Integrative Data Analysis for Biological Discovery

Sai Moturu
Arizona State University, USA

Lance Parsons
Arizona State University, USA

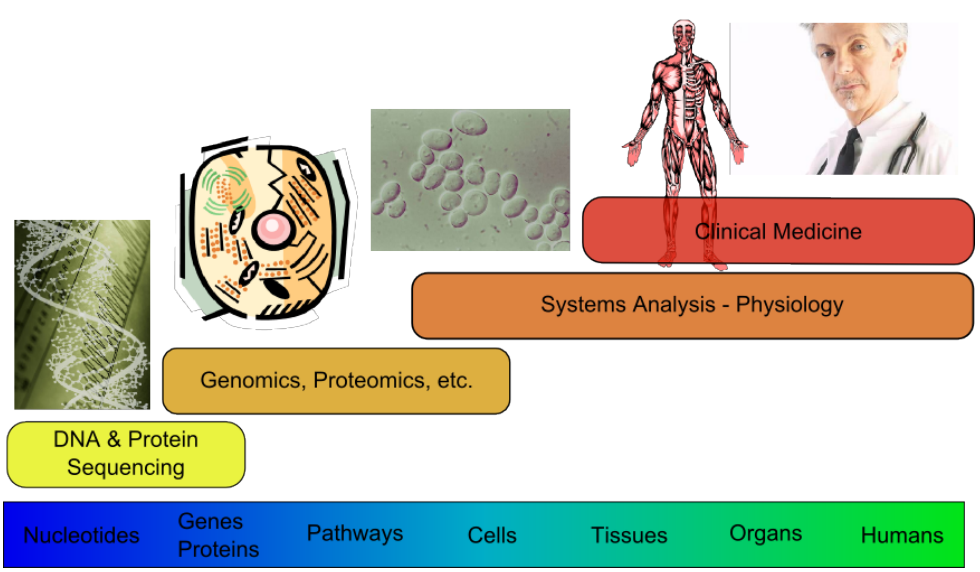
Zheng Zhao
Arizona State University, USA

Huan Liu
Arizona State University, USA

INTRODUCTION

As John Muir noted, "When we try to pick out anything by itself, we find it hitched to everything else in the Universe" (Muir, 1911). In tune with Muir's elegantly stated notion, research in molecular biology is progressing toward a systems level approach, with a goal of modeling biological systems at the molecular level. To achieve such a lofty goal, the analysis of multiple datasets is required to form a clearer picture of entire biological systems (Figure 1). Traditional molecular biology studies focus on a specific process in a complex biological system. The availability of high-throughput technologies allows us to sample tens of thousands of features of biological samples at the molecular level. Even so, these are limited to one particular view of a biological system governed by complex relationships and feedback mechanisms on a variety of levels. Integrated analysis of varied biological datasets from the genetic, translational, and protein levels promises more accurate and comprehensive results, which help discover concepts that cannot be found through separate, independent analyses. With this article, we attempt to provide a comprehensive review of the existing body of research in this domain.

Figure 1. Complexity increases from the molecular and genetic level to the systems level view of the organism (Poste, 2005).



Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

BACKGROUND

The rapid development of high-throughput technologies has allowed biologists to obtain increasingly comprehensive views of biological samples at the genetic level. For example, microarrays can measure gene expression for the complete human genome in a single pass. The output from such analyses is generally a list of genes (features) that are differentially expressed (upregulated or downregulated) between two groups of samples or ones that are coexpressed across a group of samples. Though every gene is measured, many are irrelevant to the phenomenon being studied. Such irrelevant features tend to mask interesting patterns, making gene selection difficult. To overcome this, external information is required to draw meaningful inferences (guided feature selection). Currently, numerous high-throughput techniques exist along with diverse annotation datasets presenting considerable challenges for data mining (Allison, Cui, Page & Sabripour, 2006).

Sources of background knowledge available include metabolic and regulatory pathways, gene ontologies, protein localization, transcription factor binding, molecular interactions, protein family and phylogenetic information, and information mined from biomedical literature. Sources of high-throughput data include gene expression microarrays, comparative genomic hybridization (CGH) arrays, single nucleotide polymorphism (SNP) arrays, genetic and physical interactions (affinity precipitation, two-hybrid techniques, synthetic lethality, synthetic rescue) and protein arrays (Troyanskaya, 2005). Each type of data can be richly annotated using clinical data from patients and background knowledge. This article focuses on studies using microarray data for the core analysis combined with other data or background knowledge. This is the most commonly available data at the moment, but the concepts can be applied to new types of data and knowledge that will emerge in the future.

Gene expression data has been widely utilized to study varied things ranging from biological processes and diseases to tumor classification and drug discovery (Carmona-Saez, Chagoyen, Rodriguez, Trelles, Carazo & Pascual-Montano, 2006). These datasets contain information for thousands of genes. However, due to the high cost of these experiments, there are very few samples relative to the thousands of genes. This leads to the curse of dimensionality (Yu & Liu 2004). Let M be the number of samples and N be the number of genes. Computational learning theory suggests that the search space is exponentially large in terms of N and the required number of samples for reliable learning about given phenotypes is on the scale of $O(2^N)$ (Mitchell, 1997; Russell & Norvig, 2002). However, even the minimum requirement (M=10*N) as a statistical "rule of thumb" is patently impractical for such a dataset (Hastie, Tibshirani & Friedman, 2001). With limited samples, analyzing a dataset using a single criterion leads to the selection of many statistically relevant genes that are equally valid in interpreting the data. However, it is commonly observed that statistical significance may not always correspond to biological relevance. Traditionally, additional information is used to guide the selection of biologically relevant genes from the list of statistically significant genes. Using such information during the analysis phase to guide the mining process is more effective, especially when dealing with such complex processes (Liu, Dougherty, Dy, Torkkola, Tuv, Peng, Ding, Long, Berens, Parsons, Zhao, Yu & Forman, 2005; Anastassiou, 2007).

Data integration has been studied for a long time, ranging from early applications in distributed databases (Deen, Amin & Taylor, 1987) to the more recent ones in sensor networks (Qi, Iyengar & Chakrabarty, 2001) and even biological data (Lacroix, 2002; Searls, 2003). However, the techniques we discuss are those using integrative analyses as opposed to those which integrate data or analyze such integrated data. The difference lies in the use of data from multiple sources in an integrated analysis framework. The range of biological data available and the variety of applications make such analyses particularly necessary to gain biological insight from a whole organism perspective.

MAIN FOCUS

With the increase in availability of several types of data, more researchers are attempting integrated analyses. One reason for using numerous datasets is that high-throughput data often sacrifice specificity for scale (Troyanskaya, 2005), resulting in noisy data that might generate inaccurate hypotheses. Replication of experiments can help remove noise but they are costly. Combining data from different experimental sources and knowledge bases is an effective way to reduce the effects of noise and generate more accurate hypotheses. Multiple sources provide additional information that

6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/integrative-data-analysis-biological-discovery/10952

Related Content

Pattern Discovery as Event Association

Andrew K.C. Wong, Yang Wangand Gary C.L. Li (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1497-1504).*

www.irma-international.org/chapter/pattern-discovery-event-association/11018

Mining Chat Discussions

Stanley Loh Daniel Licthnowand Thyago Borges Tiago Primo (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1243-1247).*

www.irma-international.org/chapter/mining-chat-discussions/10981

Multiclass Molecular Classification

Chia Huey Ooi (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1352-1357).*

www.irma-international.org/chapter/multiclass-molecular-classification/10997

The Personal Name Problem and a Data Mining Solution

Clifton Phua, Vincent Leeand Kate Smith-Miles (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1524-1531).*

www.irma-international.org/chapter/personal-name-problem-data-mining/11022

Program Mining Augmented with Empirical Properties

Minh Ngoc Ngo (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1610-1616).*

www.irma-international.org/chapter/program-mining-augmented-empirical-properties/11034