Contribution ID: **138**

Type: **Oral contribution**

Using the grid to solve a bioinformatics Challenge: Locating nucleosome positions

Tuesday, 23 September 2008 15:00 (15 minutes)

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. The problem becomes even harder when a complex of several proteins bind to DNA, an in the case of the nucleosome core particle. The nucleosome involves an eight protein complex binding to 147 bp of DNA. To understand selective binding we need to compare many potential binding sequences. Given that any of the four nucleic acid bases can occupy each position within the bound DNA, there are roughly 10ˆ86 potential sequences to test. We have been able to simplify this task by dividing the DNA into overlapping fragments containing five nucleotide pairs. Each such fragment can have 1024 sequences. By minimizing each sequence in turn for each fragment (allowing for local DNA and protein side chain relaxation), and then moving one step along the nucleosome-bound DNA, we can reconstruct the binding energies of all possible sequences with approximately 280,000 optimizations. Each optimization uses the JUMNA program developed in our team and takes, on average, one hour. This implies that the whole task would require roughly 22 years on a single processor.

This problem was overcome using a grid platform to distribute the independent minimizations. We have used the production grid set up by the EU-EGEE project, which brings together 41,000 CPUs and 5 PB of storage amongst 200 sites world-wide. The distribution of the minimization tasks all over the grid have crunched the execution time from 22 years to roughly 11.5 days, using at best 1,850 CPUs simultaneously. This performance was obtained with a pilot job system developed in our team. This system deploys adaptative agents on the grid. Each agent (i) ensures that the remote computing environment fulfills the requirement of our JUMNA program, (ii) compiles and optimizes JUMNA, and (iii) recursively fetches sets of data to compute. This system avoids having failed tasks due to bad remote computing environments, and decreases failed jobs due to unclear reasons.

The scientific results are being analyzed to quantify the optimal postions of nucleosomes within the chromosomes of the human genome. The preliminary results are very encouraging and in accord with known experimental data, notably concerning nucleosome organization upstream of transcription start sites. It should be emphasized that these are the first predictions made using all-atom energy calculations, in contrast to much faster, but also much less precise, estimates based on DNA sequence properties.

This work is supported by the CNRS, by the French Agence Nationale de la Recherche through the projects HIPCAL (ANR-06-CIS6-005) and HUGOREP (NT05-3_41825) and by the sixth European Framework Program through the project Enabling Grid for E-science II (EGEE, EU-FP6 INFSO-RI-031688).

Primary authors: MICHON, Alexis (CNRS IBCP); Dr BLANCHET, Christophe (CNRS IBCP); Dr ZAKRZEWSKA, Krystyna (CNRS IBCP); Dr LAVERY, Richard (CNRS IBCP)

Presenter: Dr BLANCHET, Christophe (CNRS IBCP)

Session Classification: Bioinformatics Community Meeting