sparse NMR data with evolutionary couplings. Nature Methods. (in press);.

[15] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014 Jan 1;42(1):D222–30.

[16] Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods. 2011 Dec 25;9(2):173–175.

[17] Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC bioinformatics. 2010 Aug 18;11:431–2105–11–431.

[18] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology. 1999 Sep 17;292(2):195–202.

[19] Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC bioinformatics. 2009 May 26;10:159–2105–10–159.

[20] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000 Jan 1;28(1):235–242.

[21] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011 Dec 6;108(49):E1293–301.

[22] Lapedes A, Giraud B, Jarzynski C. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. 2012 07;Available from: http://arxiv.org/abs/1207.2484.

[23] Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical reviewE, Statistical, nonlinear, and soft matter physics. 2013 Jan;87(1):012707.

[24] Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy models. PLoS computational biology. 2015;(in press).

[25] Brunger AT. Version 1.2 of the Crystallography and NMR system. Nature protocols. 2007;2(11):2728–2733.

[26] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012 Jun 22;149(7):1607–1621.

[27] Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. Crystal structure of the entire respiratory complex I. Nature. 2013 Feb 28;494(7438):443–448.

[28] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004 Dec 1;57(4):702–710.

[29] Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 2015 Jan;43(Database issue):D376–81.

[30] Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009;77 Suppl 9:128–132.

Well folded proteins

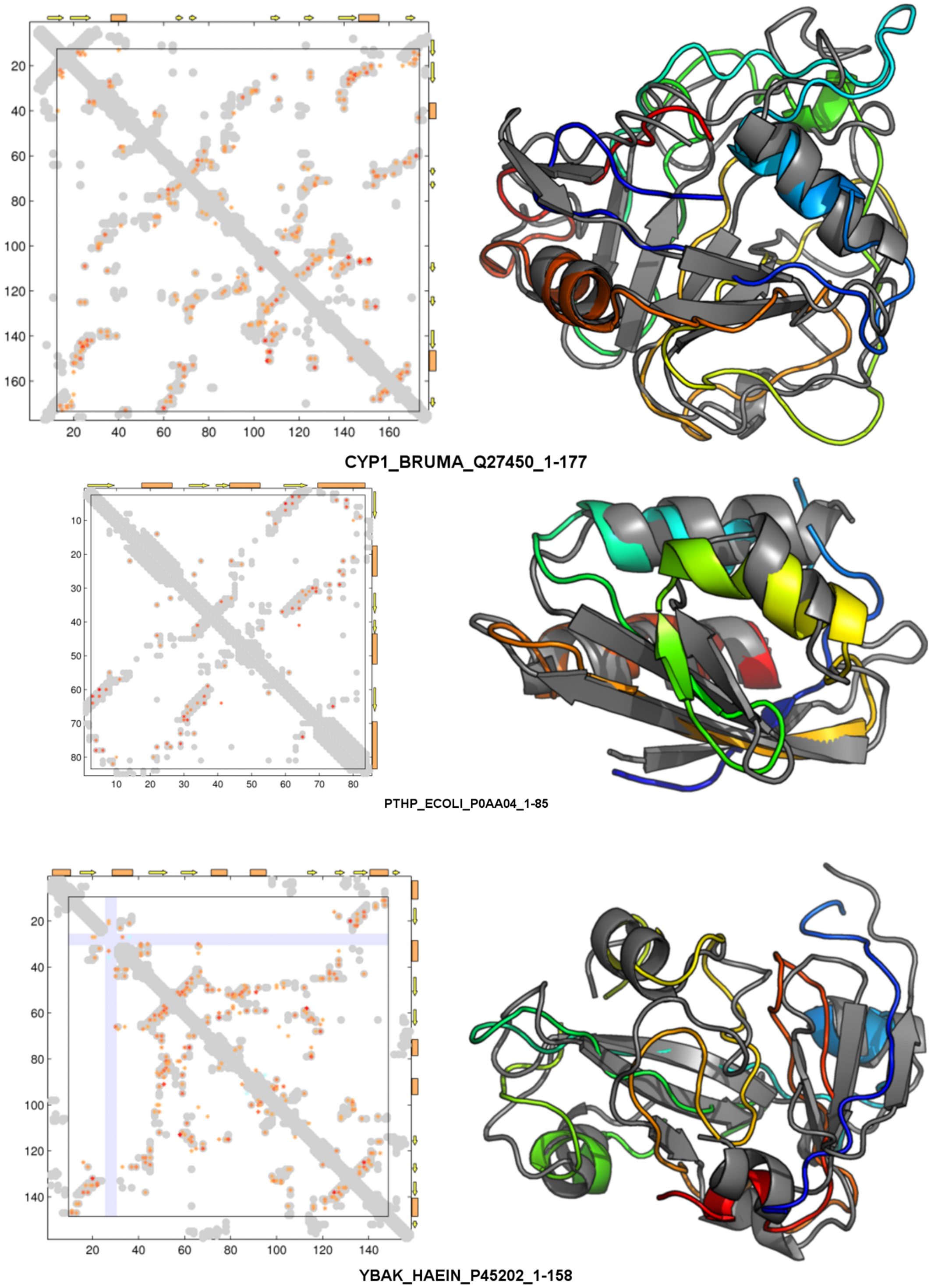
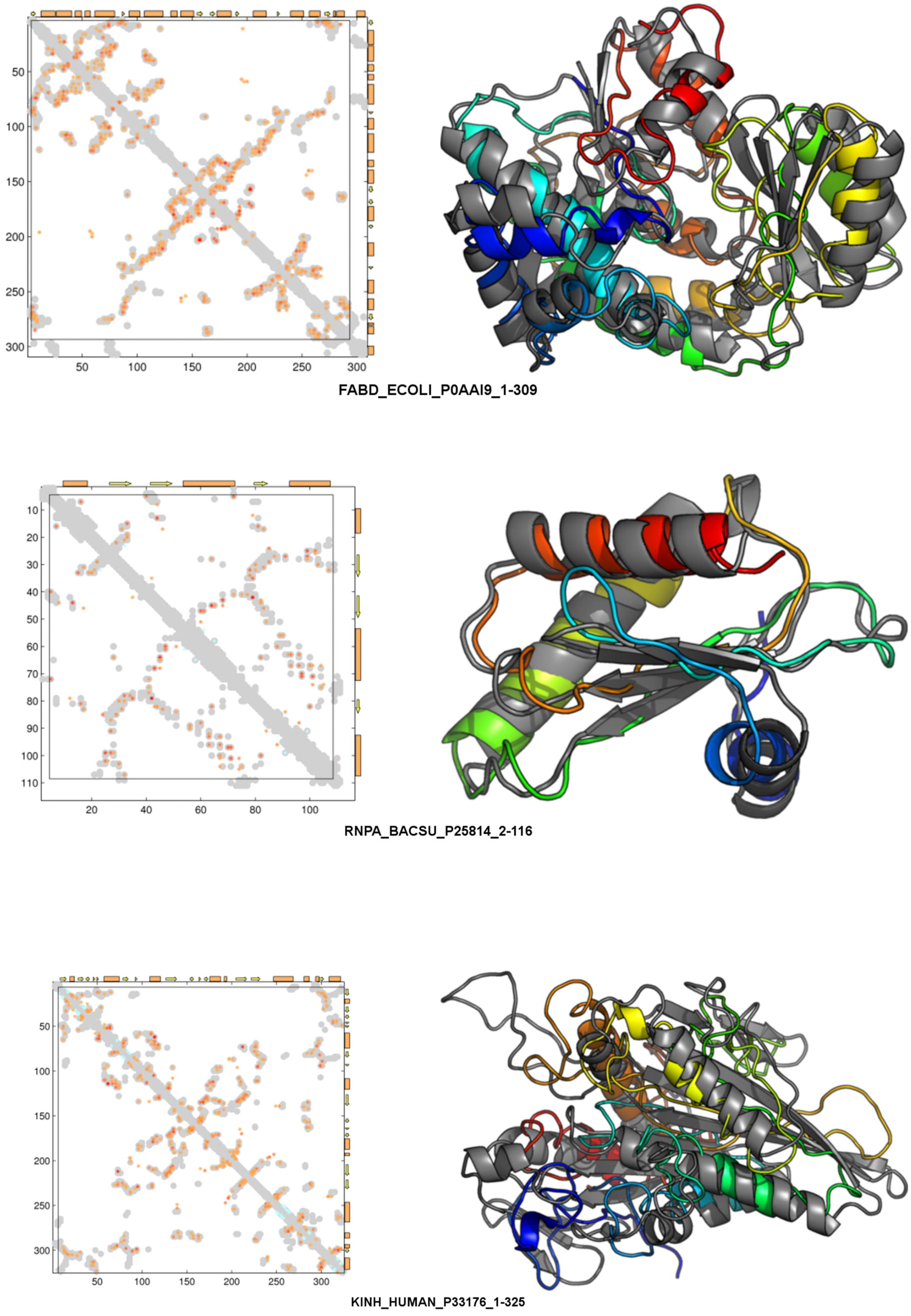


Figure 4: Well folded proteins. A selected subset of proteins in Table 1 with TM-score above 0.5 (best predicted by TM-score, sorted by Meff/L). Protein names (example): Uniprot Name_Species (CYP1_BRUMA), Uniprot ID (Q27450), residue range from-to (1-177). Left: Quality of contact prediction is higher the more predicted contacts (red to orange as EC value decreases) match the contacts derived from the experimental structure (grey). No experimental information is available in segments of the protein missing in the reference (PDB) structure (light blue ribbons). Contact patterns parallel or antiparallel to the diagonal are contacts between secondary structure helices (orange rectangles) or beta strands (arrows). Right: predicted 3D structure (rainbow-colored cartoon) superimposed against reference experimental structure (grey ribbon).

Well folded proteins (cont'd)



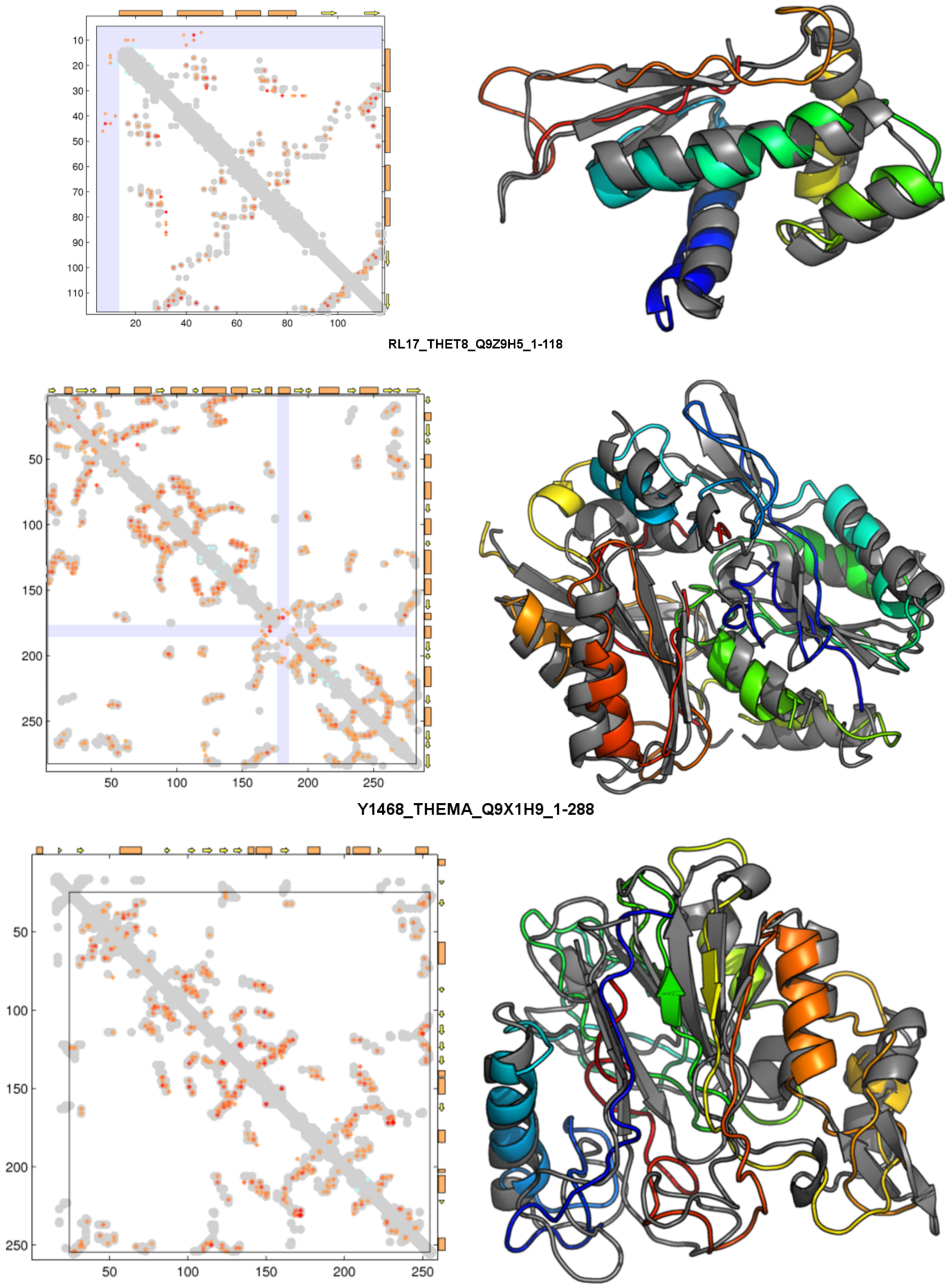
FABD_ECOLI_P0AAI9_1-309

RNPA_BACSU_P25814_2-116

KINH_HUMAN_P33176_1-325

Figure 5: Well folded proteins (cont'd). Coloring as above.

Well folded proteins (cont'd)



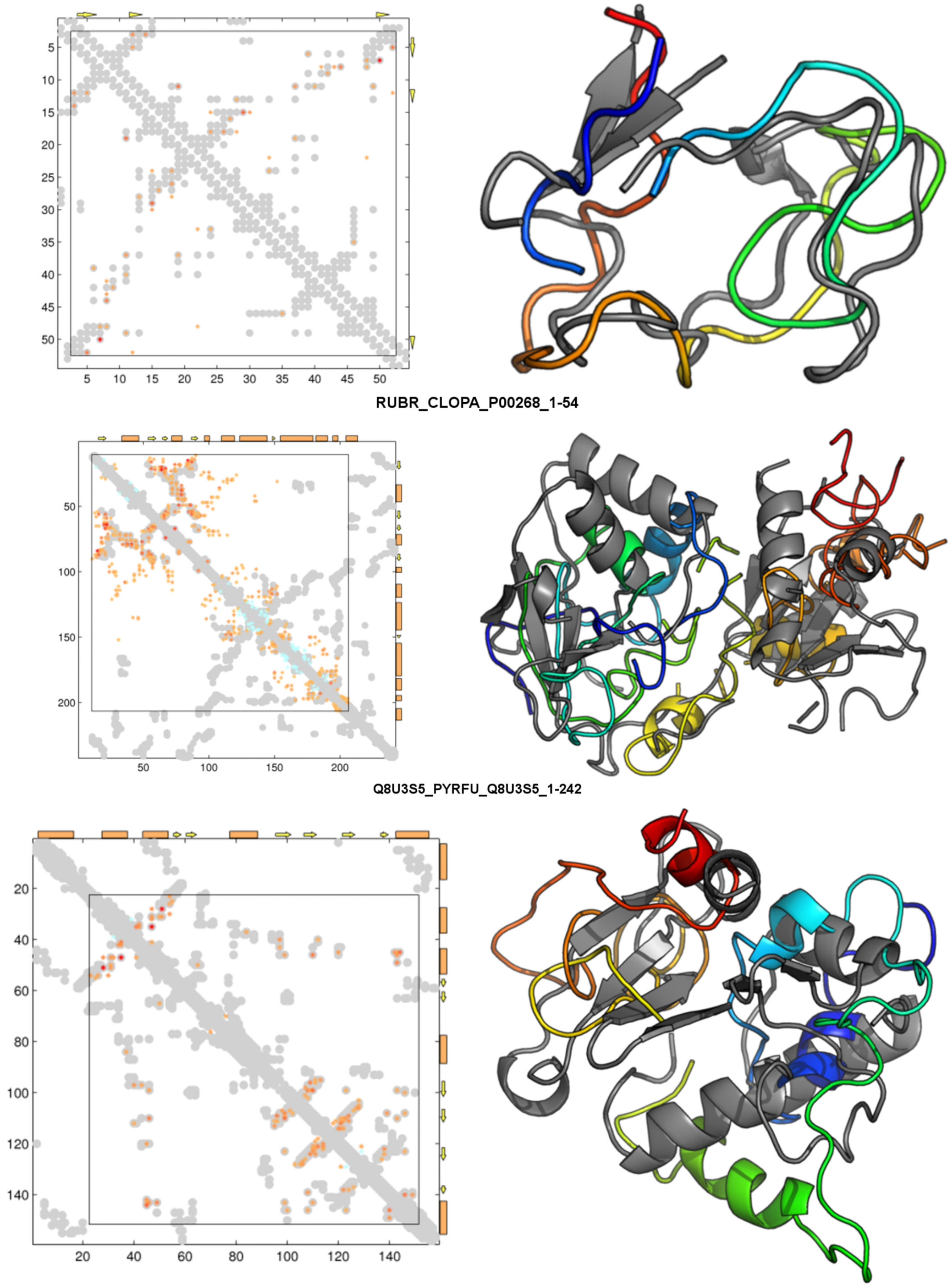
RL17_THET8_Q9Z9H5_1-118

Y1468_THEMA_Q9X1H9_1-288

Q9K0A8_NEIMB_Q9K0A8_1-259

Figure 6: Well folded proteins (cont'd). Coloring as above.

Less well or badly folded proteins



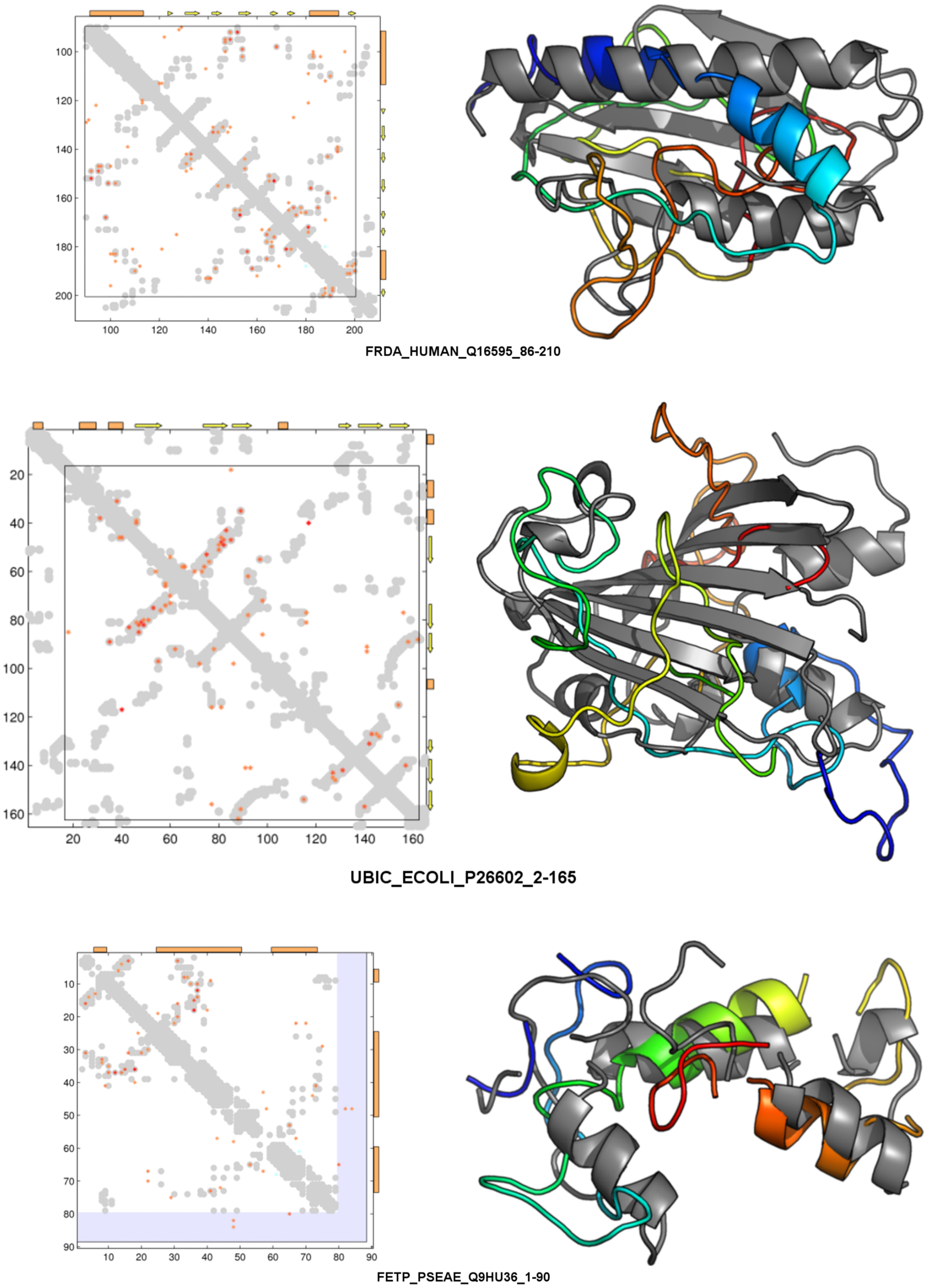
RUBR_CLOPA_P00268_1-54

Q8U3S5_PYRFU_Q8U3S5_1-242

Q926X2_LISIN_Q926X2_1-159

Figure 7: Less well or badly folded proteins. A selected subset of proteins in Table 1 with TM-score below 0.5 (least well predicted by TM-score, sorted by Meff/L). Graphical representation as in previous figure.

Less well or badly folded proteins (cont'd)



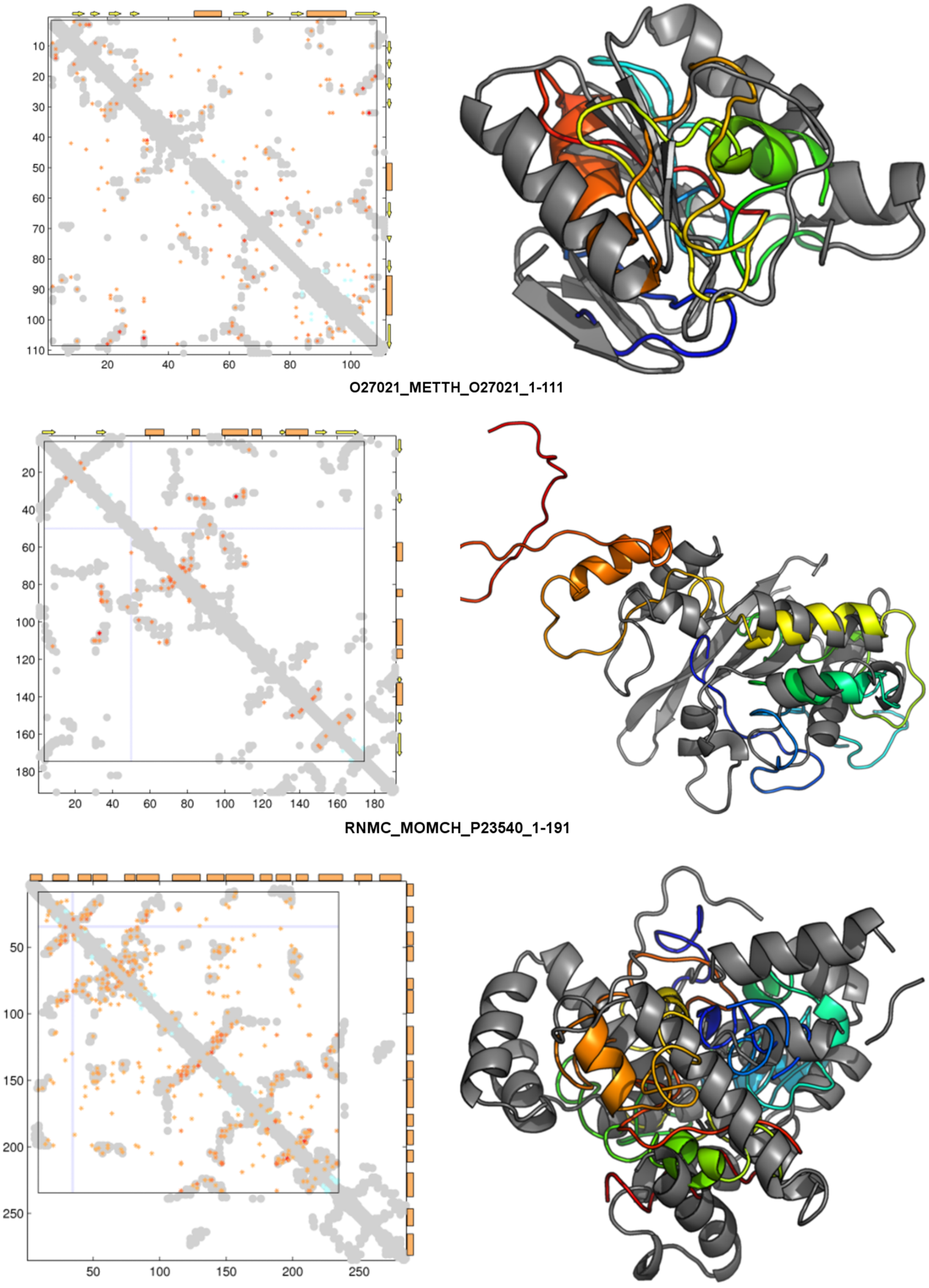
FRDA_HUMAN_Q16595_86-210

UBIC_ECOLI_P26602_2-165

FETP_PSEAE_Q9HU36_1-90

Figure 8: Less well or badly folded proteins (cont'd). Coloring as above.

Less well or badly folded proteins (cont'd)



O27021_METTH_O27021_1-111

RNMC_MOMCH_P23540_1-191

Q0PBQ7_CAMJE_Q0PBQ7_1-285

Figure 9: Less well or badly folded proteins (cont'd). Coloring as above.