Analyzing and Visualizing Genomic Complexity for the Derivation of the Emergent Molecular Networks

Theodoros Koutsandreas, National Hellenic Research Foundation, Athens, Greece
Ilona Binenbaum, National Hellenic Research Foundation, Athens, Greece
Eleftherios Pilalis, e-NIOS Applications PC, Kallithea, Greece
Ioannis Valavanis, e-NIOS Applications PC, Kallithea, Greece
Olga Papadodima, National Hellenic Research Foundation, Athens, Greece
Aristotelis Chatziioannou, National Hellenic Research Foundation, Athens, Greece

ABSTRACT

Modern genomic studies, accumulation of biological information in repositories, plus novel analytical and data-mining methodologies, comprise the backbone for the holistic explanation of intricate phenotypes, interrogated by high-throughput experiments. Recent developments in web platforms architecture, in conjunction with novel, browser-centric, visualization techniques pose a powerful framework for the development of distributed web applications, which execute complex analytical tasks, display the results in user-friendly interface and produce comprehensive, visualization charts. In this paper, the presented client-server application targets the systemic interpretation of input gene lists, through the fusion of established statistical methodologies and information-mining techniques, while interactive visualization modules aid the intuitive interpretation of results. Two publicly available datasets, related to Crohn’s and Parkinson’s disease are used to present application analytical efficiency, robustness and functionalities.

KEYWORDS

Big Data in Biomedicine, Biomarkers, Gene Signatures, Integrative Bioinformatics, Pathway Analysis, Target Prioritization

INTRODUCTION

In this paper, a powerful gene analytics web application is introduced, which overall combines computational methodologies and data visualization techniques, in order to deliver comprehensible illustrations of cellular complexity, for voluminous, high dimensional, molecular datasets, that address genome in its functional entirety. Moreover, by linking the individual genes, with the critical cellular processes that orchestrate physiological response, through the exploitation of appropriate controlled vocabularies, which ideally possess a tree structure and aid application of deductive logic and knowledge discovery, it achieves a systemic interpretation of the mechanism interrogated, while at the same time it manages to prioritize functionally genes with a pronounced regulatory involvement in the important cellular processes underlying the phenotype studied.

The completion of Human Genome Project in April 2003 (Venter et al., 2001) marks the dawn of the genomic era, disclosing the astonishing complexity of the genomic regulation and functionality that underlies phenotypic diversity. Ever since, dramatic scientific progress has been made, delivering...
advanced experimental and analytical methodologies that leverage the noted complexity and aid the elucidation of intricate biological problems. In particular, advancements in DNA microarray (Gresham, Dunham, & Botstein, 2008), or more recently next generation sequencing technologies (Goodwin, McPherson, & McCombie, 2016) have generated an avalanche of high-throughput experimental genomic data, holding the promise for the decipherment of complex regulatory mechanisms, underlying the genesis and progression of complex diseases or more generally pathologies, through the creation of stable links between the comprehension of aberrant homeostatic response and the phenotypic cues pertinent with disease manifestation. Moreover, the development of several biomedical ontologies and databases, structuring and categorizing the knowledge amassed concerning the study of biomolecules in different temporal, spatial and organizational scales, has contributed to the rapid expansion of the digitized informational universe, through the avid accumulation of semantic terms, describing biological complexity of inter- and intra-cellular mechanisms (Ashburner et al., 2000), encompassing metabolic and signaling pathways (Croft et al., 2014; Kanehisa & Goto, 2000) or phenotypes (Eppig et al., 2014; Köhler et al., 2014) among others. This highly heterogeneous and therefore idiosyncratic ecosystem, poses significant challenges, when it comes to its efficient integration and meaningful interpretation into concrete, mechanistic modeling approaches. This enables a deeper understanding of the mechanism under investigation on one hand, while at the same time provides a rational framework for the testing of novel hypotheses, as proposed from a computational, data-driven research perspective.

High-throughput and large-scale genomic analysis leads to extended lists of genes, with differentiated status among the various phenotypic categories observed, or more broadly the examined stratifications. These profiles set the genomic fingerprint of the examined phenotypes. Statistical enrichment analysis (Subramanian et al., 2005) is an efficient strategy to systematically associate the extracted gene list with various functionalities, as these are encapsulated in terms of biomedical vocabularies, revealing the respective mechanistic impression of a biological phenotype at the given level of molecular dissection that the vocabulary describes. Various enrichment analysis approaches exploit biomedical ontologies (vocabularies with hierarchical tree structure) and gene repositories integrating mappings between genes and functional or structural descriptions, through the application of well-known statistical methods (Fisher’s exact test, Hypergeometric distribution, Chi-squared test) with the scope to rigorously reveal the over-represented entities, correlated with the input list (Huang, Sherman, & Lempicki, 2009).

Yet the derivation of a set of semantic terms through enrichment analysis, is only the first step towards a system-level interpretation. The formulation of a concrete network of genes (interaction networks), that captures critical aspects of phenotypic causality, remains an important goal. A final list of genes, prioritized and imprinted in a scale free network (Réka, 2005), indicates potential biomarkers and therapeutics targets for diseases and disorders. Gene prioritization tools start the analysis from a large pool of candidate genes, trying to detect and rank the most promising of them, under various criteria. For instance, they exploit well-curated genomic data sources, in order to glean potentially important information such as mappings of seed genes with diseases and specific keywords, or other semantic entities appropriate for application of text mining techniques, exploitation of sequential information for the conduct of sequence similarity profiling inquiries or analysis of protein-protein networks. However, all that techniques are biased by the level of prior knowledge about the examined phenotype, namely the construction of adequate see

Furthermore, the advances in ‘omics’ and the entailed genome-wide studies in conjunction with the rapid development of open source data visualization modules, render the comprehensive depiction of high-throughput genomic results a compelling yet feasible task (Gehlenborg et al., 2010). The exploitation of smart visualization tools, coupled to the intelligent data-mining or functional analysis methodologies, enable the intuitively meaningful mapping of complex relations between key
A Conceptual Framework for Rock Data Integration in Reservoir Models Based on Ontologies