Data relationship mining in life science databases

Matthias Lange, Nese Sreenivasulu, Andreas Stephanik, Uwe Scholz

Institute for Plant Genetics and Crop Plant Research (IPK) Gatersleben, Germany
{lange|srinivas|stephanik|scholz}@ipk-gatersleben.de

Nowadays, publicly available databases play a central role in the so-called ”in-silico” biology. This term describes the manifold bioinformatics methods for analysing data to derive information in the field of interest in general. Some examples for those analyses are prediction of transcription factor binding sites, sequence homology searches, gene expression analysis, pattern recognition, or comprehensive data mining on integrated data towards system biology. In order to interpret wet-lab data, it is common practice to combine and correlate the data with the knowledge stored in publicly available databases [Ste02, SGBB01]. Thus, primary molecular biological data directly obtained in the lab, is enriched by the results of analysis methods and finally stored in various databases [BK03]. Current reports enumerate about 550 databases in the field of molecular biology [Gal05].

The analysis and interpretation of wet-lab data using in-silico methods can be regarded as an iterative process of tool application and database exploration. To illustrate such a process, we will give an example: a list of sequences, in this example Expressed Sequence Tags (ESTs), has to be assigned to classes of metabolic functions. For this purpose, the scientist may use the well-standardized EC numbers as unique database identifiers and use the KEGG [KGKN02] database to retrieve the related metabolic pathways for a list of given EC numbers. In order to map the EST sequences to enzymes of pathways, the scientist can apply a sequence homology search tool such as BLAST [AGM$^+$90] against sequences annotated with EC numbers. Then, the EC numbers have to be extracted from the hit descriptions and the pathway to be queried from KEGG. Such homology searches result in links to several different databases. These direct hits are the entry point to collect more detailed information and frequently do not contain the actual answer to the query motivating question. In fact, transitively related, e. g. database links must be tracked. In our concrete example, the user has to manually navigate through the hyperlinked databases. For finding the right way through the hyperlinked data, he has to benefit from his browsing experience or learn which links on which web site should be followed to find the transitively linked pathway in KEGG for our example. In particular, the direct hits in databases supporting no systematic or controlled vocabulary can not be used for automatic functional categorization or quantitative analysis scenarios. Hence, the basis for an in-silico data mining analysis should be an integrated database together with controlled vocabularies [DR04, GYMP04, KAN$^+$05]. While the integration of data is a widely addressed problem, here we primarily focus on efficient traversing within all the integrated data.

In this context, we built a materialized integrated database, called ”DBOra” containing about 35 million entries of proteomics data. A database mediation system called *BioDataServer* (BDS) [BLSS04] was used to support the database integration. The basic idea of this database integration approach is to import integrated data as a materialized copy using a pre-modelled database schema. The result is an integrated database with 81 interconnected tables. Using Oracle9i, 50 tables their selves represent database links, which make the developed schema and thus the database as the basis for the subsequently described data relationship mining.

To meet the requirement of mining transitive relations in public life science databases, we combined materialized database integration with a graph-based approach for data relationship mining. At modelling time, we possess knowledge about possible relationships between our entities. This knowledge is used to construct *entity relationship graphs* of interlinked public data items.

This relation can be used to "walk" through interconnected data and to discover transitive links and especially "hidden", not obviously recognizable, transitive relations between life science data items. If we want to apply this to a large amount of data items, an efficient algorithm must be used.

A natural join over all tables in the schema for computing the transitive relationships is not applicable because of a polynomial complexity: If all of our i tables were naturally joined ($\bowtie$) and a cascading merge join were used[1] (complexity of each merge join is $O(n)$), the overall complexity would be:

$$O_{join} = O(n)_{table_1 \bowtie table_2} * O(n)_{table_2 \bowtie table_3} \ldots * O(n)_{table_{i-1} \bowtie table_i} = O(n)^i$$

The task is now to reduce this complexity by pre-computing all spanning undirected graphs over all attributes of interconnected data as basis for efficient transitive relationship queries. To do so, so-called *data linkage graphs* (DLGs) are computed. We used the recursive breadth-first search (BFS) algorithm as basis for our algorithm, which was used for the shortest path algorithm [Dij59]. As known from the literature, the complexity is $O(n)$.

The computed graphs can be used directly for the mentioned applications for automatic functional EST classification. For this, a simple start-destination-search in the graph has to be performed. This search retrieves all graphs that connect a given start-node to an end-node. The complexity of such a query is derived from the complexity of the two selection operations for selecting the graphs containing the queried start and end node, and additionally their union. Using indexes for the selection, the complexity is $O(log\ n)$. Applying merge join on sorted graph identifiers, the complexity is $O(n)$. Therefore, the overall complexity of the above query is:

$$O_{graphquery} = O(log\ n) + O(log\ n) + O(n) = O(n)$$

If we do not take in account that we need to compute the DLG with the BFS complexity (because it is necessary only once per database update), this is a significant reduction of the polynomial natural join complexity.

To bring the concept of DLGs to a biological application, we built functional classes by using controlled vocabulary. For this purpose, we decided to take the KEGG metabolic pathway names as controlled vocabulary. Thus, we distinguish 11 classes of KEGG "superpathways" and 200 single pathways.

Using the DLGs and the controlled vocabulary, we were able to annotate 18,607 out of 111,090 in-house sequenced barley ESTs to metabolic pathways. By using BLAST results, only 1,906 functional descriptions containing an EC number were found. Hence, compared to BLAST, the usage of DLGs increased sensitivity of functional annotation 10-fold.

[1]The cascading sort-merge join is the commonly used join implementation in relational database management systems [EN00].

To give a reliable basis for EST annotation, one has to consider potential quality problems of the presented approach. During this process, we checked the reliability of Blast results by considering the secured logarithmic E-value and the length of the aligned sequence segment with the top database hit. In addition, we considered information about percentage of identity, percentage of similarity and position of the alignment. Our manual analysis resulted in identification of 104 non-redundant false positive cases.

Apart this application of data relationship mining, it is possible to perform more comprehensive applications of the presented concepts. The one described was used to give a proof-of-principle. More use cases and further EST categorizations, e.g. to the Gene Ontology vocabulary, are embedded in the "Crop EST Information System" (CR-EST) [KLF+05], which is available using the URL `http://pgrc.ipk-gatersleben.de/cr-est`.

References

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 3:403–410, 1990.

[BK03] F. Bry and P. Kröger. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. *Distributed and Parallel Databases*, 13(1):7–42, 2003.

[BLSS04] S. Balko, M. Lange, R. Schnee, and U. Scholz. BioDataServer: An Applied Molecular Biological Data Integration Service. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Data Integration in the Life Sciences; First International Workshop, DILS 2004 Leipzig, Germany*, volume 2994 of *Lecture Notes in Bioinformatics*, pages 140–155. Berlin et al: Springer, 2004.

[Dij59] W. Dijkstra, E. A note on two problems in connection with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[DR04] H.-H. Do and E. Rahm. Flexible integration of molecular-biological annotation data: The genmapper approach. In *Procceedings of the 9th International Conference on Extending Database Technology*. Springer LNCS, 2004.

[EN00] N. Elmasri and S. B. Navathe. *Fundamentals of database systems*. Reading, Mass. et al: Addison-Wesley, 3rd international edition edition, 2000.

[Gal05] Michael Y. Galperin. The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Research*, 33(suppl_1):D5–24, 2005.

[GYMP04] K. C. Gunsalus, W.-C. Yueh, P. MacMenamin, and F. Piano. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Research*, 32(90001):D406–410, 2004.

[KAN+05] A. Kahraman, A. Avramov, L. G. Nashev, D. Popov, R. Ternes, H.-D. Pohlenz, and B.m Weiss. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, 21(3):418–420, 2005.

[KGKN02] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG database at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002. http://www.genome.ad.jp/kegg/.

[KLF+05] C. Künne, M. Lange, T. Funke, H. Miehe, I. Grosse, and U. Scholz. CR–EST: a resource for crop ESTs. *Nucleic Acids Research*, 33(suppl_1):D619–621, 2005.

[SGBB01] R. Stevens, C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–188, 2001.

[Ste02] L. Stein. Creating a bioinformatics nation. *Nature*, 417:119–120, 2002.