SAM-T08, HMM-based Protein Structure Prediction

Kevin Karplus

April 20, 2009

Abstract

The SAM-T08 web server is a protein-structure prediction server that provides several useful intermediate results in addition to the final predicted 3D structure: three multiple sequence alignments of putative homologs using different iterated search procedures, prediction of local structure features including various backbone and burial properties, calibrated E-values for the significance of template searches of PDB, and residue-residue contact predictions.

The server has been validated as part of the CASP8 assessment of structure prediction as having good performance across all classes of predictions.

`http://compbio.soe.ucsc.edu/SAM_T08/T08_query.html`

Keywords: protein structure prediction, secondary structure, burial, residue-residue contact prediction, CASP8, SAM_T08, neural network

Contact:Kevin Karplus
Biomolecular Engineering
Baskin School of Engineering
University of California
Santa Cruz, CA 95064
karplus@soe.ucsc.edu
phone: 1-831-459-4250

1 Structure prediction server

The SAM-T08 web server is a protein-structure prediction server, the latest in a series of servers that started in 1999 with SAM-T99 [13, 6, 14, 15, 16]. The input to the server is an amino-acid sequence in FASTA format (limited to $\leq$700 residues), and the primary output is a 3D model in PDB format. In addition to providing 3D models, the SAM-T08 web site provides a large number of intermediate results, which are often interesting in their own right: multiple sequence alignments of putative homologs, prediction of local structure features, lists of potential templates of known structure, alignments to templates, and residue-residue contact predictions.

The example sequence used in this paper, and provided by the server if the user does not supply one, is T0437, one of the CASP8 prediction targets. An ensemble of NMR structures for T0437 is now available in PDB file 2k3i [17]. The figures in this paper are taken from our CASP8 prediction made 6 June 2008, before the NMR structures were released. Full details of the prediction can be found at `http://www.soe.ucsc.edu/~karplus/casp8/T0437/decoys/SAM_T08/`

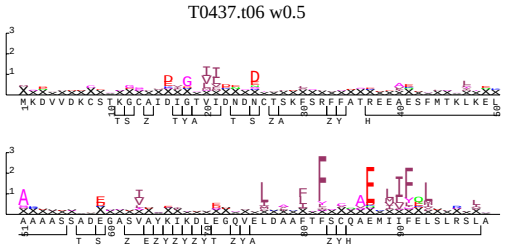
2 Multiple Sequence Alignments and Sequence Logos

Before starting to make multiple sequence alignments and hidden Markov models, the web site first does a quick blastp search of a non-redundant version of the PDB dataset (downloaded weekly from Dunbrack's PISCES server) [1, 25]. This search is not used in subsequent steps, but can be useful for determining whether there are any very close templates and whether those templates are subsequently used in the model building.

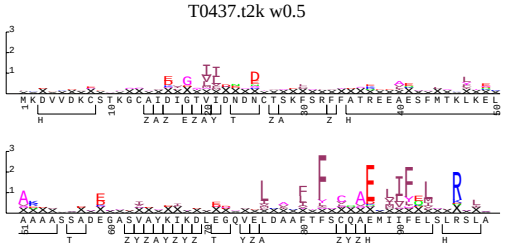
The process proper starts by doing three different iterated searches to find and align putative homologs from NR, NCBI's non-redundant database of protein sequences [19]. The first search, T06 from CASP7 in 2006, is the most sensitive, but can become contaminated with unrelated sequences about 0.5% of the time. The next search, T04 from CASP6 in 2004, is slightly less sensitive, but has about the same probability of contamination. The T04 and T06 searches use similar iterations and usually produce similar results, but occasionally come up with different alignments or different sets of homologs, due to differences in parameter settings. The T2K search, from CASP4 in 2000, is the least sensitive, and so includes mainly closely related homologs. The less sensitive search is often useful for help in choosing templates when there are many homologous proteins of known structure.

The multiple sequence alignments (MSAs) are provided in machine-readable format (A2M [9]), and in a somewhat more human-readable HTML format.[1] Because there are often over 20,000 sequences in the multiple alignment, trying to view the alignments in traditional ways is often not very illuminating. To alleviate this problem, the server provides sequence logos for the alignments, where the height of each bar indicates how conserved the residues are and the letters in the bar give the probability distribution for the amino acids at that position. The

[1] We use NCBI Entrez Utilities to retrieve taxonomy information about the sequences when making the HTML files. Because the XML files we retrieve are truncated by Entrez Utilities when they get too long, crashing the standard perl XML parser we are using to read them, our HTML files are sometimes not created when too many sequences are found. This is the most obvious known bug in the server.



(a) Sequence logo for SAM_T06 multiple sequence alignment, based on 119 sequences. The SAM_T04 alignment is nearly identical. The sequence logo shows which residues are most highly conserved in this family of proteins. The groups of conserved residues are typical of motifs that are preserved through evolution.



(b) Sequence logo for SAM_T2K multiple sequence alignment, based on 99 sequences. Note that R96 is conserved in the narrower set of homologs found by T2K, but is not conserved in the T06 alignment.

Figure 1: Sequence logos from multiple-sequence alignments for CASP8 target T0437. In both sequence logos, the residues of the target sequence are consistent with the conserved residues in the multiple sequence alignment, indicating that the iterated searches were not contaminated by unrelated proteins.

pattern of conserved residues is often of use for making conjectures about function and binding sites, even when no confident tertiary structure prediction can be made.

All three searches are provided separately, so that the sequence logos can be examined for contamination and results checked for consistency. The searches are combined later in the process.

3 Local Structure Prediction

After the iterated searches, the MSAs are used as inputs to neural networks that predict various local structure properties: twelve backbone structure alphabets and three burial structure alphabets. The twelve backbone alphabets are str4, str2, alpha, bys, pb, n_notor, n_notor2, n_sep, o_notor, o_notor2, o_sep, and dssp_ehl2. Many of these alphabets have been described previously [12, 11, 2], but some are new and are so far described in detail only on the Frequently Asked Questions (FAQ) page for the web site. The most familiar are the dssp_ehl2 alpha-

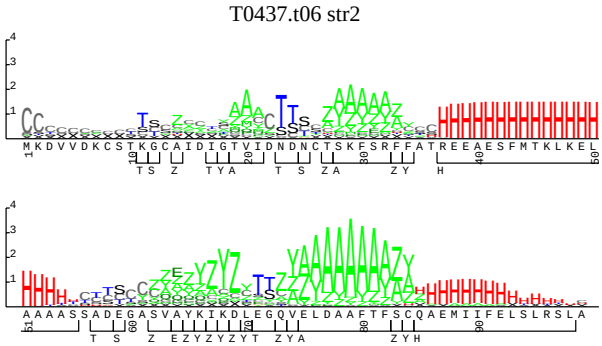


Figure 2: Sequence logo for the prediction of the str2 backbone alphabet based on the T06 multiple sequence alignment. The strong predictions here are generally accurate, but the NMR models did not include the predicted hairpin turn around D24—the region before C26 had no structure in the NMR models. The predictions are often this good for simple globular proteins, but can be thrown off by metal-binding sites and disulfide bridges, which are common in very small proteins.

bet, which has just three letters (E for beta strands and bridges, H for helices, and L for everything else), which is a reduction of the DSSP alphabet [10]. The str2 alphabet, which has been our most valuable backbone alphabet, is an extension of DSSP to distinguish between different types of beta-strands (Figure 2). The str4 alphabet is an attempt to use different ways of classifying loop residues and strand residues, but turned out to be somewhat less useful than str2. The alpha alphabet classifies residues according to their C_α-C_α-C_α-$C_\alpha$torsion angles, the bys alphabet is a classification of residues by ϕ and ψ angles by Bystroff [4], and the pb alphabet is de Brevern's protein blocks [5].

The notor and sep alphabets classify residues according to the hydrogen bond at the N or O atom. The notor and notor2 alphabets classify the Hbonds according to the $C_{i-1} - N_i - O_j - N_{j+1}$ torsion angle for donor N_i and acceptor O_j with special cases for alpha helices ($i = j + 4$) and 3-10 helices and turns ($i = j + 3$). The notor2 alphabets have a few more special cases for $i = j + 5$ and common multiple hydrogen-bond patterns. The sep alphabets classify the Hbonds according to the separation $i - j$.

The three burial alphabets predict the number of C_β atoms within 14 Ångstroms (7 classes), the number of C_β atoms within 8 Ångstroms at least 9 residues apart along the chain (14 classes), and a somewhat more complicated count of nearby residues (near-backbone-11 [2], 11 classes). The near-backbone-11 measure has been the most useful of these burial predictions (Figure 3). The burial alphabets are organized so that "A" is the least buried class with increasing burial as the letters go

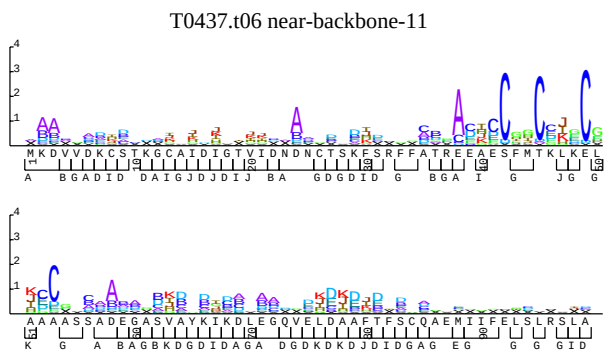


Figure 3: Sequence logo for the prediction of near-backbone-11 burial based on the T06 multiple sequence alignment. The residues with A, B, or C are predicted to be highly exposed, while those with H, I, or J are predicted to be highly buried. The burial predictions are excellent for this target.

through the alphabet.

For each MSA and each structure alphabet, several outputs are provided: a table of the probability vector over the alphabet for each position in the sequence; a sequence logo summarizing the probability vectors, showing the prediction and strength of prediction at each position of the sequence (Figures 2 and 3); and a summary sequence giving the most probable letter at each position. For users wanting a quick approximate view of the local structure prediction, a consensus prediction of the three-state alphabet (E=strand, H=helix, L=loop) is provided. To aid in viewing the local structure predictions in the final tertiary prediction, rasmol scripts for coloring the model according to the predicted local structure are provided.

4 Template selection and alignment

Our templates come mainly from our *template library*, a large representative subset of PDB chains, for which we have precomputed a set of hidden Markov models (HMMs). We have separate template libraries for the different iterated search methods. As of 28 Jan 2008, the template libraries contained 19621, 17732, and 15967 chains for T2K, T04, and T06, respectively, while a non-redundant PDB set contained 36643 chains.

After the local structure predictions are done, the SAM (Sequence Alignment and Modeling) tool suite [8] is used to build HMMs from the MSAs and predicted local structures. The HMMs are used to search PDB for potential templates for structure prediction. HMMs in the template library are used to score the target sequence, and all the resulting scores are merged into a best-scores-all.html file that summarizes the best hits, sorted by E-values. The table also includes links to the PDB [3] and Proteopedia [7] web sites for each template, as well as links to the Structural Classification of Proteins (SCOP [18]) website, when available.

The E-values are moderately well calibrated (off by no more than a factor of 10 in cross-validation tests, unpublished), so that E-values less than 0.01 indicate that a good structural template is available for at least part of the target protein and E-values greater than 1 indicate that the method will be using mainly ab-initio and fragment methods to generate the structure, and that the tertiary structure is thus much less reliable.

It is important for users to check the E-values, as the method always produces a full-length model, even when no good template is available. For T0437, the target 2jz5A has an E-value of 8.6e-12, indicating a very confident similarity. Even the initial blastp over PDB finds this template, though the E-value is only about 0.1, so the confidence is not as high.

For longer multi-domain proteins, the server may have a good template-based model for one domain, and poor, ab-initio models for the others. In those cases it is often wise to split the target into separate domains and predict them separately. The server does not do this automatically.

For each of the top templates, the server provides several alignments between the target and the template, which are used in subsequent tertiary prediction, but which could also be useful for transferring information (such as binding-site residues) from the templates to the target. Of the various alignments, the t06-local-str2+near-backbone-11-0.8+0.6+0.8-adpstyle5 alignments are generally the most reliable. These are constructed by local alignment to a 3-track HMM that has an amino acid profile, str2 predictions, and near-backbone-11 predictions as the three tracks, with track weights 0.8, 0.6, and 0.8 respectively, and using posterior decoding aligment (SAM parameter adpstyle=5). Although there are often better alignments in the pool, they do not come consistently from the same method.

Crude models are generated from the top alignments, and superimposed in a pdb file. The undertaker-align.pdb.gz file can be viewed with any molecular modeling software to see what parts of the protein are coming from the templates and whether the templates agree on the structure of that portion of the protein.

After the major alignments have been made, short gapless alignments (fragment lists) are made to provide reasonable local structures for building the final model.

5 Contact Prediction

We have two distinct ways of predicting what residues may be in contact: ab-initio contact prediction using neural networks and information about correlated mutations in the MSAs [23], and distance constraints are extracted

from the best alignments, for use in constraining the tertiary structure prediction [20].

The neural network predictions are most useful when there are no templates found with E-value $\leq$ 1. The server presents three different neural network predictions. The 647_47 prediction is the network validated at CASP7 [23]. The 730_47 prediction does not use any paired-column statistics, but just local structure prediction at the individual residues. The 648_17.730_47 prediction is a two-stage one that filters the 730_47 predictions using paired-column statistics. In our testing, the two-stage method works best when the T06 MSA has enough diversity of sequences for correlated mutations to be detected. For ORFans and target sequences for which only very similar sequences are aligned in the T06 MSA, the 730_47 predictions are somewhat better (unpublished).

The constraints extracted from alignments are most useful when templates are found with low E-values, as the constraints are used in model generation for selecting templates and to keep the models from drifting too far from the templates. The constraints have also been used for model quality assessment in evaluating models from other servers [20, 2], but that application is not provided by the web server.

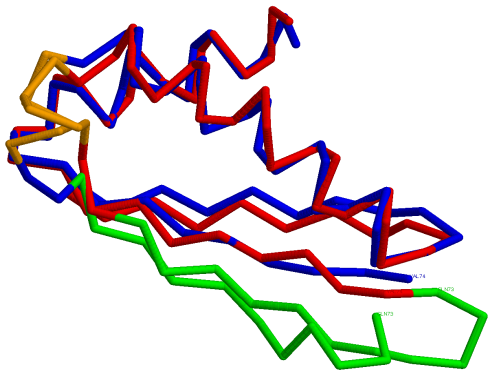


Figure 4: Predicted model in blue and NMR structure in red superimposed, Only residues C26-L98 are shown, since the NMR models had no structure before C26. Residues V63-Q73 (shown in green) are misaligned in the prediction, resulting in the gap before V74, instead of the proper hairpin. Residues S56-G60 (shown in orange) were not evaluated at CASP, because the ensemble of NMR models had quite different structures for those 5 residues. The picture was generated by the rasmol molecular viewing software [22]. The SAM-T08-human prediction was substantially better than the SAM-T08-server prediction, but it was based on Zhang-Server_TS5, the second-best server model (Zhang-Server_TS2 was the best server model in our evaluation).

6 Model generation

Finally, the undertaker program is run to generate an all-atom model using the templates, the local structure predictions, the distance constraints, and the contact predictions. The prediction for our example is compared with an NMR structure in Figure 4.

For compatibility with the CASP experiment, 5 models are produced: model1 is the polished model output from undertaker, model2 is the initial model after undertaker examines the templates but before attempting to remove clashes or gaps, and models 3 through 5 are incomplete models based on simple side-chain replacement on the top 3 templates. A few other models are available in the "decoys" subdirectory, including intermediate models in the optimization process, and models which have had the sidechains repacked by rosetta [21], or which have had energy minimization by gromacs [24], though neither of these programs is used for the primary output.

All the results, including all intermediate files, are kept available on the web site for at least one week, and can be downloaded as a gzipped tarball from a link at the bottom of the page.

7 Validation at CASP8

The SAM-T08 server was validated as part of the CASP8 protein-structure prediction experiment in summer 2008. CASP (Critical Assessment of Structure Prediction) is a community-wide experiment held every two years. Predictors are given the sequences of proteins whose structures are being solved, but whose structures have not been publicly released, and are required to register their predictions within 3 days for servers, or 3 weeks for human-assisted methods. The predicted models are compared with the experimental models to determine which methods are really working.

Detailed results of the testing can be found on the CASP8 web page http://predictioncenter.org/casp8/results.cgi as well as on several unofficial evaluation sites (list available at http://www.reading.ac.uk/bioinf/CASP8).

Several different metrics have been used to evaluate the quality of predictions, and rankings of servers depend heavily on which metrics are used, what set of targets are compared, and whether whole-chain comparisons or domain-based comparisons are made.

Although the SAM-T08 server was not the best server on the commonly used metrics that measure just the positions of the C_α atoms (GDT_TS and TM-score, for example), it did quite well (ranking 2 through 21 out of 70 servers overall, depending on the evaluation used).

In Zhang's ranking of the servers by TM-score of domains on the hard targets, (http://zhang.bioinformatics.ku.edu/casp8/13D.html), the SAM-T08-server ranks third, after Zhang-Server and BAKER-ROBETTA, while on the easy targets

(where differences are smaller and many servers produce almost identical models), SAM-T08-server ranks 21st. If hydrogen bond scoring is included, SAM-T08-server moves to 2nd place overall, and 4th on the easy targets.

Using a contact-based measure, Nick Grishin ranked SAM-T08-server 5th or 6th on all targets (http://prodata.swmed.edu/CASP8/evaluation/DomainsAll.First.html).

In the official evaluations, the SAM-T08 models were seen to have unusually good stereochemistry for homology models, even though the C_α traces were not the best (based on assessor's presentations at CASP8 conference, not published yet). On the common backbone accuracy measures, the SAM-T08-server ranked 9th through 14th among servers (http://predictioncenter.org/casp8/groups_analysis.cgi), except on the "high-accuracy" server targets, where it was in the middle of the pack (31st out of 70).

The SAM-T08 server generally ranked less well on the very easy targets (where most of the methods produced almost indistinguishable results) and better on the harder targets. Performance relative to other servers seemed to peak for those targets that had templates available, but for which finding and aligning the template was difficult, as we have focused our efforts most on fold recognition and alignment.

The SAM-T08 server uses the same protocol for all targets, whether they have highly similar templates available or not, but the method is tuned for the difficult targets, rather than the easy ones.

With a few notable exceptions (such as target T0442 domain 2), the SAM-T08 server did substantially better than the older SAM-T06 server in all evaluations. The older SAM-T02 server does not produced models, just alignments, and had substantially poorer performance than either of the more recent servers. The selection of templates and alignments by the hidden Markov models has not improved substantially—the models built directly from the top alignment: SAM-T08-server_TS3, SAM-T06-server_TS2, and SAM-T02-server_AL1 are of variable quality, but not showing consistent improvement. The selection and optimization of models by undertaker, however, is showing substantial improvement from SAM-T02 to SAM-T06 to SAM-T08.

Acknowledgments

Over the years dozens of people have contributed to the tools of the SAM-T08 servers. Some of the more notable contributors (in alphabetical order) include John Archie, Bret Barnes, Christian Barrett, Sugato Basu, Jonathan Casper, Melissa Cline, Mark Diekhans, Chris Dragon, Birong Hu, Richard Hughey, Rachel Karchin, Sol Katzman, Firas Khatib, Anders Krogh, Martin Madera, Yael Mandel-Gutfreund, Martin Paluszewski, George Shackelford, Kimmen Sjölander, Don Speck, Grant Thiltgen, and Spencer Tu.

Comments on drafts of the paper by John Archie, Richard Hughey, Thomas Juettemann Josue Samayoa, and Chirag Sharma, are particularly appreciated.

This work was supported in part by NIH grant 1 R01 GM068570-01.

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] John Archie and Kevin Karplus. Applying undertaker cost functions to model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, 75(3):550–555, 2009. published online 30 Sep 2008.

[3] F.C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, November 1977.

[4] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, 301(1):173–190, August 2000.

[5] A.G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function and Genetics*, 41:271–287, November 2000.

[6] Daniel Fischer, Christian Barrett, Kevin Bryson, Arne Elofsson, Adam Godzik, David Jones, Kevin Karplus, Lawrence A. Kelley, Robert M. MacCallum, Krzysztof Pawłowski, Burkhard Rost, Leszek Rychlewski, and Michael Sternberg. CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Genetics*, Supplement 3(1):209–217, 1999.

[7] Eran Hodis, Jaime Prilusky, Eric Martz, Israel Silman, John Moult, and Joel L Sussman. Proteopedia—a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. *Genome Biology*, 9:R121, 2008. doi:10.1186/gb-2008-9-8-r121.

[8] Richard Hughey, Kevin Karplus, and Anders Krogh. SAM: Sequence alignment and modeling software system, version 3. Technical Report UCSC-CRL-99-11, University of California, Santa Cruz, Computer Engineering, UC Santa Cruz, CA 95064, October 1999. Available from `http://www.soe.ucsc.edu/research/compbio/sam.html`.

[9] Richard Hughey and Anders Krogh. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Computer Applications in the Biosciences*, 12(2):95–107, 1996. Information on obtaining SAM is available at `http://www.soe.ucsc.edu/research/compbio/sam.html`.

[10] Wolfgang Kabsch and Chris Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[11] Rachel Karchin, Melissa Cline, and Kevin Karplus. Evaluation of local structure alphabets based on residue burial. *Proteins: Structure, Function, and Genetics*, 55(3):508–518, 5 March 2004. doi:10.1002/prot.20008.

[12] Rachel Karchin, Melissa Cline, Yael Mandel-Gutfreund, and Kevin Karplus. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins: Structure, Function, and Genetics*, 51(4):504–514, June 2003.

[13] Kevin Karplus, Christian Barrett, Melissa Cline, Mark Diekhans, Leslie Grate, and Richard Hughey. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Genetics*, Supplement 3:121–125, 1999.

[14] Kevin Karplus, Rachel Karchin, Christian Barrett, Spencer Tu, Melissa Cline, Mark Diekhans, Leslie Grate, Jonathan Casper, and Richard Hughey. What is the value added by human intervention in protein structure prediction? *Proteins: Structure, Function, and Genetics*, 45(S5):86–91, 2001.

[15] Kevin Karplus, Rachel Karchin, Jenny Draper, Jonathan Casper, Yael Mandel-Gutfreund, Mark Diekhans, and Richard Hughey. Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins: Structure, Function, and Genetics*, 53(S6):491–496, 15 October 2003.

[16] Kevin Karplus, Sol Katzman, George Shackelford, Martina Koeva, Jenny Draper, Bret Barnes, Marcia Soriano, and Richard Hughey. SAM-T04: what's new in protein-structure prediction for CASP6. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):135–142, September 2005.

[17] J.L. Mills, K.K. Singarapu, A. Eletski, D.K. Sukumaran, D. Wang, M. Jiang, C. Ciccosanti, R. Xiao, J. Liu, M.C. Baran, G.V.T. Swapna, T.B. Acton, B. Rost, G.T. Montelione, and T. Szyperski. Solution NMR structure of protein yiiS from *Shigella flexneri*. Northeast Structural Genomics Consortium target SfR90. PDB file doi:10.2210/pdb2k3i/pdb, 2008.

[18] A. G. Murzin. Structural classification of proteins: new superfamilies. *Current Opinion in Structural Biology*, 6(3):386–94, June 1996.

[19] Non-redundant Protein Database. Distributed by anonymous FTP from ftp.ncbi.nih.gov/blast/db/., 2008.

[20] Martin Paluszewski and Kevin Karplus. Model quality assessment using distance constraints from alignments. *Proteins: Structure, Function, and Bioinformatics*, 75(3):540–549, 2009. published online 20 Sep 2008.

[21] Carol A. Rohl, Charlie E. M. Strauss, Kira Misura, and David Baker. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383:66–93, 2004.

[22] Roger Sayle and E. James Milner-White. RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences*, 20(9):374–376, September 1995.

[23] George Shackelford and Kevin Karplus. Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):159–164, 2007. doi:10.1002/prot.21791.

[24] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: Fast, flexible and free. *Journal of Computational Chemistry*, 26:1701–1718, 2005.

[25] G. Wang and R. L. Dunbrack, Jr. PISCES: a protein sequence culling server. *Bioinformatics*, 19:1589–1591, 2003.