On Sequence-Structure Alignment using Stochastic Grammars

Yasubumi Sakakibara

Inverse folding has been proposed as a practical and promising approach to computer-aided protein structure prediction. A simple method to accomplish inverse folding is structure prediction by homology. This method is performed by computing an alignment of two protein sequences based on measures of evolutionary distance. However, there exists some limitation on prediction accuracy of this method because evolutionary distance is only an indirect metric for structural similarity [Sander and Schneider, 1991]. Recently, a new method [Ouzounis et al., 1993], called sequence-structure alignment, has been proposed which deals with structural aspects directly for structure prediction. The metrics used for alignment are modified so that they take measures of structure conservation rather than mutation probability into account. If a family of protein sequences and structures is available, we have more advantage. In this case, then we propose stochastic grammars as statistical models for sequence-structure alignment problems. We have already shown that stochastic context-free grammars (SCFGs) are quite adequate for the problems of folding, aligning and modeling families of homologous RNA sequences [Sakakibara et al., 1994]. We can automatically learn SCFG parameters from unaligned, unfolded training sequences, and the distance measure defined by these SCFG parameters has the advantage that gap penalties and residue-residue contact preferences are specific to the position in the model. We compare our method based on stochastic grammars with previous proposed methods for sequence-structure alignment. The main purpose of this talk is to stimulate discussions in the workshop.

References

[Ouzounis et al., 1993] C. Ouzounis, C. Sander, M. Scharf and R. Schneider. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from three-dimensional structures. Journal of Molecular Biology, 232:805–825, 1993.

[Sakakibara et al., 1994] Y. Sakakibara, M. Brown, I. S. Mian, R. Underwood, and D. Haussler. Stochastic Context-Free Grammars for Modeling RNA. In Proceedings of 27th Hawaii International Conference on System Sciences (HICSS'94), (IEEE Computer Society Press, Los Alamitos, CA), Maui, Hawaii, 1994. Also to appear in NAR.

[Sander and Schneider, 1991] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. PROTEINS: Structure, Function, and Genetics, 9:56–68, 1991.