Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Integrating information from existing databases for enhanced function annotation of genes, genomes and networks

Lokesh Pati Tripathi[1, *]

[1]National Centre for Biological Sciences, TIFR, GKVK Campus, Bangalore-560065, India; Lokesh Pati Tripathi* - E-mail: lokesh@ncbs.res.in; Phone: 91 80 23666251; * Corresponding author

received November 19, 2007; accepted November 14, 2007; published online December 11, 2007

Abstract:
Uncovering functional associations for genes and gene products remains one of the most significant challenges in biology. The classical approaches, such as homology detection, are mainly suited for predicting approximate molecular function of a protein and should be used in context with other methods. Several studies have emerged that employ knowledge-based procedures to extract functional data for genes from a variety of biological sources. However, data derived from a single biological resource often provides only a limited perspective on their functional associations largely due to systematic bias in the underlying data. The post-genomic era has witnessed the emergence of knowledge-based studies that aim to decipher functional associations by combining several biological evidence types. These are expected to provide better insights into the functional aspects of diverse genes, genomes and networks.

Keywords: homology; post-genomic; functional associations; knowledge-based; evidence types

Description:

One of the most significant challenges is to assign reliable function association to genes and gene products. Many proteins (and domains) in eukaryotic genomes are part of multi-member families; they participate in many cellular processes and are located in different parts of cells. In recognition to this, the Gene Ontology (GO) consortium annotates information about molecular function, biological process and cellular component to describe "function" of a protein. **[1]** In the post-genomic era, there have been attempts at function annotation that procure data from varied sources. Biological data from a single type of data source, though useful, is often limited in the extent to which it may help uncover functional associations; either because of a systematic bias towards specific genes, gene families and pathways and/or incorporation of false positives during data acquisition. With focus shifting from genes and proteins to biological systems, integrating information from multiple data types is seen as a more robust and accurate means of unraveling functional associations. **[2]**

Several attempts have been undertaken to obtain biological data from multiple data types and implement statistical frameworks for their integration. These studies employ data sources such as global gene expression patterns **[3]**, yeast two-hybrid data **[4]**, genomic characteristics **[5]**, genome-wide RNAi screens **[6]**, literature information *etc*. Approaches that combine information from multiple data types with information from peer-reviewed scientific literature are particularly successful in providing functional associations for genes and gene products. **[7]** Most of these frameworks output profiles or clusters of genes based on their similarities within a particular data source and their interpretation is largely dependent on expert knowledge. For instance, microarray technology allows simultaneous study of expression patterns of thousands of genes under specific conditions. Expression data is analysed (using approaches such as clustering) to identify sets of genes with similar expression patterns that are assumed to function in similar physiological processes. Clustering aims to partition genes such that genes with similar expression patterns fall into the same groups called clusters. Since gene clusters are often inclined to be enriched in specific functional categories, identification of such clusters may be used to assign putative functional associations to uncharacterised genes within those clusters. Approaches such as hierarchical clustering, *k*-means, self organizing map (SOM), principal component analysis (PCA) have been employed to identify sets of co-expressed genes and tools are available for visualisation of these clusters. Different datasets may often provide overlapping or complementary information due to hierarchy in the definition of function of a gene **[8]**; integrating knowledge from various data types thus, provides a uniform view of functional associations and is most useful when coupled with expert knowledge. Few such

attempts have proved to be highly successful in annotating prokaryotic genomes **[9]** and there have been few attempts in eukaryotic genomes as well. **[3]**

In a recent study, knowledge from structure-function analysis of 3-D structures and sequences, gene expression profiling, text mining, protein-protein interactions and knowledge-based computational tools (Figure 1) have been extensively employed to manually assign either of the three GO **[1]** categories to the putative members of trypsin-like serine proteases (SPs) family encoded in the genome of *Drosophila melanogaster*. **[10]** Through this approach, functional information was obtained for 190 gene products containing serine protease like domains. This approach provides significant functional information for 30 of 190 gene products and to assign a putative function to these with high confidence. Of these ten are supported by literature curation and four are supported by Flybase annotations, while annotations for 16 gene products are entirely derived from analysis of large-scale datasets employed in the study (see supplementary data). A large scale involvement of many *Drosophila* SPs and SPHs (that are likely to be proteolytically inactive due to mutations in the residues of the serine protease catalytic triad) was observed in development and immune response, which would explain the diversity observed for this gene family in *Drosophila*. The approach also helps uncover putative functional associations between genes involved in different metabolic pathways. For example, *Drosophila* SP CG3066 is a monophenol monooxygenase activator involved in activation of melanization chiefly in response to fungal infection and believed to be involved in a possible cross-talk between melanization and Toll pathway. Time-series expression data suggests that expression of CG3066 is correlated with Easter and Snake, members of Toll signaling pathway in *Drosophila*. Also, studies suggest high similarity in the putative active sites of CG3066 and Easter. Thus, a probable role for CG3066 in association with the components of the Toll pathway may be associated in early embryogenesis.

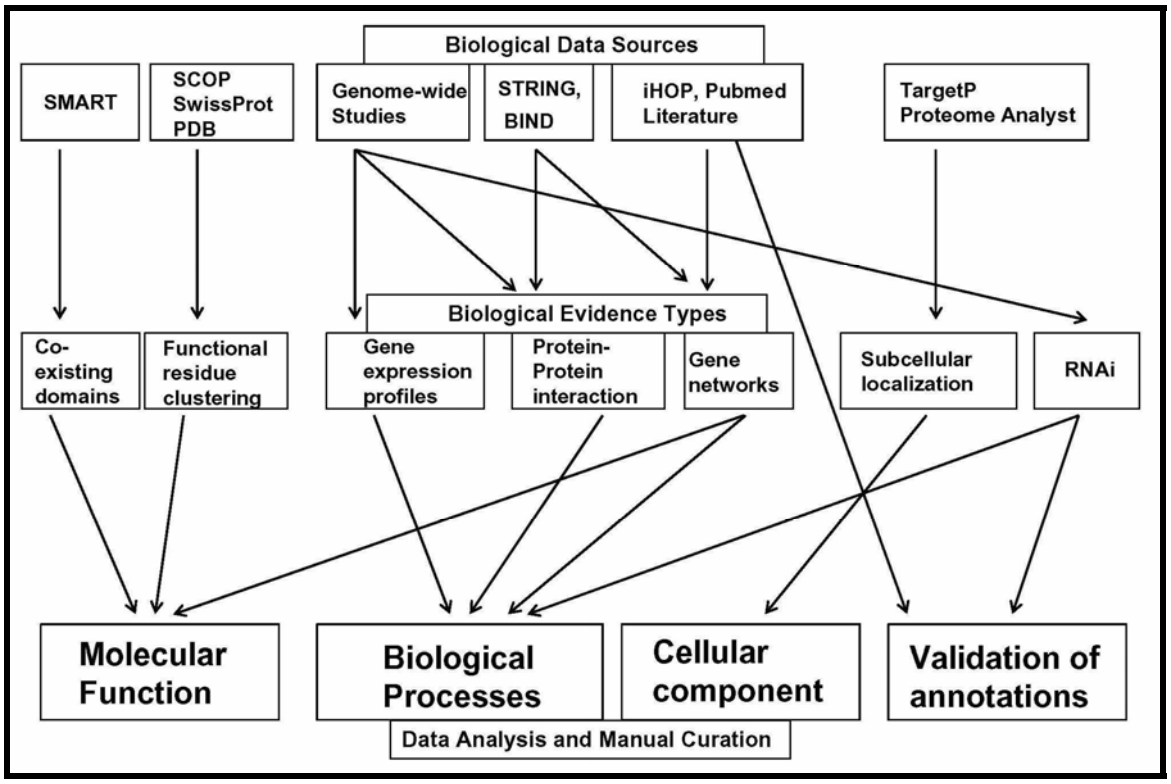

Figure 1: A schematic representation of biological data sources and evidence types employed for the enhanced function annotation of *Drosophila* SPs

Thus, integration of biological data from diverse sources provides an effective means for large-scale function annotations of genes, multi-member gene families and networks. The evolution of such tools is likely to gain further momentum as enormous amounts of high-throughput experimental data from diverse sources are likely to become available in the near future.

Supplementary data are available at http://caps.ncbs.res.in/download/Bioinformation/.

Acknowledgement:
L. T. is a Senior Research Fellow of the Council of Scientific and Industrial Research (CSIR), India. The author would also like to thank Prof. R. Sowdhamini for critical reading and inputs for the manuscript.

References:

[01] S. E. Lewis, *Genom Biol.,* 6: 103 (2005) [PMID: 15642104]
[02] J. Li, *et al., Bioinformatics,* 22: 2037 (2006) [PMID: 16820427]
[03] R. Jansen, *et al., Science,* 302: 449 (2003) [PMID: 14564010]
[04] P. Uetz, *et al., Nature,* 403: 623 (2000) [PMID: 10688190]
[05] E. M. Marcotte, *et al., Nature,* 402: 83 (1999) [PMID: 10573421]
[06] L. M. Cullen and G. M. Arndt, *Immunol Cell Biol.,* 83: 217 (2005) [PMID: 15877598]
[07] L. J. Jensen, *et al., Nat Rev Genet.,* 7: 119 (2006) [PMID: 16418747]
[08] L. J. Jensen and L. M. Steinmetz, *FEBS Lett.,* 579: 1802 (2005) [PMID: 15763555]
[09] C. V. Mering, *Nucl Acids Res.,* 33: 433 (2005) [PMID: 15608232]
[10] P. K. Shah, *et al., Gene,* 407: 199 (2007) [PMID: 17996400]

Edited by P. Kangueane
Citation: Tripathi, Bioinformation 2(4): 132-134 (2007)