Genomic and proteomic data integration for comprehensive biodata search

Arif Canakoglu✉, **Marco Masseroli**
Dipartimento di Elettronica ed informazione, Politecnico di Milano, Milan, Italy

Motivation and Objectives

With high-throughput technologies in the life sciences, particularly in molecular biology, the amount of data available has grown exponentially. Yet, such data are stored in several different formats and spread into numerous databanks (Galperin and Fernández-Suárez, 2012). This scenario makes even more difficult to find and retrieve the data required to answer the scientists' questions, which usually are complex and regard multiple biological entities and several of their aspects. Consequently, in the last few decades biological data integration has become a major focus in bioinformatics. Data integration is essential to comprehensively evaluate and search information from different databanks. For example, no single data source exists that supplies association data between protein interactions and genetic disorders.

There are several approaches, with related implementations, to integrate heterogeneous data from different sources, such as information linkage (e.g. SRS (Etzold et al., 1996), NCBI Entrez (Tatusova et al., 1999)), federated databases (e.g. BioKleisli (Davidson et al., 1997), DiscoveryLink (Haas et al. 2001)), multi-databases (e.g. TAMBIS (Stevens et al., 2000), BACIIS (Miled et al., 2002)), mediator based solutions (e.g. BioDataServer (Freier et al., 2002), Biomediator (Cadag et al., 2007)) and data warehousing (e.g. EnsMart (Kasprzyk et al., 2004), BioWarehouse (Lee et al., 2006)). Data warehousing is the most convenient one when the data are very numerous and offline processing is a necessity to mine integrated data efficiently and comprehensively. Using such an approach, we created an integrative data warehouse, where integration is performed based on a predefined modular data model that provides a unified reconciled global view of the integrated data. Data warehouse creation and updating is performed by supervised automatic procedures, which also control variation of the integrated data in the original data sources (Davidson et al., 1995). The used modular data model supports both easy data warehouse extensibility, with the integration of new data sources, and effective automatic querying on the integrated data for their search and extraction.

Methods

We built a Genomic and Proteomic Knowledge Base (GPKB), which is a relational, integrative and multi-organism data warehouse containing heterogeneous genomic and proteomic annotation data. We import them from several well known public databases, including Entrez Gene, UniProt, IntAct, MINT, BioCyc, KEGG, Reactome, GO, GOA and OMIM, The very numerous data integrated, which regard biomolecular entities (mainly genes and proteins) and their biomedical features and associations, are all checked for data correctness and consistency (Ghisalberti et al., 2010). By leveraging imported similarity and historical evaluation data available, we identify different IDs from different data sources as representing the same entity. This enables us to classify and extract different attributes available also from different data sources as referring to the same entity, feature or association, rather than as distinct attributes of different entities or of their features or associations.

For the GPKB, we designed a modular global data schema with abstraction and generalization of the main data features. It is characterized by a multi-level data architecture, which includes source-import level, instance-aggregation level and concept-integration level.

Leveraging on such data schema, we defined query templates to extract the integrated data. These query templates allow extracting the user required data from any version of the GPKB automatically. This supports different Web applications and services connected to the GPKB in automatically searching and extracting data from the data warehouse for different goals, including gene and protein annotation inference, annotation enrichment analysis and user query support for biomedical knowledge discovery.

The performed inference of gene and protein annotations is based on the "transitive closure" concept. It is inspired by Swanson work (Swanson, 1986) that is based on the transitive closure of het-

erogeneous extensive annotation data. The inference procedure is controlled by Standard Query Language (SQL) templates, which are applied to any relational biomedical molecular database.

Results and Discussion

With the data downloaded on May 28th, 2012, among others, the GPKB contained 9,537,645 genes of 9,631 organisms, 38,960,202 proteins of 338,004 species, 19,522 protein domains and 824,797 protein domains annotations, 28,889 biochemical pathways and 171,372 pathway annotations (77,812 gene and 93,560 protein annotations), 35,252 Gene Ontology terms and 64,185,070 Gene Ontology annotations (1,272,168 gene and 62,912,902 protein annotations), 10,212 human genetic disorders and their 27,705 gene annotations. Furthermore our GPKB integrates also other types of data regarding DNA sequences, transcripts, enzymes, small molecules of biological interest, and clinical synopses. In total it contains more than 103,006,922 gene annotations and 183,209,462 proteins annotations.

The great amount of biomolecular features and their association data that the GPKB contains makes it a unique valuable resource which can be used for different applications, in silico experiments and knowledge discoveries.

The created automatic query templates make possible to easily search and extract each of the integrated data, offering an efficient base for various data mining algorithms and applications. As an example, by leveraging the multi-source integrated date, we inferred new gene annotations through transitive closure on various association data regarding the features of the gene encoded proteins. The same approach enabled us also to infer possible associations between protein-protein interactions and genetic disorders. Towards this aim, protein-protein interaction data files downloaded from MINT (Licata et al., 2012) and IntAct (Kerrien et al., 2012) databases were automatically parsed. Data of 46,154 human protein-protein interactions (out of the contained 254,048 protein-protein interactions of 397 different organisms' proteins), regarding 12,178 distinct human proteins, were imported in the data warehouse. These human proteins, which represent 3.7% of all the 326,766 human proteins in the data warehouse, are encoded by 11,232 different human genes. By applying the transitive closure concept on the interacting protein encoding gene and genetic disorder related gene association data, we identified 1,130 gene-gene interactions and found 1,136 human protein-protein interactions possibly associated with 628 genetic disorders (such as Alzheimer, Cystic fibrosis, Diabetes mellitus, Parkinson, etc.). Such genetic disorders resulted related to 86 clinical synopses and 3,481 phenotypes.

The created Genomic and Proteomic Knowledge Base, that is updated quarterly, can be freely accessible through an easy-to-use Web interface available at http://www.bioinformatics.dei.polimi.it/GPKB/ where all integrated data in the GPKB can be comprehensively searched.

Acknowledgements

This research is part of the "Search Computing" project (2008-2013), funded by the European Research Council (ERC), under the 2008 call for "IDEAS Advanced Grants".

References

1. Cadag E, Louie B, et al. (2007), Biomediator data integration and inference for functional annotation of anonymous sequences. Pac Symp Biocomput. 343-354.
2. Davidson SB, Overton C, et al. (1995), Challenges in integrating biological data sources. J. Comput. Biol. 2(4):557-572.
3. Davidson SB, Overton C, et al. (1997), BioKleisli: a digital library for biomedical researchers. Int. J. Digit. Libr. 1997, 1(1):36-53. doi:10.1007/s007990050003
4. Etzold T, Ulyanov A, et al. (1996): SRS: Information Retrieval System for molecular biology data banks. Meth. Enzymol. 266:114-128.
5. Freier A, Hofestädt R, et al. (2002), BioDataServer: a SQL-based service for the online integration of life science data. In Silico Biol. 2(2):37-57.
6. Galperin MY, Fernández-Suárez XM (2012), The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res. 40(Database issue):D1-D8. doi:10.1093/nar/gkr1196
7. Ghisalberti G, Masseroli M, Tettamanti L (2010), Quality controls in integrative approaches to detect errors and inconsistencies in biological databases. J Integr Bioinform 7(3):199,1-13. doi: 10.2390/biecoll-jib-2010-119
8. Haas LM, Rice JE, et al. (2001), DiscoveryLink: a system for integrated access to Life Sciences data sources. IBM Systems Journal 40(2):489-511.
9. Kasprzyk A, Keefe D, et al. (2004), EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 14(1):160-169. doi:10.1101/gr.1645104
10. Kerrien S, Aranda B, et al. (2012), The IntAct molecular interaction database in 2012. Nucleic Acids Res. 40(Database issue):D841-846. doi:10.1093/nar/gkr1088
11. Lee TJ, Pouliot Y, et al. (2006), BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics, 7:170,1-14. doi:10.1186/1471-2105-7-170

12. Licata L, Briganti L, et al., (2012), MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40(Database issue):D857-D861. doi:10.1093/nar/gkr930
13. Miled ZB, Li N, et al. (2002), Complex life science multidatabase queries. Proc. IEEE 90(11):1754-1763. doi:10.1109/JPROC.2002.804683
14. Stevens R, Baker P, et al. (2000), TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics, 16(2):184-185.
15. Swanson DR (1986), Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 30(1):7-18
16. Tatusova TA, Karsch-Mizrachi I, et al. (1999), Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics 15(78):536-543.