Cross-Link Analysis and Experiment Planning for Elucidation of Protein Structure

Xiaoduan Ye,[1] **Janusz M. Bujnicki,**[2]
Alan M. Friedman,[3] **Chris Bailey-Kellogg**[4]

Keywords: Protein structure prediction, protein-protein complexes, experiment design, cross-linking mass spectrometry, disulfide trapping, structural genomics.

Emerging high-throughput experimental techniques for the characterization of protein and complex structure yield noisy data with sparse information content, placing a significant burden on computation to predict, optimize, and interpret the information provided. One such experiment employs residue-specific chemical cross-linkers to confirm or select among proposed structural models by testing consistency of cross-linking data with respect to model geometry (Fig. 1). Recent applications include those by Young et al., using high-resolution mass spectroscopy alone to correctly discriminate threading models of fibroblast growth factor [7]; Scaloni et al., analyzing the binding mode of the calmodulin-mellitin complex [4]; and Sorgen et al., determining the arrangement of transmembrane helices in lac permease [6].

We have developed a mechanism for analyzing cross-linking information with respect to a set of models, predicting the ability of experiments to discriminate among those models, and optimizing a set of experiments accordingly. A probabilistic framework selects models based on consistency with data (Fig. 1(c)), mediated by the geometric feasibilities of the cross-links for the models [3], represented by "cross-link maps" (Fig. 1(b)), and the experimental conditions, represented with noise and capture rates. We formalize model discriminability in terms of differences in cross-link maps, and formulate experiment planning problems to select sets of experiments that maximize such differences, accounting for the key factors of discriminability, coverage, balance, ambiguity, and cost. We have developed a greedy algorithm, generalizing those for related SetCover problems [1], that effectively navigates the design space defined by these terms.

We are applying this mechanism in a study of the bacteriophage lambda Tfa chaperone protein, and have planned dicysteine mutations for model discrimination by disulfide formation. Fig. 2 summarizes our planning results on 103 Tfa models, including three high-quality threading models from our fold recognition meta-server [2] and 100 decoys from the *ab initio* folding program Rosetta [5]. We are currently carrying out a minimized, balanced, and least ambiguous set of six experiments to discriminate the three threading models with discriminability two.

Our mechanism is very general, and we plan to study additional applications not tested here, for example in the discrimination of protein-protein complexes, incorporating different types of experimental data, and planning combinatorial possibilities of cross-linkers and mutations. Our methods provide the experimenter with a valuable tool for understanding and optimizing cross-linking experiments.

[1]Department of Computer Sciences, Purdue University. ye@cs.purdue.edu
[2]International Institute of Molecular and Cell Biology, Warsaw, Poland. iamb@genesilico.pl
[3]Department of Biological Sciences, Purdue University. afried@bilbo.bio.purdue.edu
[4]Corresponding author. Department of Computer Sciences, Purdue University. 250 N. Univ. St., West Lafayette, IN 47907, USA. cbk@cs.purdue.edu

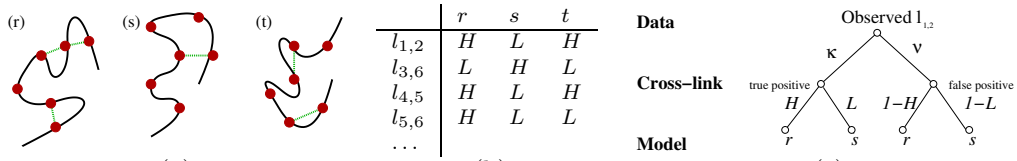

Figure 1: Model discrimination by cross-linking. (a) Different predicted models of a protein have different patterns of feasible cross-links (green dotted lines). (b) Cross-link maps represent feasibilities with conditional relationship for cross-links (rows) given models (columns), here shown as either high (H) or low (L). (c) Experimental identification of a cross-link $l_{1,2}$ provides evidence for and against models r and s, based on consistency with cross-link maps and modulated by the capture and noise rates of the experimental method (here uniformly κ and ν, respectively). Other terms arise from other cross-links, both observed and unobserved.

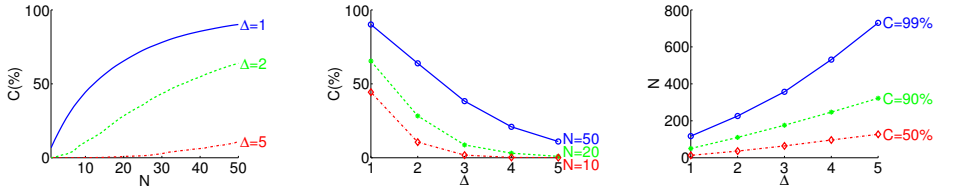

Figure 2: The relationship between coverage percentage C (%), discriminability Δ, and number of experiments N in disulfide experiments planned for 103 Tfa models. Pairs of parameters are varied, while the third is blocked at the indicated values. Our greedy plans are roughly within a factor of two of a simplistic lower bound (data not shown).

References

[1] D.S. Johnson. Approximation algorithms for combinatorial problems. *J Comput System Sci*, 9:256–278, 1974.

[2] M.A. Kurowski and J.M. Bujnicki. Genesilico protein structure prediction meta-server. *Nucleic Acids Res*, 31(13):3305–7, 2003. `http://genesilico.pl/meta`.

[3] S. Potluri, A.A. Khan, A. Kuzminykh, J.M. Bujnicki, A.M. Friedman, and C. Bailey-Kellogg. Geometric analysis of cross-linkability for protein fold discrimination. In *Pac Symp Biocomp*, pages 447–458, January 2004.

[4] A. Scaloni et al. Topology of the calmodulin-melittin complex. *J Mol Biol*, 277:945–958, 1998.

[5] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268:209–25, 1997.

[6] P.L. Sorgen, Y. Hu, L. Guan, H.R. Kaback, and M.E. Girvin. An approach to membrane protein structure without crystals. *PNAS*, 99(22):14037–14040, 2002.

[7] M.M. Young et al. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, 97:5802–5806, 2000.