Large Scale Structural Genome Annotation with 123D Threading

Theo Mevissen, Friedrich Rippmann[1], Ralf Zimmer, and Thomas Lengauer
Institute for Algorithms and Scientific Computing (SCAI)
GMD – German National Research Center for Information Technology
Schloß Birlinghoven, D-53754 Sankt Augustin, Germany
E-mail: {Theo.Mevissen,Ralf.Zimmer,Thomas.Lengauer}@gmd.de
[1]Merck KGaA, Darmstadt, E-mail: Friedrich.Rippmann@merck.de

Abstract

We address the problem of analysing and interpreting large amounts of sequence data from genome projects and EST sequencing via structural and functional annotation methods implemented on multiprocessor machines and clusters of workstations. Proprietary sequence information is produced by a few genomics companies and provided to the large pharmaceutical companies, which all exploit very similar raw data for finding possible drug target proteins in a competitive way. This implies that the speed and sensitivity of screening and analysis methods are of utmost importance for success in identifying truly novel targets via bioinformatics approaches.

Currently, several function prediction methods are in everyday use based on sequence similarity in order to detect distant homology [2]. However, most sequences do not show significant similarity to any sequence of known structure. In these cases, "threading" tries to predict the structure of these sequences by exploiting the wealth of known structural information thereby broadening the range of detectable similarity.

One such threading approach developed at GMD is the program 123D [1] of the software package ToPLign [4]. A parallel version of 123D has been developed and transferred to Merck [3] and found immediate application in a number of ongoing research projects. The parallel 123D software allowed for the tentative functional annotation of several complete small bacterial genomes and of a large number of partial gene sequences (so-called EST assemblies) of hitherto unknown structure and function. These functional annotations via the prediction of a three-dimensional structure could lead to target ideas which subsequently could be validated and fed into the drug discovery process chain.

As the problem of threading a number of sequences to a large set of folds is easily split into independent subtasks the main problem is to manage resources of thousands of processes efficiently. An almost perfect increase of speed was achieved with up to 24 processors on SUN/SGI shared memory machines as well as workstations clusters build up from Sun, SGI and Linux/Intel Pentium PCs. Thereby, running times of screening tasks for small genomes with 4000 sequences have been reduced from more than three days to about 3h.

The prediction data are available over the WWW [3] to allow for the inspection of top scoring folds, the distribution of scores in specific experiments, and the close inspection of underlying alignments and models. The alignments and models can be further analysed with the ToPLign system.

The work has been supported by the BMBF (Project TargId) and the EU (TTN/DECHEMA).

References

[1] N. Alexandrov, R. Nussinov, and R. Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In L. Hunter and T. E. Klein, editors, *Pacific Symposium on Biocomputing'96*, pages 53–72. World Scientific Publishing Co., 1996.

[2] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. Predicting function: From genes to genomes and back. *Journal of Molecular Biology*, 283(4), 1998.

[3] F. Rippman and R. Zimmer. Multi-processor protein structure prediction threading with ToPLign. Technical report, Merck KGaA and GMD-SCAI, October 1998.

[4] R. Thiele, R. Zimmer, and T. Lengauer. Recursive dynamic programming. *Journal of Molecular Biology*, 290(3):757–779, 1999.