Contribution ID: **129** Type: **Oral**

Grid Solving a Bioinformatics Challenge: a First Step to Anchoring the Nucleosome

Tuesday, 12 February 2008 11:20 (20 minutes)

The nucleosome involves a complex of eight proteins (histones) binding to 147 base-pairs of DNA. Simulating a nucleosome core bound to a single DNA sequence would require treatment of roughly 250,000 atoms and many months of computer time. To understand selective binding we need to compare many potential binding sequences and hence perform many such simulations. Given that any of the four nucleic acid bases can occupy each position within the bound DNA, there are roughly 10^86 potential sequences to test. We have been able to reduce this task by dividing the DNA into overlapping fragments containing four nucleotide pairs (4^4=256 sequences for each pair). By minimizing each sequence in turn for each fragment, and then moving one step along the nucleosome-bound DNA, we can reconstruct the binding energies of all possible sequences with approximately 36,000 optimizations using the JUMNA program (developed in our team). The whole task would take roughly four years on a single processor.

3. Impact

We have used the production grid set up by the EGEE-II project. We have submitted 35,840 energy minimizations as individual jobs on the grid. This means that each job had gone through the submission processes, and thus paid the overhead inherent to the grid architecture and internal processes: from the submission through the user interface (UI), via the scheduling step on the resource broker (RB) to the execution on the computing element (CE), a cluster with several worker nodes (WN). The whole computing task was launched through 12 RBs, which have scheduled all the jobs on 23 CEs. The total cumulated computing time was about 1,275 days, with a job duration of 51 minutes on average. The full calculation was completed after 4 days and 16 hours, running up to 1039 jobs simultaneously. This was 271 times faster than using a single machine.

URL for further information:

http://gbio-pbil.ibcp.fr

4. Conclusions / Future plans

Using the EGEE grid to obtain a first indication of the binding specificity of the nucleosome turned out to be rather efficient. The results have demonstrated the sustainable status of the EGEE grid for large-scale experiments with a real laboratory workflow. We are planning to continue our study with an improved model that will require 140,000 energy minimizations, corresponding to roughly 16 years of sequential CPU time.

Provide a set of generic keywords that define your contribution (e.g. Data Management, Workflows, High Energy Physics)

Bioinformatics, Molecular simulation, Large scale experiment

1. Short overview

How proteins find their targets amongst millions (or more) of competing sites is still largely an unsolved problem. Understanding this process in detail is however central to understanding the mechanisms underlying gene expression. A better understanding of site-specific targeting is also a vital step towards rational re-engineering of proteins for therapeutic purposes.The problem becomes even harder when a complex of several proteins binds to DNA, as in the case of the nucleosome core particle.

Primary author: Dr BLANCHET, Christophe (CNRS IBCP)

Co-authors: Mr MICHON, Alexis (CNRS IBCP); Dr ZAKRZEWSKA, Krystyna (CNRS IBCP); Dr LAVERY, Richard (CNRS IBCP)

Presenter: Dr BLANCHET, Christophe (CNRS IBCP)

Session Classification: Life Sciences

Track Classification: Scientific Results Obtained Using Grid Technology