

# High-throughput annotation of genomic datasets with Genephony

Angelo Nuzzo and Alberto Riva

**Abstract** — We describe the initial implementation of Genephony, an online tool for the creation and manipulation of very large genomic datasets. Genephony allows the user to easily create “sets” containing biological entities (e.g. genes, SNPs, pathways, microarray probesets, etc) and to combine them in a variety of ways in order to generate new sets. Sets are stored in a dynamic workspace through which the user can freely navigate. Relying on an extensive underlying database of genomic information, the system makes it easy to integrate, annotate and interpret the results of high-throughput experiments, providing automated operations that would be otherwise impractical if performed manually. Our expectation is that Genephony will become a useful tool for translational research, high-throughput biology, and for all knowledge-intensive data manipulation tasks in computational biology.

**Index Terms** — Databases in Bioinformatics and Biotechnology, High-throughput Biology, Web-based Tools.

## I. INTRODUCTION

ONE of the main challenges resulting from the widespread adoption of high-throughput technologies in the biological sciences is that researchers are increasingly faced with the task of manipulating and interpreting very large datasets. These datasets are usually contained in files that can reach sizes of hundreds of megabytes or even gigabytes for a single experiment. In addition to the technological issues related to the storage and efficient retrieval of the data they contain, and to the problem of developing analysis methodologies able to cope with several thousand variables at once, we believe there is a range of “basic” data manipulation operations that, although conceptually simple, become extremely difficult to perform when working with very large datasets. For example, the simple task of associating a probeset identifier from a gene expression microarray with the gene it corresponds to, and then to add information about the chromosomal location of the gene, its GeneOntology classification, and possibly its known disease associations, can quickly become impractical if performed manually on more than a small number of genes, in addition to requiring familiarity with several different data-

bases and informational resources. These apparently simple manipulations are nevertheless essential when trying to extract new knowledge from the results of a high-throughput experiment, in an automated and efficient way.

Many of the software tools that the researchers are familiar with are not designed to handle very large datasets: for example, the number of rows in a Microsoft Excel table is limited to 65,536, which is one order of magnitude smaller than the number of probesets in the most recent generation of genotyping microarrays. At the same time, it is not realistic to expect the average researcher to become proficient in programming or even simple database design. In order to take advantage of the wealth of information produced by the new technologies and to move towards a “systems biology” perspective that encompasses a large number of variables at once, researchers increasingly need user-friendly, powerful and flexible tools for large-scale data integration, manipulation and annotation.

This paper describes the initial implementation of Genephony, an online tool for the creation and manipulation of very large datasets of genomic information. The current prototype provides information about genes, transcripts, arbitrary genomic regions, SNPs (including frequency and genotype data from HapMap), pathways, GeneOntology classes, HomoloGene clusters, PubMed and OMIM entries, micro-RNAs, predicted transcription factor binding sites, and probeset accessions numbers for both gene expression and genotyping microarrays. The current version is limited to human genome data, but its underlying database can be easily extended to other organisms.

## II. SYSTEM DESIGN AND USAGE

The user begins a Genephony session by *creating* one or more initial datasets (either by manually entering appropriate identifiers or by uploading them from files). New datasets can then be generated by *deriving* them from an existing one or by *combining* two of them. For example, two distinct sets of genes can be combined to generate a third one containing the genes that appear in both of them (possibly indicating that the same gene was identified by two different experiments). From this third set the user could then derive the set of SNPs belonging to the genes it contains. Finally, datasets can be *filtered* (to display only a subset of their contents), *browsed*, or

A. Nuzzo is with the Dipartimento di Informatica e Sistemistica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (e-mail: angelo.nuzzo@unipv.it).

A. Riva is with the Department of Molecular Genetics and Microbiology and with the University of Florida Genetics Institute, University of Florida, PO Box 103610, Gainesville, FL 32610-3610 (Corresponding author; e-mail: ariva@ufl.edu).

exported in a variety of common formats. All datasets are kept in a “workspace” through which the user can freely navigate: the system keeps track of how each set was generated, and of which other sets were generated by it. It is therefore always possible to reconstruct the path through which any single dataset was generated.

The user interface is designed to be intuitive and easy to use. When creating new datasets, users only need to enter one or more identifiers, either by typing them in a text field or uploading them from files. In most cases the system will be able to automatically recognize the kind of identifier being used and will populate the dataset with objects of the appropriate type. The system will also accept files compressed with the ‘gzip’ or ‘ZIP’ utilities, allowing for faster upload times and reduced disk usage (in the case of a ZIP archive, the system will automatically ask the user to choose which one of the files contained in the archive should be parsed). All uploaded files are stored on the server in a user-owned directory, together with information about how they were parsed by the system. They can therefore be easily re-parsed if necessary, for example to create a dataset based on the contents of a different column, or in case the underlying database is updated resulting in improved annotations.

The main Genephony window is composed of two panels: a smaller one that displays all the sets currently present in the workspace (allowing the user to easily “focus” on one of them by simply clicking on its name), and a larger one displaying information about the set the user is working with. All available operations are clearly indicated by buttons that are always visible. The system keeps track of all the possible ways in which datasets can be combined and/or derived from each other, and presents all applicable options to the user in a detailed, readable form. Very complex sequences of data manipulation steps can therefore be performed with just a few clicks, and no knowledge of the structure of the underlying database is required.

### III. PERFORMANCE

The Genephony server is optimized to efficiently handle very large datasets, both in terms of memory usage and speed. When a dataset is created it initially contains only the unique identifiers of the objects that belong to it; the objects themselves are actually created only when they are needed (for example when the user needs to export the data). The queries used to populate the datasets are optimized for speed by combining multiple retrievals into a single query, using an adaptive strategy to deal with the limitations of the database communication protocol.

The entire system (except for the relational database server, based on MySQL) is implemented using the Common Lisp programming language, a dynamical, high-performance, ob-

ject-oriented development environment explicitly designed to handle large quantities of complex data structures. As an example of the performance of the system, we were able to upload a gzip-compressed file (compressed size 4.5MB, uncompressed 15.9MB) containing over 318,000 SNP identifiers, and to create the corresponding set of SNPs, in 42 seconds of real time on a medium-range desktop machine.

### IV. DISCUSSION

The efficient and effective integration of heterogeneous data and knowledge is one of the most pressing current problems in computational biology and bioinformatics, due to the exponential increase in the amount of available knowledge and data, and the growing trend towards a high-level, “system-wide” view of biological systems. Compared to similar systems such as DAVID [1] and Galaxy [2], Genephony strives to be more general and easier to use. While DAVID is mainly devoted to the annotation of large sets of genes, and Galaxy is oriented towards comparative studies, Genephony allows the user to generate and combine datasets in a free, exploratory fashion, without constraining him/her on a predefined path. Genephony does not currently include data analysis or visualization tools, but its datasets can be easily exported for use in other existing programs for these purposes.

### V. CONCLUSIONS

We described the initial implementation of Genephony, an online tool to handle large datasets aimed at non-technical researchers who need to annotate, integrate and explore genomic data resulting from large-scale experiments. Features that are currently planned for development include: a) basic statistical tools (for example, to perform enrichment analysis on newly generated sets); b) a link with resources providing information on phenotypes and/or disease associations (e.g. OMIM); and c) a backdoor that will provide Genephony with the ability to receive commands from other programs, using a Remote Procedure Call protocol like XML-RPC or SOAP.

The system is robust, efficient and extremely easy to use, and we hope it will become a useful resource for translational researchers who need to effectively deal with large amounts of genomic data and knowledge. While the system is still under development, a demonstration pre-release is available at <http://bioinformatics.ufl.edu/tools/>.

### VI. REFERENCES

1. Giardine B, Riemer C, Hardison RC *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005 15:1451-55.
2. Dennis G, Sherman BT, Hosack DA *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 2003; 4(9): R60.