Computational Proteomics on the Grid

E. HUEDO[1], U. BASTOLLA[1], R. S. MONTERO[2] and I. M. LLORENTE[2,1]

[1] *Centro de Astrobiología (CSIC-INTA). 28850 Torrejón de Ardoz, Spain.*

[2] *Dpto. de Arquitectura de Computadores y Automática. Universidad Complutense, 28040 Madrid, Spain.*

Abstract

The large number of protein sequences, provided by genomic projects at an increasing pace, constitutes a challenge for large scale computational studies of protein structure and thermodynamics. Grid technology is very suitable to face this challenge, since it provides a way to access the resources needed in compute and data intensive applications. In this work, we concentrate on the grid-aware implementation of a protein structure prediction algorithm.

Keywords Bioinformatics, Grid, Adaptive Scheduling and Execution.

One of the main challenges in Bioinformatics concerns the analysis of the huge amount of protein sequences. The structure and function of a protein is coded in its amino acid sequence, but deciphering it has turned out to be a very difficult problem, which is still waiting for a complete solution. In this work we consider a protein structure prediction algorithm, that computes the gapped alignment between the target sequence and each structure in the Protein Data Bank (PDB). When a close relative of the target structure is present in the PDB, the algorithm recognizes it and produces a good alignment between sequence and structure. In such cases, the algorithm can be used to estimate thermodynamic parameters of the target sequence, such as the folding free energy.

We want to apply this protein structure prediction algorithm to a large number of families of proteins performing the same function in different organisms, extending a previous study which showed that folding efficiency is lower in proteins of intracellular bacteria than in their free-living relatives[4)].

Grid computing[2)] constitutes an appropriate platform to execute high throughput computing (HTC) applications like the one described above. This kind of applications comprises the execution of a high number of tasks, each of which performs a given calculation (gapped alignment) over a subset of parameter values (the protein sequences to be analyzed), and potentially shares common files. One of the most challenging problems to efficiently execute HTC applications is the fact that Grids present unpredictable changing conditions, namely: dynamic resource availability, load, and cost, and a high fault rate.

Adaptive Grid scheduling[1)] is generally accepted as the cure to the dynamicity of the Grid. We have modified several components of the Grid*W*ay framework[3)] as well as some application characteristics to achieve the adaptive functionality required in this dynamic scenario. We have developed a resource broker that reflects application preferences and the dynamic characteristics of Grid resources. Also, the application has been modified to generate restart files, and so to be able to restart the execution from a given point.

In particular, we have applied the structure prediction algorithm to 88 sequences of the *Triose Phosfate Isomerase* enzyme, expressed in different organisms. The experiment was conducted in the UCM-CAB research testbed[3)]. The execution time for the 88 jobs was 7.15 hours, which supposes a mean job turnaround time of 4.88 minutes. Compared to the single host execution on the fastest machine in the testbed, these results roughly represents a 50% reduction in the overall execution time. These promising experiments show the potentiality of the Grid for the study of large numbers of protein sequences, and suggests the possible application of this methods on a genome scale.

Acknowledgment

This research was supported by MCyT (research grant TIC 2003-01321), INTA, and CSIC (I3P Programme).

References

1) F. Berman et al. Adaptive Computing on the Grid Using AppLeS. *IEEE Transactions on Parallel and Distributed Systems*, 14(5):369–382, 2003.

2) I. Foster and C. Kesselman. *The Grid: Blueprint for a New Computing Infrastructure.* Morgan-Kaufman, 1999.

3) E. Huedo, R. S. Montero, and I. M. Llorente. A Framework for Adaptive Execution in Grids. *Intl. J. Software–Practice and Experience*, 2003. to appear.

4) R. van Ham et al. Reductive Genome Evolution in Buchnera Aphidicola. *Proc. Natl. Acad. Sci. USA*, 100:581–586, 2003.