Sizing Up: Bioinformatics in a Grid Context

Fotis E. Psomopoulos* and Pericles A. Mitkas
Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, Thessaloniki, 54-124, Greece.
*Correspondence to: fpsom@issel.ee.auth.gr

Abstract: A Grid environment can be viewed as a virtual computing architecture that provides the ability to perform higher throughput computing by taking advantage of many computers geographically distributed and connected by a network. Bioinformatics applications stand to gain in such environment both in regards of computational resources available, but in reliability and efficiency as well. There are several approaches in literature which present the use of Grid resources in bioinformatics. Nevertheless, scientific progress is hindered by the fact that each researcher operates in relative isolation, regarding datasets and effort, since there is no universally accepted methodology for performing bioinformatics tasks in a Grid. Given the complexity of both the data and the algorithms involved in the majority of cases, a case study on protein classification utilizing the Grid infrastructure, may be the first step in presenting a unifying methodology for bioinformatics in a Grid context.

Background: There is already a lot of research in progress towards applying parallel computing techniques on bioinformatics methods, such as multiple sequence alignment, gene expression analysis and phylogenetic studies [1] among others. Moreover, there is a recent trend in utilizing the Grid as a platform for complex approaches to bioinformatics problems, such as reverse-engineering gene-regulatory networks [2] and the Grid Basic Local Alignment Search Tool (Grid-Blast) [3], and in building specific frameworks over the Grid that target genomics and proteomics issues, such as the Biotechnology Information and Knowledge Grid (BioGrid) and the Bioinformatics and Genomics Grid for European Research (BIG-GER). However, there is still a lack of a common analysis framework for genomic and proteomic data, which could exploit the benefits of a grid environment.

Challenges: The advancement in the technologies is providing an exponentially rising amount of data. For this reason there is a shift in research from hypothesis-driven to data-driven studies. This shift is presenting scientists with new challenges in data analysis. On one hand, most proteomics studies have a limited number *n* of instances (i.e. samples or records). Usually *n* ranges from tens to hundreds of cases, opposed to the thousands or millions of instances in a typical engineering or finance application. On the other hand, high-throughput techniques in life sciences yield several thousand variables per case, in sharp contrast with the traditional data mining scenarios. This problem is known as the *curse of dimensionality*, or *small-n-large-p problem*. The curse of dimensionality is especially evident when dealing with protein classification studies. Overall, the major methods for protein classification can be distinguished into three basic groups: pairwise sequence comparison algorithms [4], generative models for protein families [5-6] and discriminative classifiers [7]. Despite the differences in the actual methods, they all share a common factor; building an accurate protein classification system depends critically upon choosing a good representation of the input sequences.

Methods: In order to cope with the dimensionality issue, most machine learning algorithms focus on specific groups of proteins or reduce either the size of the original data set or the number of attributes involved. Grid computing could potentially provide an alternate approach to this issue, by combining multiple approaches in a seamless way. There are several recent attempts in integrating in a seamless way these two fields: data mining and grid computing [8]. Nevertheless, a unifying methodology that couples them together, and at the same time taking into account the specific needs and constraints of the protein classification problem will potentially open the way for close collaboration between domain experts and computer scientists.

References

[1] "*Parallel Computing for Bioinformatics and Computational Biology*", Wiley Series on Parallel and Distributed Computing, A. Zomaya, Wiley-Interscience, 2006, pp. 193– 210.
[2] M. Swain, T. *et al.*, "*Reverse-engineering gene-regulatory networks using evolutionary algorithms and grid computing*", J. of Clinical Monitoring and Computing, vol. 19, no. 4-5, pp. 329–337, 2005.
[3] A. Krishnan, "*Gridblast: a globus-based high-throughput implementation of blast in a grid computing framework*", Concurrency and Computation: Practice and Experience, vol. 17, no. 13, pp. 1607–1623, 2005.
[4] S. F. Altschul *et al.*, "*A basic local alignment search tool*", Journal of Mol. Biol., vol. 215, pp. 403–410, 1990.
[5] J. Park *et al.*, "*Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods*", J. of Mol. Biol., vol. 284, pp. 1201–1210, 1998.
[6] F. E. Psomopoulos and P. A. Mitkas, "*A protein classification engine based on stochastic finite state automata*", in Lecture Series on Computer and Computational Sciences VSP/Brill, vol. 4B, Loutraki, Greece, Oct. 2005, pp. 1371–1374.
[7] J. Weston, *et al.*, "*Semi-supervised protein classification using cluster kernels*", Bioinformatics, vol. 21, no. 15, pp. 3241–3247, 2005.
[8] V. Stankovski, *et al.*, "*Grid-enabling data mining applications with dataminninggrid: An architectural perpective*", Future Generation Computer Systems, vol. 24, pp. 259–279, 2008.