PUC

ISSN 0103-9741

Monografias em Ciência da Computação
n° 27/09

A Conceptual Data Model Involving Protein Sets from Complete Genomes: a biological point of view

Cristian Tristão
Antonio Basílio de Miranda
Sergio Lifschitz

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22451-900
RIO DE JANEIRO - BRASIL

Monografias em Ciência da Computação, No. 27/09 ISSN: 0103-9741
Editor: Prof. Carlos José Pereira de Lucena October, 2009

A Conceptual Data Model Involving Protein Sets from Complete Genomes: a biological point of view

Cristian Tristão, Antonio Basílio de Miranda[1], Sergio Lifschitz

[1] FIOCRUZ – Rio de Janeiro, RJ

ctristao@inf.puc-rio.br, antonio@fiocruz.br, sergio@inf.puc-rio.br

Abstract. This work involves the comparison of protein information in a genomic scale. The main goal is to improve the quality and interpretation of biological data, besides our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith-Waterman algorithm in a pair wise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available in the public database RefSeq. Comparisons were run through a computational grid and the complete result reaches a volume of over 900 GB. In this context, the database system design is a critical step in order to extract the expected information from the comparisons' results. This paper describes database conceptual design issues for the creation of a database that represents a dataset of sequence cross-comparisons. We show that our conceptual schema enables users to query from simple to rather complex queries, providing a conceptual framework that can be further implemented in any object-relational database.

Keywords: Sequence Alignment, World Community Grid, Smith-Waterman, Biological Databases, Conceptual Data Model.

Resumo. Este trabalho envolve a comparação de informações de proteínas em uma escala genômica. O principal objetivo é melhorar a qualidade e interpretação de dados biológicos, além da nossa compreensão dos sistemas biológicos e suas interações. As comparações foram obtidas, após a aplicação do algoritmo de Smith-Waterman, realizando-se rigorosas comparações do tipo "par à par" em todas as proteínas preditas codificadas em genomas, incompletos e completamente sequenciados, disponíveis no banco de dados público RefSeq. As comparações foram executados através de um *grid* computacional e o resultado completo atinge um volume superior à 900 GB. Neste contexto, o projeto do sistema de banco de dados é uma etapa crítica, a fim de extrair as informações esperadas a partir dos resultados das comparações. Este artigo descreve os problemas encontrados durante o projeto do banco de dados conceitual para a criação de um banco de dados que representa um conjunto de dados de comparações de seqüências. Nós mostramos que o nosso esquema conceitual permite aos usuários consultar desde simples consultas a counsultas mais complexas, proporcionando um *framework* conceitual que pode futuramente ser implementadas em qualquer banco de dados objeto relacional.

Palavras-chave: Alinhamento de sequencias, World Community Grid, Smith-Waterman, Bancos de Dados biológicos, Modelo Conceitual de Dados.

In charge of publications:

Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC-Rio Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22451-900 Rio de Janeiro RJ Brasil
Tel. +55 21 3527-1516 Fax: +55 21 3527-1530
E-mail: bib-di@inf.puc-rio.br

1. Introduction

The availability of complete genome sequences of numerous organisms, combined with the computational progress occurred in the last few decades provides an opportunity to use holistic approaches in the detailed study of the genome structure, as well as gene prediction and functional classification.

Among these approaches, we are mainly interested in the comparative genome analysis (or comparative genomics), which consists in the analysis and comparison of genetic material from diverse species (or strains), aiming at investigating their internal organization, and evolution of the compared genomes (and the corresponding species). In addition, we are looking forward to revealing the function of genes and non-coding regions in these genomes.

This work reports results of an ongoing research project, called Protein World Database (PWD) [2]. It is an initiative, among Functional Genomics and Bioinformatics Laboratory – Fiocruz [3], World Community Grid [4], and Bioinformatics Laboratory – PUC-Rio [5], dedicated to the comparison of protein information on a genomic scale in order to improve the quality and interpretation of biological data, consequently, our understanding of biological systems and their interactions. Stringent comparisons were obtained after the application of the Smith-Waterman (SW) algorithm [7] in a pair wise manner to all predicted proteins encoded in both completely sequenced and unfinished genomes available in the public database RefSeq [8] (version 21).

Rigorous dynamic programming algorithms, such as SW, ensure the determination of the optimal alignment between pairs of sequences. However, due to their computational complexity, these algorithms are usually not suitable for the comparison of a large set of sequences. Therefore, we have considered for PWD some distributed computing resources provided by the World Community Grid. We have shared computer's idle resources from all over the world to calculate the sequence similarity level among almost 4 million proteins. We have run in our experiments SSEARCH [6], a robust implementation of the SW algorithm.

Once our experiments were finished, the result data available has reached a huge volume of data, over 900GB. Due to its size, the database system used for storage, analysis and management of these data is a fundamental factor to maximize knowledge generation from the results yielded by the PWD. Indeed, among others, one must consider data persistency, high availability and efficient access.

When dealing with VLDBs (very large databases), its conceptual design becomes a relevant step to avoid performance bottlenecks, enable efficient querying and also database maintenance. This paper discusses conceptual modeling issues, and a proposed conceptual database schema for this particular research project that helps with the comprehension of the data involved, and enables a first study about queries and other manipulation.

This paper is organized as follows: Chapter 2 shows an overview about the comparative genome analysis, fundamental knowledge area of this research work. A description about the database conceptual modeling used and objects involved is

presented in Chapter 3. Some additional conceptual design issues, alternatives and modeling decisions adopted in the conceptual schema are described in Chapter 4. Finally, in Chapter 5, conclusions and future works are showed.

2. Background

Comparative genomics comprehends the comparison of two or more genomes, including genomic sequences and also their predicted protein content, the relative positions of their genes (gene order) and other genomic context features that may be of functional (or regulatory) importance. It also includes the study of gene structure and organization, the presence and location of repetitive sequences, polymorphisms and several other characteristics that may help to differentiate genomes [1].

A detailed analysis of the predicted protein contents of an organism is an important step in genome analysis, and it has been applied to several studies with different objectives. Cancer, for instance, is a class of diseases where modifications in the expression pattern of several genes confer new biological properties to the cell. A better understanding of these alterations may provide new insights for the development of diagnostic and treatment procedures.

An important task is the identification of all protein-coding genes and their location in the genome sequence, as well as the characterization of their functions. Genomic sequences are scanned, searching for protein-coding genes, using computational gene models. For each new genome, each predicted gene is conceptually translated into a protein sequence; the predicted collection of protein sequences is the predicted proteome of the organism. Each predicted protein is used as a "query sequence" in similarity searches against repositories of biological sequences. Significant matches are added to the genomic sequence together with the gene position and its product description. More sophisticated methods for the search of gene families are also used for annotation. Collectively, these methods provide predictions for the proteome of a newly sequenced organism [1].

Additional information about a proteome can be obtained through the comparison of the set of protein sequences against itself, which identifies *paralogs* (genes originated after duplication events) through the comparison among different proteomes for the identification of *orthologs* - genes originated after speciation by studying fusion or fission events or new domain arrangements - and by studying the evolution of cellular, metabolic and regulatory functions.

3. Database Conceptual Modeling

In our experiments, a set of 3,812,663 proteins from RefSeq version 21- consisting of all predicted proteins encoded in 458 completely sequenced and unfinished genomes - and 254,609 proteins from Swiss-Prot [9] version 51.5 were compared, in a pair wise manner, with the program SSEARCH. We have configured SSEARCH with standard parameters, and an E-value cut-off equal to one.

For each significant match, a report is generated containing information from sequence identifiers to bit scores and E-values. The output format is given in Figure 1.

A pair of protein sequences satisfying the required conditions to be stored was called a *hit* and defined by a pair of identifiers [query_gi, subject_gi]. The resulting matrix contains only hit information. A hit is defined by identifiers of the two sequences compared (for example, Figure 1 has query_gi = 67523787 e subject_gi = 67540134), and stores the validation measures of the pair wise comparison, besides additional information about the alignment, like similarity and coverage. The alignment itself is not stored.

```
query gi, subject gi, SW score, bit score, e-value, % identity, alignment length, query start, query end, subject start, subject end, query gaps, subject gaps

67523787,67540134,2166,488.8,2.6e-138,0.336,1320,35,1275,67,1367,79,19
```

Figure 1. An example of a report produced for a significant match in PWD. Only the values are actually stored. The upper line presents descriptors of the listed values

Our main problem here was to define a database system that would help us for future querying and general data accesses. The goal is to store the results obtained in such a way that one could use these data together with other external data sources and generate relevant information. However, the whole system must consider the usual impedance mismatch among users offering a simple rather complete way to obtain the required information.

Figure 2 presents a first and basic conceptual schema that can be directly produced from the output results. We have represented it with a conventional Entity-Relationship (ER) diagram, including min-max cardinalities. There are actually 3 possible combinations of hits involving translated ORFs and Proteins. All minimal cardinalities are zero as not all pair wise comparisons generate significant hits.

Results stored at the initial matrix contain only sequence identifiers and alignment information. The first step of this conceptual design was the creation of an entity that characterizes protein sequences. Information about the catalogued proteins compared in PWD includes the protein definition, its length and organism, and possible external references as protein identifiers (RefSeq and/or SwissProt). As the database is kept up to date and updates occur, we identify those proteins that have participated in registered comparisons. These are the main attributes of the *Protein* entity.

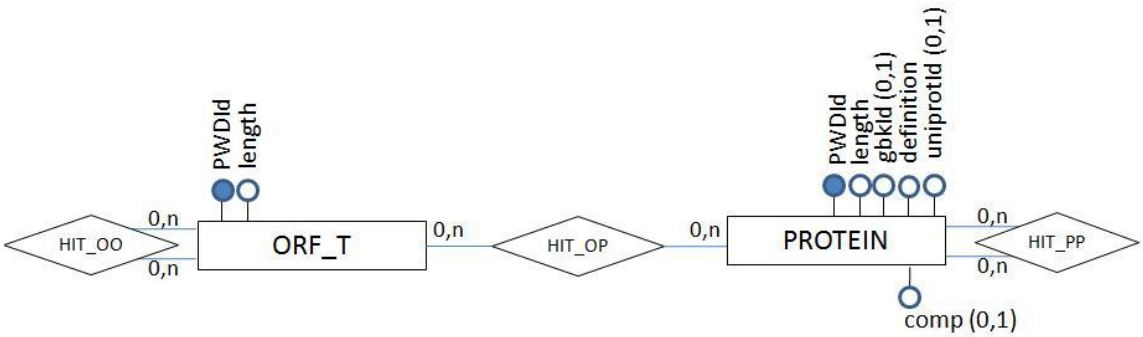

Figure 2. First Approach for a Conceptual Diagram for PWD Project

The amino acid sequences *translated ORF* are represented by another entity – *ORF_T* – because they do not possess previous identifiers. Information about these sequences includes the original organism, location and length. Three types of distinct relationships between hits, proteins and translated ORFs are defined:

1. hit_OO – result of a comparison between translated ORFs;
2. hit_OP – result of the comparison between translated ORFs with proteins derived from SwissProt (proteins derived from RefSeq were not compared with translated ORFs);
3. hit_PP – result of the comparison of RefSeq proteins with RefSeq and SwissProt proteins.

These relationships possess attributes that specify the pair wise results of PWD: *query gi*, *subject gi*, *SW score* (brute score of the comparison), *bit score* (normalized score), *e-value* (alignment significance), *%identity*, *alignment length*, *query start*, *query end*, *subject start*, *subject end*, *query gaps*, *subject gaps*.

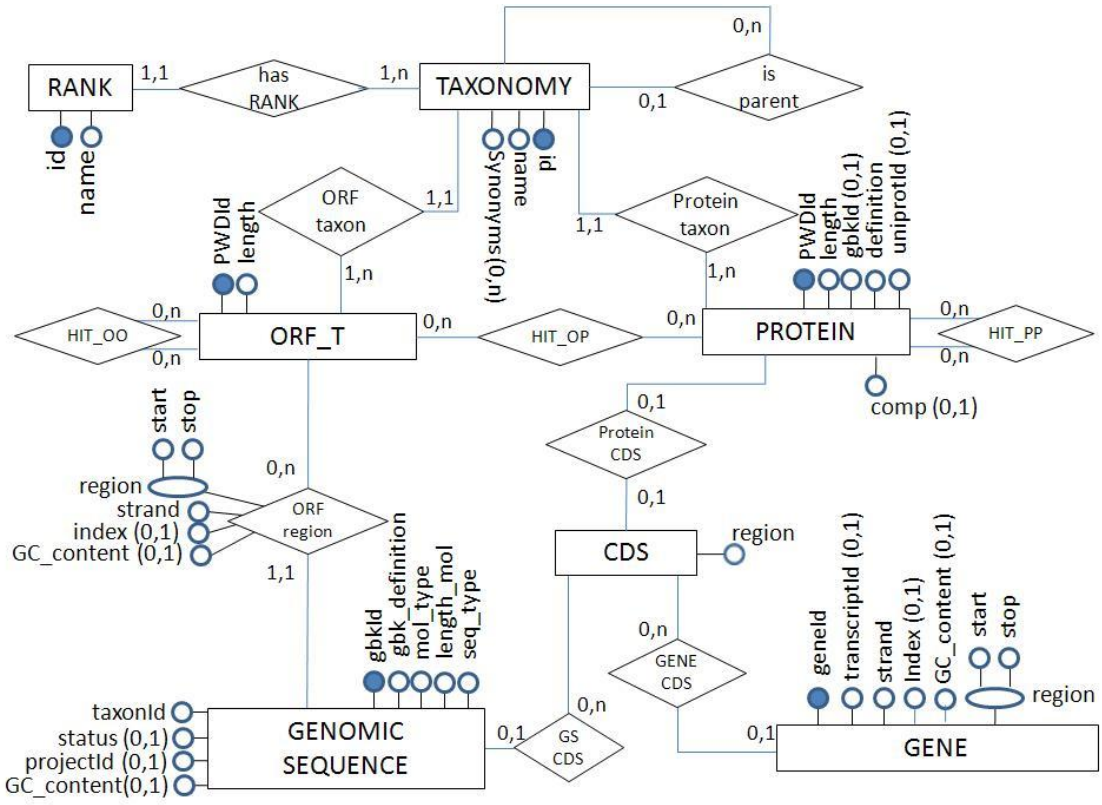

Figure 3. A Conceptual Diagram for PWD Project

However, we have some general and specific goals with this PWD project that cannot be solved only with the SSEARCH results and the corresponding output. There is a need of external data sources if one wants to check on the availability and feasibility. With respect to comparative genomics, hits represent only the result of protein-related genome comparisons. Further interesting questions depend on the protein physical position at its genomic context. For instance, the structure

organization and their genes relative position (gene order), and many other genomic features that may be of functional importance.

Therefore, Figure 3 gives an overview of our particular extended conceptual schema. In what follows, we will discuss some of the data modeling alternatives and our design choices, which have guided us until the final conceptual schema.

We must observe that genes, transcripts, ORFs and genomic sequences are nucleotide sequences, while proteins and translated ORFs are amino acid sequences. The relationships between proteins (independently of their origins) and nucleotide sequences are constructed based on information from RefSeq version 21. Our PWD project will consider recent data sources as RefSeq version 33.

3.1. Conceptual model objects

With Figure 3 in mind, we may explain in more details its entities, relationships and attributes. A protein is generated from a gene, which is a genomic sequence region. A protein coding gene is transcribed and produces a primary transcript that, after some processing, generates a mature transcript containing the protein coding sequence (CDS). The mature transcript is formed by the concatenation of sub-sequences containing information for the protein (exons) and untranslated regions (UTRs).

An ORF is a series of nucleotides extending up to the first termination codon. ORFs may not code for proteins. This way, all coding sequences (CDS) are ORFs but not all ORFs are coding for proteins.

Entity *Protein* represents the amino acid sequence of the protein with the nucleotide sequence of CDS, and CDS with the gene and the genomic sequence containing it, keeping only an external reference to the transcript. Thus, CDS is an entity whose basic property is to keep up with the relationship between entities *Protein*, *Gene* e *Genomic Sequence*. This is done through the positioning of given gene coding regions (exons) in the coordinate system of the genomic sequence that contains it. Each exon in a gene corresponds to a CDS sub-sequence, defined by an initial and a final position mapped into the coordinate system of the genomic sequence.

The nucleotide sequence of a protein coding gene is part of a genomic sequence possessing coding (exons) and non-coding (introns and untranslated regions – UTR) subregions. Reading and transcription of the gene generating the mRNA, which will be further processed and translated into an amino acid sequence, occur in a specific direction in vivo (5′ to 3′). *In silico*, when considered the coordinate system of the genomic sequence containing the gene in question, the reading sense of the gene will be "+" if the gene is represented in the strand with growing coordinates (start < stop), otherwise "-" (or complementar) when the gene is represented in the complementary strand of the genomic sequence (stop < start).

The entity *Gene* possesses also an NCBI identifier – Entrez Gene [10]. The region of its genomic sequence, defined by a start and a stop position, and a reading sense, its order in relation to the other genes in the genomic sequence, a transcript identifier (from RefSeq), and the comparison content. An amino acid sequence ORF_T is analogous, and relates to the genomic nucleotide sequence through an ORF_region

delimited by a start and a stop position inside the genomic sequence, with the RefSeq identifier of the genomic sequence, the reading sense, its position in relation to neighboring genes and the sequence itself.

A genomic nucleotide sequence derived from RefSeq contains the genes (containing CDSs) coding for the amino acid sequence of proteins. These genomic sequences possess a status, property that refers to the current stage of the sequencing project. Possible values for this property are Complete, which typically means that each chromosome is represented by a single scaffold of very high sequence quality; Assembly, which typically means that scaffolds have been constructed that are not yet at the chromosome level and/or are of draft sequence quality; and In Progress, which indicates that either the sequencing project is at the pre-assembly stage or the assembled/completed sequences have not yet been submitted to GenBank/ EMBL/ DDBJ [15].

RefSeq genomic sequences with the prefix NC_ (complete genomic molecules including genomes, chromosomes, organelles, and plasmids) include both automated processing and expert review for some of the records, and the coordinate system, gene positioning and annotation are more stable. Prefixes NT_, NW_, NZ_ (contig or scaffold and unfinished WGS) indicate records that are not individually reviewed; updates are released in bulk for a genome. Assembly, annotations and gene positioning are provisional. These sequences must be differentiated and carefully processed. Relationships *Protein*, *CDSs*, *Gene*, *Genomic Sequence* may be incomplete or even absent.

The *Genomic Sequence* entity possesses a RefSeq identifier, definition and sequence length, the type of organic molecule (DNA/RNA), status, type of sequence (chromosome, organelle, plasmid), an optional identifier of the respective genome project, GC content and an identifier of the original taxon.

4. Additional Conceptual Design Issues

Organism taxonomy is a powerful organizing principle in the study of biological systems. Inheritance, homology by common descent, and the conservation of sequence and structure in the determination of function are all central ideas in biology that are directly related to the evolutionary history of any group of organisms.

The classification used in PWD project is the same as the used by NCBI [11]. Each entry in the NCBI database is a *taxon*, also referred to as a "node". The "root node" (taxid1) is at the top of the hierarchy. The path from the root node to any other particular taxon in this database is called its "lineage"; the collection of all of the nodes beneath any particular taxon is called its "subtree".

In the conceptual model, the organism from which the genomic sequence was derived is the leaf node, defining the sequenced species (or an inferior rank). It contains the taxID identifier from NCBI (a stable unique identifier for each taxon), the scientific and common names and synonyms. Each tree node has a rank , a parent node and may have descendent nodes. The taxonomic lineage may be obtained through a tree traversal from leaf nodes up to the root.

4.1. Limits and Extensions

We have first considered a database system exclusively oriented for the PWD project. Thus, the idea was to consider an entity called Seq_AA that would represent all compared amino acid sequences – including annotated proteins and translated ORFs. This entity would relate with the hit's matrix, and we would be able to specialize Seq_AA with either RefSeq or SwissProt as attributes. This entity would also be limited to sequences compared within the PWD project.

Within this particular conceptual schema, the amino acid sequences would relate with an entity *Coding_region*, and the latter with a nucleotide sequence that contains *Seq_NT source*. The problem of this representation is that it would be artificially adapting a biological concept, as ORFs were considered even if not coding regions. Moreover, another entity, the *Coding_region*, containing *Seq_NT source*, also represented a wrong concept by dealing equally with both genomic and transcript sequences. Furthermore, we have noticed that among the compared protein sequences, there are mRNA and genomic sequences, two different types of nucleotide sequences originally.

We have discussed some alternatives and decided to adopt the conceptual model depicted in Figure 3 due to the following reasons and modeling challenges:

- It is important to enable the database system to support updates as new genome sequences become available. It would be an error to limit the database only to the PWD project.
- The translated ORFs sequences, as a group of artificial (possible) proteins, brought many design problems. It becomes clear now that the sequences do not share the same characteristics with proteins, and need to be represented by and independent entity.
- Even if proteins with different origins could have distinct characteristics, with a conceptual viewpoint they all could be grouped as a single entity - *Protein*.

Among all compared sequences in our project, there are proteins that are originated from genomic projects including gene annotations, mRNA and CDS; those whose origins are only mRNA and proteins that are directly obtained from its sequencing process, without any reference to its original nucleotide. This situation brings another challenge for conceptual modeling and 3 types of amino acid sequences with specific characteristics and distinct research goals may be defined. This issue is better discussed withing a given logical model. We have proposed a relational schema in [17] that explores further this question.

5. Conclusions

We have discussed in this paper the database conceptual modeling used and objects involved to storage PWD comparisons' result. We main goal have been to create a database conceptual model to improve the quality and interpretation of biological data, besides our understanding of biological systems and their interactions.

An extended version of this conceptual model is showed by [13]. It involves the addition of annotation procedures, additional attributes and external data sources, such as Pfam [12] – protein domains -, KEGG [13] – metabolic pathways and controlled vocabulary based on Gene Ontology [14].

We are currently implementing the logical model using PostgreSQL [16] as the underlying DBMS. Many different loading scripts are being developed and will be available to the public. More details of the logical model can be obtained in [17]. As future work we will be able, then, to do a clustering analysis in order to check upon proteins that have presented relevant hits.

References

[1] Mount, D.W.: Bioinformatics: Sequence And Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2004)

[2] ProteinWorld DB, http://bioinfo.pdtis.fiocruz.br/ProteinWorldDB/

[3] Functional Genomics and Bioinformatics Laboratory – Fiocruz, http://www.dbbm.fiocruz.br/labwim/bioinfoteam/index.pl?action=home

[4] World Community Grid, http://www.worldcommunitygrid.org/

[5] Bioinformatics Laboratory - PUC-Rio, http://www.inf.puc-rio.br/~blast/

[6] Pearson W.R.: SSearch. Genomics 11, 635--650 (1991)

[7] Smith T.F., Waterman M.S.: Comparison of Biosequences. Adv. Appl. Math. 2, 482--9 (1981)

[8] NCBI Reference Sequences, http://www.ncbi.nlm.nih.gov/RefSeq/

[9] The UniProt Knowledgebase, http://www.uniprot.org/

[10] NCBI Entrez Gene, www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[11] NCBI Taxonomy Database, http://www.ncbi.nlm.nih.gov/Taxonomy/

[12] The Pfam Protein Families Database, http://pfam.sanger.ac.uk/

[13] KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

[14] The Gene Ontology, http://www.geneontology.org

[15] International Nucleotide Sequence Database Collaboration, http://www.insdc.org/

[16] PostgreSQL, http://www.postgresql.org/

[17] TRISTÃO, C.; LIFSCHITZ, S. Protein world database: geração do esquema lógico e processo de ETL. Technical Report MCC 28/09, Departamento de Informática, PUC-Rio, 2009.