An Integrative Bioinformatics Approach for Knowledge Discovery

Lourdes Peña-Castillo, Sieu Phan, and Fazel Famili

Institute for Information Technology,
National Research Council Canada, Ottawa, Ontario, K1A 0R6, Canada
{lourdes.pena-castillo,sieu.phan,fazel.famili}@nrc-cnrc.gc.ca

Abstract. The vast amount of data being generated by large scale omics projects and the computational approaches developed to deal with this data have the potential to accelerate the advancement of our understanding of the molecular basis of genetic diseases. This better understanding may have profound clinical implications and transform the medical practice; for instance, therapeutic management could be prescribed based on the patient's genetic profile instead of being based on aggregate data. Current efforts have established the feasibility and utility of integrating and analysing heterogeneous genomic data to identify molecular associations to pathogenesis. However, since these initiatives are data-centric, they either restrict the research community to specific data sets or to a certain application domain, or force researchers to develop their own analysis tools. To fully exploit the potential of omics technologies, robust computational approaches need to be developed and made available to the community. This research addresses such challenge and proposes an integrative approach to facilitate knowledge discovery from diverse datasets and contribute to the advancement of genomic medicine.

Keywords: Bioinformatics, knowledge discovery, genetic diseases.

1 Introduction

Large-scale omics experiments are continuously being performed providing a constant source of a huge amount of data. The sheer accumulation of this data have required the development of computational approaches to deal with the collection, storage, integration, analysis, visualization and dissemination of these data sets. The goal of these computational approaches is to enable researchers to advance their understanding of the physiological purpose or molecular role of human proteins, with a special emphasis in better comprehending the molecular basis of genetic disorders. Integration of heterogeneous data sources is of paramount importance to obtain a global view of the molecular associations to multifactorial diseases such as Alzheimer's disease, diabetes, and cancer. These molecular associations could lead to new therapeutic targets, new diagnostic tests, new drug design, and ultimately the transformation of the medical practice.

Current efforts have shown the feasibility and utility of integrating and analysing heterogeneous omics data to identify molecular associations to patho-genesis [1,2]

M. Ulieru, P. Palensky, and R. Doursat (Eds.): IT Revolutions 2008, LNICST 11, pp. 254–257, 2009.

and to predict gene function [3]. Most of these initiatives are data-centric in the sense that their focus is to provide access to high-quality data and to allow researchers to explore this data. These initiatives are highly valuable resources and provide a crucial service to the research community; however, since their focus is to disseminate data and to facilitate the analysis of this data, they limit researchers to a specific application domain (e.g., cancer instead of other genetic disorders) and to specific data sources. To fully exploit the potential of omics technologies, robust computational approaches need to be developed and made available to the community. The recently NCI-launched initiative, caBIG [4] has among its goals the development and dissemination of a compendium of freely available software applications. However, to the best of our knowledge, in caBIG's current tool collection an integrative approach offering a unified framework from data integration to knowledge discovery is missing.

Here, we propose the development of an integrative computational approach to facilitate knowledge discovery from heterogeneous omics data sets that provides a unified framework from data integration to the application of multiple machine learning algorithms to derive classification or prediction models. In addition, this approach is suitable for a variety of application domains (e.g., various diseases).

2 An Integrative Approach to Facilitate Discovery of Biological Knowledge

Our pipeline provides an unified framework to support the following tasks:

1. creation of a data collection which may include data from high throughput studies and available annotation data,
2. query-based dynamic data integration,
3. selection of candidates (genes, functions, pathways, protein complexes, etc.) of interest, and
4. model construction.

Our integrative approach for biological knowledge discovery distinguishes itself from other approaches in:

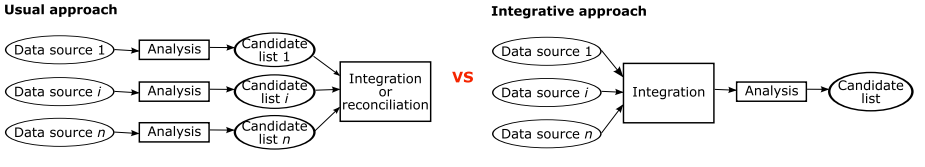

Fig. 1. Comparison of the usual integration approach vs our integrative approach

- performing integration of the heterogeneous data sources before analysis as shown in Figure 1, and
- allowing various abstraction levels during analysis as depicted in Figure 2.

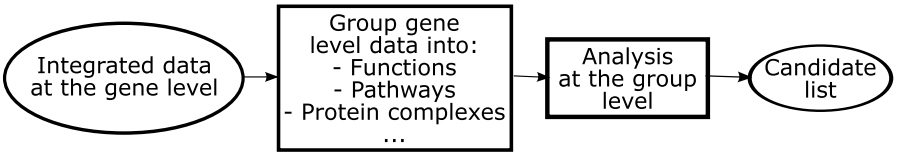

Fig. 2. Analysis at various abstraction levels

By performing integration of diverse evidence available before analysis, we increase the sensitivity of our approach, improve the confidence in the list of candidates since the list is based on several sources of evidence, and are able to observe biological relationships; for instance, we observe a mutation affecting a certain gene and the differential expression of that gene. By analyzing and selecting candidates at various abstraction levels, our list of candidates directly consists of the groups (genes, functions, pathways, etc) of interest.

The main advantages of our approach are a sound methodology for data integration and dealing with missing data; various abstraction levels during analysis (e.g., gene-based or pathway-based); three-state classification models (models include an inconclusive state for borderline cases), and strategies for evaluation of multiple models.

3 Conclusion

To fully exploit the potential of omics technologies, robust computational approaches need to be developed and made available to the community. Several current efforts are focusing on the development of these computational approaches. We propose an integrative computational approach supporting the complete data analysis workflow from data integration to knowledge discovery. Our approach will allow researches to use diverse data sets of their choice and be suitable for various application domains. The end goal of our project is to facilitate the identification of molecular associations to pathogenesis. These associations may lead to the identification of new targets for better diagnosis tests, drugs and, ultimately, to personalized medical care.

References

1. Rhodes, D., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B., Barrette, T., Anstet, M., Kincead-Beal, C., Kulkarni, P., et al.: Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. Neoplasia 9(2), 166 (2007)
2. The Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216), 1061–1068 (2008)

3. Peña-Castillo, L., Tasan, M., Myers, C., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W., et al.: A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology 9(suppl. 1:S2) (2008)
4. Wolfson, W.: caBIG: Seeking Cancer Cures by Bits and Bytes. Chemistry & Biology 15(6), 521–522 (2008)