Progress at Last

Ron Elber[1,*]
[1]Department of Chemistry and Biochemistry, Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX 78712, USA
*Correspondence: ron@ices.utexas.edu
DOI 10.1016/j.str.2011.11.005

Refinement of protein structures from a correct topology to atomically detailed resolution has proven remarkably difficult. Jian et al. (in this issue of *Structure*) illustrate a significant advance in this task by carefully incorporating into the refinement process many body interactions extracted from fragment statistics.

The article by Jian et al. (2011) in this issue of *Structure* starts with the statement, ''Template based modeling represents the most accurate method in protein structure prediction.'' This statement is correct, but ''most accurate'' is a relative term and it is not clear if the accuracy achieved by the ''most accurate method'' is sufficient. A desired level of accuracy is that of atomic resolution (better than 1Å), a level that is still out of reach if we start from a structure of about 4Å RMSD from the native coordinates. Template based modeling is a two-step process. In the first step, an experimental structure is selected, which we believe is a good starting point to create a model for the target sequence. In the second step, the model is adjusted (refined) to obtain a better accuracy. It is expected that the refinement step will include only minor changes to the overall fold. Most of the time, templates are at 4Å–5Å from the native fold. The task of the second step is to improve the resolution of the model from the typical difference of 4Å–5Å between the initial template and the true native structure to the 1Å of atomic resolution. This task has proven elusive. Given the significant progress and documented successes in the first step of template based modeling, identifying starting structures, it is surprising that a second step with a significant impact on the quality of the structure has been hard to achieve. What is the problem?

Refinement is difficult because well-established procedures that are used to identify templates do not work as effectively in the second step. Statistical learning of experimentally determined protein structures is clearly the way to identify templates. Log-odd ratios and inequality constraints were used to learn contact interactions and local biases from the Protein Data Bank. These learning models made significant headway in the determination of acceptable templates. However, even with the significant growth in the number of experimentally solved protein structures, the experimentally determined folds formed a sparse set. The number of diverse templates separated from each other by 1Å–4Å is small, and it is not clear if it is sufficient for meaningful statistical learning of structural adjustments between homologous proteins. It is, therefore, not a surprise that refinement based on statistical learning of pair interactions has not shown significant progress. It is interesting that a past study (Májek and Elber, 2009) illustrated that even optimal pair potentials select a substantial number of incorrect structures. An informed speculation would be that interactions of more than two bodies are necessary in the refinement process. While higher order potentials were developed, the statistics to estimate the parameters of these functions using log-odd ratios is poor, and to the best of my knowledge, did not lead to significant improvement of homologous models.

An alternative to the use of statistical potentials is the use of physics-based simulations. The arguments against this approach are that of cost and accuracy. For a refinement step that requires relatively small adjustments in structure, there is hope that cost could be reduced substantially and only accuracy will remain a concern. Unfortunately, solving one of the two problems is not good enough. Accuracy was proven to be a concern, and a direct application of physics-based modeling did not lead to systematic and significant improvements in the quality of the structures of the templates with respect to the native folds.

The manuscript by Jian et al. (2011) is a major contribution to the field for two reasons. First, it proposes a new technical idea deviating from the approaches mentioned above. Having something different at hand is more likely to yield new promising results. The approach of Jian et al. (2011) is based on statistics of protein fragments incorporated into molecular dynamics, or physics-based simulations. The use of fragments allows the incorporation of many body effects into the energy function, and at the same time, retains a consistent statistical model that is not strongly influenced by the sparseness of the data. The second observation that makes this paper different is the practical success. The observation that structures with a TM score as low as 0.5 with respect to the native fold are consistently improved in the new calculation is remarkable. The simultaneous careful implementation of software to optimize the hydrogen bonding pattern and to eliminate steric clashes further adds to the improvement of the structures, which is also evident from the comparison to CASP8 and CASP9 targets. It should still be noted that the improvements are not the final solution to the problem because they are typically small and lead to improvement only in the second or third digit of quality measures, such as the GDT. A small improvement is, however, better than no improvement. It is therefore expected that fragment based biases as an addition to atomically detailed energy functions will find their way to other homology modeling and refinement programs and that the present investigation will quickly become influential.

REFERENCES

Májek, P., and Elber, R. (2009). Proteins, Structure, Function and Bioinformatics *76*, 822–836.

Jian, Z., Yu, L., and Zhang, Y. (2011). Structure *19*, this issue, 1784–1795.