A data model for comparative genomics

L. A. Digiampietri, C. M. B. Medeiros, and J. C. Setubal

Instituto de Computação, Universidade Estadual de Campinas
CP 6176, Campinas, SP 13084-971 BRAZIL
luciano.digiampietri@ic.unicamp.br

Abstract. We present a simple data model that can be used as a building block in a comparative genomics information system for prokaryotic genomes. The model is extensible and flexible, and has as main entities the organism and the gene family. Existing systems tend to focus either on organisms or in gene families. We have applied the model to a set of eight bacterial genomes, and briefly describe the resulting system.

This paper addresses the following topics
- Biological Databases, Data Management, Data Integration, and Data Mining
- Comparative Genomics and Annotation

1 Introduction

There are currently more than 130 completed microbial genomes [4], and it is expected that in a few years there will be at least a thousand more. This advance has made necessary the development of all kinds of new tools for genomic information management, and in particular for comparative genomics. The aim of this paper is to present a simple data model that can be used in building a larger comparative genomics system. The real situation we have in mind is the one in which there is interest in the comparison of a few (2 to 20) genomes of related bacteria. Such a situation is becoming increasingly common nowadays, given that the genomes of several species of a genus or indeed the genomes of several strains of the same species have been sequenced (a few examples are: *Bordetella, Mycoplasma,* and strains of *E. coli*). We have used the data model presented here in building a system for the comparison of the genomes of eight plant-associated bacteria. A preliminary version of this system was used to obtain results described in [12].

We now define some basic background concepts. Two genes are *homologous* if they share a common ancestor. If the two genes are in the same organism, and therefore are duplications, they are called *paralogous genes.* If the two genes belong to different organisms, they are called *orthologous genes.* The term *replicon* denotes any kind of cell unit that contains genetic information (e.g. chromosomes, plasmids and mitochondria). Thus a genome may contain several replicons.

The basic object in the work presented here is the gene, which is seen as belonging to a genome as well as (in general) to a gene family (which may be a paralogous family or an orthologous family). The raw data we assume for creating our data model come from sequencing genome projects. Other sources of information may exist for genes or genomes, such as gene expression levels, but we concentrate on purely genome-derived information in order to better understand the issues in this context. That said, a comparative genomics information system should ideally support queries on both multiple genomes and multiple gene families. Examples of such queries are:

1. Given two or more genomes, what are the genes shared between them (orthologous genes) and to what families do they belong?
2. Given two or more genomes, what are the genes specific to one genome in relation to the others, and to what families do they belong?
3. Given a gene x from an organism not in the system, does it have homologs in the system? If so, how many?

The above queries are the most important ones in the family of "set operation" queries, where each genome corresponds to a set. The operations can be the standard union, intersection, difference and symmetric difference operations. The queries highlight the importance of the gene family concept in any information system designed for comparative genomics. As will be seen, the basic model presented in the paper allows handling these kinds of queries.

The paper is organized as follows. Section 2 presents some related work. Section 3 describes the data model. Section 4 contains the description of the PABdb system, an information system for comparative genomics whose data model is an extension of the basic data model. Finally, Section 5 contains some conclusions and plans for future work.

2 Related work

Genomic data integration and comparison is a hot research topic. One interesting piece of work is that of Stein et al. [9]. They describe a system (the Generic Genome Browser) designed under a generic data model (Generic Model of Organism System Database, GMOD) to manage genomes. This model is modular, and allows each user to use a subset of model components in order to adjust the system to their needs. GMOD's aims however are geared more towards providing a good platform for one genome rather than several.

An example of an information system that seeks to integrate many genomes under the same data model is the Comprehensive Microbial Resource (CMR), by Peterson et al. [6]. CMR integrates all (or most) completed microbial genomes in its database (called Omniome). The data in Omniome are divided in many levels (genome level, gene level, etc) encapsulating thus specific data inside more general concepts. This encapsulation facilitates the creation of views, queries and data understanding. Genome comparison in CMR can be done through predefined queries to its databases. However a user is tied to the genomes already in the database, so it is difficult to compare a new genome to those in the CMR.

The Generic Genome Browser and CMR are centered on managing organisms. Another approach in genomic information systems is centered around gene families. One of these is the Cluster of Orthologous Groups (COG) database, by Tatusov et al. [10]. The main entities of the COG database are protein-coding genes from various organisms. Genes are grouped according to similarity criteria. The database structure by itself does not support the storing of data related to genomes (only gene data) and thus does not allow whole genome analyses. More importantly (for our purposes), as viewed by the world outside of NCBI, COG is not a database *per se*, but rather a "WWW resource". In addition it does not include genes whose function has not been characterized. All genomes to date have presented relatively high fractions of hypothetical genes (between 10 and 40%). It is not uncommon to find an interesting and numerous paralogous family of hypothetical genes in a new genome, which furthermore have homologs in other genomes. The COG database leaves out such genes.

The works discussed in the previous paragraphs offer a basis for genomic data modeling and integration. None of them, to our mind, deals with the problem of having a simple data model that integrates both multiple genomes and multiple gene families. This is the problem that our work seeks to address.

3 A Basic Data Model

The main issues to be addressed in any information system are: data model, tools on data management, query facilities, and visualization aspects. In this section we describe our data model. The data model must contain all entities and attributes necessary to the analysis and comparison of genomes, as well the relationships among these entities. The basic entities in the comparison of prokaryote genomes are: Genome, Replicon, Gene and Gene Family. Figure 1 shows these entities and their relationships. Other entities can be added to the model depending on the particular needs and data availability of each research group. In the information system described in Section 4 a new entity called Category is added to the basic model.

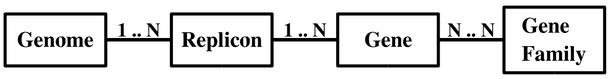

Fig. 1. Basic entities and their relationships.

Based on the model of Figure 1 we developed a database with five relations (tables), shown in Figure 2. Each entity originates a table: genome_tbl, replicon_tbl, gene_tbl and family_tbl. The family_gene_tbl table includes gene family information. The choice of the attributes to be included in each table depends on data availability for the genomes selected. The choice of attributes is based on GenBank [2] replicon files. The primary keys are in bold font. Attribute genome_id in the table gene_family_tbl is redundant, but this redundancy is there for efficiency purposes. For example, to run a query that lists a set of paralogous genes of a given family we need to know the following information: gene_id, family_id and genome_id of each gene. If the attribute genome_id is not in the table gene_family_tbl, this kind of query needs to search the tables gene_family_tbl to find family_id and gene_id values and gene_tbl and replicon_tbl to find the genome_id value of each gene of the given family. The redundancy of genome_id allows search table gene_family_tbl only.

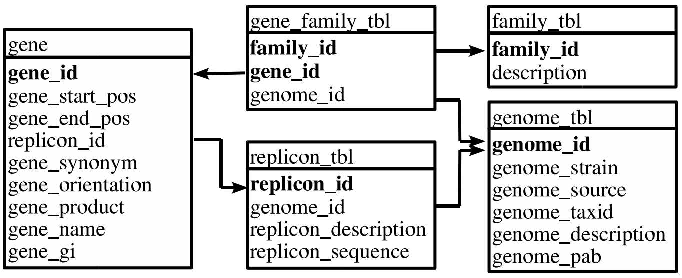

Fig. 2. Basic data model.

The implementation of the gene family concept requires either a description of families and their members or a method to create such descriptions. There is extensive research on gene families, and the COG database [10] is just one example. We consider this to be a system implementation issue, and we address it in the next section.

Given the database tables of Figure 2, the queries mentioned in the introduction can easily be expressed in SQL, as follows:

1. Given two or more genomes, what are the genes shared between them (orthologous genes)?

```
SELECT fg1.gene_id, fg2.gene_id, fg1.family_id
  FROM family_gene_tbl AS fg1, family_gene_tbl AS fg2
    WHERE fg1.family_id = fg2.family_id
      AND fg1.genome_id = 'genome_xxx'
        AND fg2.genome_id = 'genome_yyy';
```

   This SQL query yields all pair of orthologous genes between the genomes genome_xxx and genome_yyy.
2. Given two or more genomes, what are the genes specific to one genome in relation to the others, and to what families do they belong?

```
SELECT gene_tbl.gene_id, family_gene_tbl.family_id
```

```
FROM gene_tbl, replicon_tbl, family_gene_tbl
 WHERE gene_tbl.replicon_id = replicon_tbl.replicon_id
  AND replicon_tbl.genome_id = 'genome_xxx'
   AND family_gene_tbl.gene_id = gene_tbl.gene_id
    AND gene_tbl.gene_id NOT IN
     (SELECT DISTINCT(fg1.gene_id)
       FROM family_gene_tbl AS fg1, family_gene_tbl AS fg2
        WHERE fg1.family_id = fg2.family_id
         AND fg1.genome_id = 'genome_xxx'
          AND fg2.genome_id IN SET_OF_GENOMES);
```

This SQL query will return all gene and family identifiers from the genes that are present in the genome genome_xxx and that do not have orthologous genes in the genomes of the set SET_OF_GENOMES.

We point out that, as it stands today, the COG website database does not allow immediate queries of type 1 and 2 when dealing with 3 or more genomes (but it is possible to create a relational database using COG data allows this [5]).

3. Given a gene x from an organism not in the system, does it have homologs in the system? If so, how many?
We assume we know that gene x belongs to a certain family of the database, suppose family 'family_x'. Then the query becomes:

```
SELECT genome_id, SUM(1) from family_gene_tbl
  WHERE family_id = 'family_x'
    GROUP BY genome_id;
```

The SQL query returns the list of genome identifiers of all genomes that have one or more genes in the same family of gene x, and the number of orthologous genes of gene x in each genome.

4 The PABdb system

The model presented in Section 3 is extensible and flexible. These nice properties are exemplified in this section. Here, we describe an implementation of our data model extended with features in a "real life" situation. This is a system designed to compare the genomes of eight plant-associated bacteria: *Agrobacterium tumefaciens* strain C58, *Mesorhizobium loti*, *Ralstonia solanacearum*, *Sinorhizobium meliloti*, *Xanthomonas axonopodis pv. citri* strain 306, *Xanthomonas campestris pv. campestris* strain ATCC 33913, *Xylella fastidiosa* and *Xylella fastidiosa Temecula1*. We call this system the PABdb system. It has been built using the MySQL DBMS [3]. Partial results using a preliminary version of this system were published in [12]. That publication presented results derived from PABdb, but it did not describe the database itself.

One very important issue in construction of a data model is ensuring its extensibility. This allows both extending a system and its future integration with other systems. The PABdb model extends the basic model specified in Section 3. In addition, PABdb can be considered to be an information system, in the sense that it contains the four basic building blocks of an information system, i.e., database and tools for managing, querying and visualizing data.

Figure 3 shows the conceptual model for the PABdb system. A new entity called Category has been added. This entity is necessary to provide the capability of gene classification, essential in any genomic database. Note that "category" is a concept different from the "gene family" concept. A category (following Gene Ontology terminology [11]) relates to a biological process, molecular function or cellular component. Gene family is a phylogenetic concept. Two genes may have the same category but may belong to different gene families. The entity Category is not present in the basic data model of Section 3 because NCBI [4] gbk files do not contain gene categorization. The values for gene categories should come from standard systems. For plant-associated bacteria so far there is no such system, although GO [11], when properly expanded, may become one. For the PABdb system we used a category system developed for the *Xylella fastidiosa* genome project [8], then refined in later projects, and loosely based on Riley's system [7]. The assignment of functional categories to each gene was automatically made based on BLAST [1] results comparing each gene to a set of genes that had been previously categorized. In other words, a gene inherits the category of its best statistically significant BLAST hit. Table 1 shows the description of the eight basic categories.

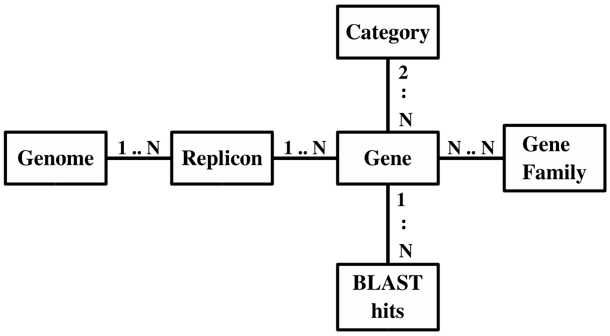

Fig. 3. Entities and relationships of PABdb system.

The data in tables genome_tbl, replicon_tbl and gene_tbl of the PABdb system all came from NCBI [4] gbk files. These files obey the GenBank [2] model, so only one integration tool (to populate the PABdb system) was necessary.

Gene families were created based on BLAST results and on an undirected graph model G. Each gene in PABdb is a vertex in G, and there is an edge between vertices a and b if and only if a is a hit of b in a BLASTP search of

PABdb (maximum e-value: 10^{-20}), or vice versa. The connected components of G are the families. We are aware that this criterion for building families is less rigorous than it should be, but it allows our system to be tested, which is the main goal of the work at this point.

Various tools or scripts were devised to help data access. Some scripts describe data tables or show their contents. A data visualization tool was created that allows plotting of orthologous genes of two genomes. The tool is a simple perl script and was easy to develop given our data model.

The system also includes pre-defined queries (with user-supplied parameter values) to facilitate searches. One of these is query (3) described in Section 1. To illustrate this query we applied it to gene gi:27382283 (a 2-hydroxyacid dehydrogenase of *Bradyrhizobium japonicum*); results are given in Table 2. Another interesting query yields a nice overview of all genes in the system according to category; results are shown in Table 3. Please note again that these results are of course highly dependent on the method for categorizing genes; the point here is not to defend a particular method, but to show that given a categorization, a simple query extracts this valuable table from our system.

ID	**Category description**
I	Intermediary metabolism
II	Biosynthesis of small molecules
III	Macromolecule metabolism
IV	Cell structure
V	Cellular processes
VI	Mobile genetic elements
VII	Pathogenicity, virulence, and adaptation
VIII	Hypothetical

Table 1. Table of the instances of the entity Category.

Organism	number of copies
Agrobacterium tumefaciens	15
Mesorhizobium loti	12
Ralstonia solanacearum	6
Sinorhizobium meliloti	16
Xanthomonas axonopodis pv. citri	3
Xanthomonas campestris pv. campestris	3
Xylella fastidiosa	1
Xylella fastidiosa Temecula1	1

Table 2. Number of copies of a 2-hydroxyacid dehydrogenase (gi:27382283).

Organism	Percentage of genes per category							
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
Agrobacterium tumefaciens	23.3	8.0	7.1	2.7	17.8	1.7	3.8	35.6
Mesorhizobium loti	19.2	7.5	6.4	2.6	12.6	3.2	3.1	45.4
Ralstonia solanacearum	17.4	7.9	7.5	3.6	9.4	4.3	5.3	44.6
Sinorhizobium meliloti	22.6	7.5	6.7	2.8	15.6	3.1	3.2	38.5
Xanthomonas axonopodis pv. citri	17.0	8.2	11.4	4.7	9.1	3.4	6.7	39.5
Xanthomonas campestris pv. campestris	16.7	8.5	11.9	4.8	9.1	4.1	6.8	38.1
Xylella fastidiosa	8.2	8.0	10.8	4.5	4.2	4.7	4.8	54.8
Xylella fastidiosa Temecula1	11.4	11.9	16.5	6.1	6.9	4.5	6.1	36.6

Table 3. Distribution of the genes of each genome by category.

5 Conclusions and future work

Information systems evolve with changes in context and user requirements. This is even more critical in bioinformatics, due to the rapidly advancing pace of the field. In order to meet this challenge, these systems must be scalable and allow exchange of data and operations between each other. One important factor towards this goal is the definition of an appropriate data model.

This paper presented a simple but flexible and extensible data model for comparative genomics. It should be considered a building block of a larger system. We used it to create a useful resource, the PABdb system. This building block can be used to build other kinds of systems for comparative genomics.

An interesting next step would be to try to use our data model in a richer context, such as one that includes data from different sources. For example, modern database systems in genomics (such as those that are being built on top of the GMOD model [9]) allow inclusion of data from gene expression level experiments. We plan to explore this kind of extension to our data model.

Acknowledgements

We are indebted to J. P. Kitajima, who created the preliminary version of the PABdb system that was used for the work described in [12]. The work described in this paper was partially financed by CAPES and the MCT-PRONEX SAI project. JCS and CMBM have CNPq fellowships. The Laboratory of Bioinformatics, in which most of this work was done, is partly funded by FAPESP.

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, (1997) 3389–3402.
2. Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., Wheeler, D.: Genbank. *Nucleic Acids Research* **1**:28 (2000) 15–28.

3. Kovler M.: MySQL. *Apress* (2001).
4. National Center of Biotechnology Information. http://www.ncbi.nlm.nih.gov
5. Perez, M., Carrião, P., Setubal, J. C.: A relational database for COG data. *Proceedings of the Second Brazilian Workshop on Bioinformatics* (2003).
6. Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K, White, O.: The Comprehensive Microbial Resource. *Nucleic Acids Research* **29**:1 (2001) 123–125
7. Riley, M.: Functions of the gene products of *escherichia coli. Microbiology Review*, **57**, (1993) 862–952.
8. Simpson A. J. G., et al. The genome sequence of the plant pathogen *Xylella fastidiosa. Nature,* **406**, (2000) 151–157.
9. Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., Lewis, S.: The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research* **72** (2002) 1599–1610
10. Tatusov, R. L., Galperin, M. Y., Natale D. A., Koonin E. V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**, Number 1 (2000) 33–36
11. The Gene Ontology Consortium: Creating the gene ontology resource: Design and implementation. *Genome Research* **11**, (2001) 1425–1433.
12. Van Sluys, M. A., Monteiro-Vitorello, C. B., Camargo, L. E. A., Menck, C. F. M., da Silva, A. C. R., Ferro, J. A., Oliveira, M. C., Setubal, J. C., Kitajima, J. P., Simpson, A. J.: Comparative genomic analysis of plant-associated bacteria. *Annual Review of Phytopathology* **40**, (2002) 169-189.