Bioinformatics Application Framework Through GenPipe Approach

Márcio Katsumi Oikawa and João Eduardo Ferreira
{koikawa,jef}@ime.usp.br

University of São Paulo
Institute of Mathematics and Statistics
Department of Computer Science

Abstract

The majority of Molecular Biology experiments depends on serial execution of different applications to generate relevant results for future analyses. The distinct input/output format standards of these applications require complex features for data importing and exporting [1]. In addition, the applications need to be viewed as "black boxes" and atomic units for the workflow definitions, which can change according to the complexity of the experiment, origining sophisticated analysis components [3].

Format differences are generally compensated through human manipulation, leading to failures and wrong results in large scale data, increasing error ratios and decreasing the quality of the final information [2].

In this research, we propose an alternative methodology to integrate heterogeneous applications using their properties and characteristics to formalize generic and atomic components [4]. This approach avoids translation failures of heterogeneous formats, implementing a intermediate XML layer between each pair of applications. The modelling for the composition and integration has originated a generic infra-structure for customizable and flexible workflows. This has been supported by a workflow definition language [5], automated through software systems, supporting and improving analysis power for the researchers. We have called this alternative methodology of the *GenPipe*.

keywords: Biological workflows, information modelling, metadata and data sharing, data semantic and control, molecular sequence analysis.

Acknowledgment

This presentation is part of the project titled "Integration of Data and Applications for Computational Molecular Biology", funded by the Research Support Foundation of São Paulo State (FAPESP), grant number 00/10062-3.

References

[1] Frederic Achard, Guy Vaysseix, and Emmanuel Barillot. XML, bioinformatics and data integration. *Bioinformatics*, 17(2):115–125, 2001.

[2] Tim Hubbard. Biological information: making it accessible and integrated (and trying to make sense of it). *Bioinformatics*, 18(90002):140S–, 2002.

[3] Marcio K. Oikawa and Joo Eduardo Ferreira. Heterogeneous database integrator to support genomic data mining. *In preceedings of International Conference on Bioinformatics 2002*, page 88, 2002.

[4] J. Payton, R. Gamble, S. Kimsen, and L. Davis. The opportunity for formal models of integration, 2000.

[5] Mor Peleg, Iwei Yeh, and Russ B. Altman. Modelling biological processes using workflow and Petri Net models. *Bioinformatics*, 18(6):825–837, 2002.