UNDERSTANDING AND PREDICTING PROTEIN STRUCTURE

Daniel Fischer
Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

Adam Godzik
Dept. of Molecular Biology, Scripps Research Institute, La Jolla, CA 92037, USA

Su Chung
Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA

S. Subbiah
The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104, USA

Richard Lathrop
Information and Computer Science, University of California, Irvine, CA, USA

Protein structure prediction from sequence remains a premiere computational problem for modern molecular biology. Just as protein structure prediction may be divided into sub-problems of main-chain and side-chain placement, so the protein structure prediction track this year has been divided into sub-tracks of protein threading (organized by Daniel Fischer and Adam Godzik) and side-chain packing (organized by Su Chung and S. Subbiah). The result is an unusually rich tour through different levels of protein structure prediction, from coarse-grained prediction of the tertiary fold to the fine-grained atomic detail of individual side-chains.

1 Protein Threading

Organized by Daniel Fischer and Adam Godzik.

All threading and folding algorithms depend crucially on the quality of the energy parameter set. For many years such sets were built by analyzing interaction regularities in known protein structures. Despite a long history of such derivations, fundamental problems with the theoretical background persist till today, with many existing derivations containg important omissions or even errors. Reva et al. tackle this problem by providing a very formal derivation protocol. While the practical difference their derivation makes can only be demonstrated by empirical tests in a predictive setting, their derivation provides a well thought out standard that other derivations can compare to.

In a second paper, Zheng et al. extend their original work presented last year in this conference. In their work they explore a new picture of a protein structure, described as a space filling aggregate of irregular tetrahedra

with vertices at C-alpha atoms. This model puts an emphasis on the spatial arrangement of residues, rather than on their position along the chain. In a natural way it also allows to develop a four body interaction potential. Although it still remains to be tested on actual threading, it is an interesting generalization of most other existing empirical interaction potentials used in threading and folding algorithms.

Delarue and Koehl return to the roots of the inverse folding problem, which addresses the question of finding a sequence compatible with a given protein structures. Using a mean field theory they develop a generalized sequence profile compatible with a given C-alpha atom trace. The sequence matrix, as it is called in the paper, is derived purely from the structural information, but can be used in the framework of standard sequence alignment programs.

Two papers in the threading session attempt to combine several prediction methods to create hybrid approaches addressing two different stages in protein structure prediction. Ortiz et al. analyze multiple aligned protein sequences to derive possible constraints on the final structure in terms in secondary structure elements and contacts between them. By combining state of the art secondary structure prediction, analysis of correlated mutations, and inverse folding, they derive a small number of distance constraints. These in turn are used to guide a lattice folding algorithm to arrive at a low resolution protein model. On three examples they show that, given a sufficient number of homologous protein sequences, it is possible to obtain final models with correct topology and 4-4.5 Angstrom root mean square deviation from the native structure.

Pawlowski et al. combine a semi-automatic comparative modelling and threading to build full atom models of proteins close to the limits of recognizable sequence similarity. For very distantly homologous proteins sequence alignments become unstable, changing dramatically with small changes in gap penalties and mutation matrices; recognition of a structurally correct alignment is impossible on the sequence level. At the same time, differences between alignments become more pronounced when full atoms models are built on their basis. Structurally correct alignments lead to higher quality models, as measured by the threading algorithm. This promising method is shown to work on two examples of notoriously difficult sequence alignments, and is also used to suggest a possible model for a currenlty unknown structure of S100A1 dimer. This paper also suggests an interesting approach to verifying the results of threading.

Mamitsuka and Abe use a stochastic tree grammar method to predict location and structure of beta sheet regions in proteins. Using a formalism developed to study natural languages, they are able to discover amino acid patterns defining localization of beta strands which go beyond simple sequence similar-

ity. This method stands half way between secondary and tertiary structure prediction method; it concentrates on predictions of secondary structure elements, but predicts their mutual structural organization within the beta sheet as well as their location along the sequence. On a set of several beta proteins this method is able to predict the correct position of most beta strands, and in two cases correctly suggests the entire protein topology.

2 Side-chain Packing, or "The Importance of Being Well-Packed"

Organized by Su Chung and S. Subbiah

Starting in the mid-70's and through the mid-80's there was much well-publicized literature by many groups that claimed to have solved the homology modeling problem to a significant degree. The general lay-biologist was led to think that, save for some long loops and solvent-exposed residues, the modeling of side-chains onto a template backbone from a homologue protein was in the main a solved problem. The combination of "sex appeal" from the advent of exciting color computer graphics and the "above-reproach" mathematically sophisticated, computationally complex energy calculations was a potent mix that could not be subjected to honest and critical assessment of (1) what was really being predicted as opposed to being merely copied and (2) what the prediction accuracy relative to random was.

In a typical modeling exercise, a protein sharing 50% sequence identity with a known homologue protein structure was modeled based on the homologue's main-chain providing the template scaffold on which the new side-chains were added. Since, roughly speaking, half the atoms in an average protein belong to either the main-chain or the C-beta side-chain location (which is in practice totally defined by the coordinates of the main chain atoms), at best only half the atoms — those in the post-C-beta side-chain positions — actually were predicted. Further, it is well-known that side-chains that are identical between homologous proteins very frequently assume the same or very similar conformation. It is also well-known that the backbone atoms between 2 proteins sharing 50% sequence identity typically differ by only about 1 Angstrom root mean square deviation (r.m.s). This is not a lot when compared with studies suggesting that when the same exact protein is solved in two different X-ray crystallography labs using different software, etc., the backbone atoms differ by as much as 0.5 Angstroms r.m.s. Thus for our typical modeling exercise where there is 50% sequence identity between target and template, simply accepting the template as a starting point for further modeling guarantees the "prediction" of half the atoms in the final completed model to about 1 Angstroms r.m.s. Moreover, the fact that 50% of the side-chain conformations can be sim-

ply adopted as is, "predicts" another one quarter of the atoms in the complete final model at significantly better than 1 Angstrom r.m.s. — possibly even 0.5 Angstrom r.m.s. Thus, barring the loops and highly solvent-exposed residues where modelers (then and now) a priori concede defeat, there was at best one quarter of all atoms that truly remained to be predicted. One could expect that even a random method might get some of these right by chance. Thus, no matter how badly this quarter of atoms were predicted, one could always count on the final complete predicted model being judged relatively accurate when compared to the known experimental answer. It also helped that there were no established numbers — in r.m.s. Angstroms, or equivalently chi angle degrees — on what constituted a totally random prediction. Any seemingly low r.m.s. error to report, accompanied by detail-obscuring superpositions of predicted model vs. known experimental truth that emphasized the general backbone similarity and ignored the details of the side-chain conformational similarity, could be trumpeted as success. The dictum — "Don't get it right; just get it in color" — was often an apt description of the state of affairs.

By 1989 some investigators began to point at the true difficulty of the problem [1,2]. Despite this dawn of pessimism, a little earlier in 1986 Ponder and Richards [3] pointed out that the exhaustive enumeration of the so-called side-chain rotamers could in principle allow side-chains to be predicted with great accuracy, using the perfect native backbone as a rigid constraint. However, in practice, this task of enumerating all the possibilities was well beyond the computing power of modern day computers. Assuming just 5 rotamer choices on average per side-chain, the typical 200 residue protein would require enumerating 5^{200} conformations. For an ab initio chi angle enumeration at 10 degree intervals the number would be even more astronomical at more than 10^{600}.

In 1991, 3 groups working independently of each other — one using an ab initio approach and two using rotamer-based ones — proved that the foresight of Ponder and Richards was indeed correct [4,6]. When a protein is stripped of all its side-chains and the perfect native backbone is used as a constraint to re-pack all the side-chain atoms, these varied methods could depend on achieving 1.25, 1.5 and 1.6 Angstroms r.m.s. accuracy on the 30-40% of the least solvent-exposed residues. Using only extremely simple and crude van der Waals packing-oriented energy functions without electrostatics or solvent considerations, these methods circumvented the astronomical combinatorics of exhaustive enumeration. At this time the r.m.s. deviation averaged over all side-chain types that could be expected for a random prediction was established to be about 3.1 to 3.3 Angstroms, which has since been confirmed [4,7]. Since then there has been a flood of different methods and improvements of earlier

ones — mostly using rotamers or database-derived information — that have confirmed our ability to crack this seemingly large combinatorial problem at even higher accuracies of 1 and sub-1 Angstrom r.m.s. for the buried side-chain residues. It is worth noting that when two X-ray labs solve the structure of the same protein, side-chain r.m.s. deviations, albeit over all side-chains and not just buried ones, can differ by as much as 1 to 1.5 Angstroms r.m.s.

Despite the success at breaking the back of the basic side-chain packing combinatorial problem, many workers, as illustrated by the contributions from Koehl and Delarue and from Desmet et al. in the following pages, are busy trying to refine the methodologies further. More recently, other workers have extended these methods to the practically more important real-life homology modeling situation; again using a rigid-backbone, but one that is a less accurate template from the homologue protein. While these studies show that even imperfect backbones can be used to produce side-chain information that is better than random, it is increasingly clear that the backbone needs to be simultaneously adjusted. While this problem has been solved recently in a very specialized instance[8], a general solution still is lacking. This problem, the problem of handling side-chain packing at loops, and the problem of allowing for solvent and electrostatic effects, all are beginning to take center-stage in theoretical side-chain packing.

In parallel with the success at defeating the combinatorics of the packing problem, methods relying on this success have been proposed to predict the packing energies and relative stabilities of mutants. In the following pages both Kono and co-workers and Lee and Levitt compare their predictions with experiments. Related algorithms that allow side-chain flexibility during protein-ligand docking are also under study.

It is clear, that in the last few years we have convincingly shown that the combinatorial problem of side-chain packing is reliably solved. This theoretical success, viewed in the context of supporting experimental evidence, suggests that to first order good side-chain packing is an important, if not the dominant, factor in the architecture of proteins. Many hurdles remain to turn this discovery into truly practical tools to aid the experimental biologist. Nevertheless this appears an opportune moment to take stock of our theoretical successes and chart the way forward that will allow predictive success to migrate from the computer to the test-tube. Recalcitrant backbone shifts, wayward loops, slippery energy potentials, troublesome electrostatic effects and even the mystical medium of water - Let's get it all right and in living color.

Acknowledgments

The session co-chairs gratefully thank the session referees, whose careful reviews of the submitted papers and insightful, judicious suggestions for improvement are materially reflected in the high quality of the presented papers: N. Alexandrov, M. Bass, G. Crippen, S. Doniach, R. Dunbrack, A. Elofson, J. Fetrow, L. Gregoret, M. Grisbkov, J. Hirst, A. Kolinski, S. LeGrand, K. Olszewski, B. Park, M. Pellegrini, F. Pettit, D. Rice, D. Ripoll, E. Shakhnovich, J. Skolnick, A. Tempczyk, M. Vasquez, M. Wilmanns, T-M. Yi, R. Zimmer. Special thanks to all crystallographers who deposited their coordinates in the international scientific databases.

This work was supported in part by a grant to S.S. from the Department of Energy, DE-FG03-95ER62135. We also acknowledge financial and other support for the session from the Bioinformatics Center at the National University of Singapore, The Program in Mathematical Biology at University of California, Berkeley, Molecular Applications Group of Palo Alto and Digital Equipment Corporation.

References

1. L. Reid and J. Thornton, Proteins 5, 170 1989.
2. N. L. Summers and M. Karplus, J. Mol. Biol. 210, 785 (1989).
3. J. W. Ponder and F. M. Richards J. Mol. Biol. 193, 775 (1987).
4. C. Lee and S. Subbiah, J. Mol. Biol. 217, 373 (1991).
5. P. Tuffery et. al., J. Biomol. Struct. Dynam. 8, 1267 (1991).
6. L. Holm and C. Sander, J. Mol. Biol. 218, 183 (1991).
7. R. Tanimura, A. Kidera and H. Nakamura, Prot. Sci. 3, 2358 (1994)..
8. P. Harbury, B. Tidor and P. Kim, Proc. Natl. Acad. Sci USA 92, 8408 (1995).