Chapter 7
Database Systems in Biology

Elisa Pappalardo
University of Catania, Italy

Domenico Cantone
University of Catania, Italy

ABSTRACT

The successful sequencing of the genoma of various species leads to a great amount of data that need to be managed and analyzed. With the increasing popularity of high-throughput sequencing technolgies, such data require the design of flexible scalable, efficient algorithms and enterprise data structures to be manipulated by both biologists and computational scientists; this emerging scenario requires flexible, scalable, efficient algorithms and enterprise data structures. This chapter focuses on the design of large scale database-driven applications for genomic and proteomic data; it is largely believed that biological databases are similar to any standard database-drive application; however, a number of different and increasingly complex challenges arises. In particular, while standard databases are used just to manage information, in biology, they represent a main source for further computational analysis, which frequently focuses on the identification of relations and properties of a network of entities. The analysis starts from the first text-based storage approach and ends with new insights on object relational mapping for biological data.

1. INTRODUCTION

In the last few years, a large amount of biological data has been produced thanks to the development and application of new technologies, such as sequencing, microarrays, spectrometry; the vast majority of these data are publicly available through data repository.

Bioinformatics represents an emerging field of the biological sciences, that uses computational power and mathematical structures to manage, analyze and understand biological information to solve biological questions (Hogeweg and Hesper,

DOI: 10.4018/978-1-4666-3946-1.ch007

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

1978). Bioinformatics mainly addresses two problems: organizing data and model inference on the data. Databases store and hold data, while specific algorithms allow one to extract knowledge from annotated information (Rashidi and Buehler, 2000). As for any database applications, the development of biological databases involves several issues, as the decisions about which resources consider and which discard, services to implement, relations between entities, maintenance, and several other related problems. On the other hand, the complex structure of the biological information enclosed in databases needs a proper model to be extracted.

Therefore, the collaboration of biologists and IT professionals is needed to face the problem of implementing successful databases to extract meaningful information.

In this review chapter, we introduce some design principles for the effective design of biological databases. Subsequently, we present the current available data resources available and, finally, we outline some conclusions.

2. BIOLOGICAL DATABASES DESIGN PRINCIPLES

Designing a database is a complex task, which involves many related and conflicting aspects. The design and development of biological databases is not very different from the development of the "traditional" ones, as business or government databases: the first step is understanding the information to store in the database, and then translating them into a robust framework, debugging and maintaining the system (Birney and Clamp, 2004). Of course, this is not a simple task, since modeling a database structure requires a careful and complex analysis, and involves many problems.

2.1. Data Conservancy

One of the basilar problems encountered when designing biological databases concerns the data to store: data can be integrated across a unique species, or refer to multiple species (Paton and Goble, 2001). The first category of data is collected in deep databases; some examples are FlyBase (Drysdale and Crosby, 2005), that collects genetic and molecular data of the insect family Drosophilidae, the Mouse Genome Database (MGD) (Blake et al., 2003), and SCPD (Zhu and Zhang, 1999), SGD (Christie et al. 2004), YEASTRACT (Teixeira et al. 2006), YPD (Hodges et al. 1998) for Saccharomyces cerevisiae. While deep databases collect data of a single species, broad databases store biological information across multiple species; these data pertain to nucleotide sequences as in EMBL (Stoesser et al. 1999) and GenBank (Benson et al. 1998); protein sequences, as in SWISS-PROT (Bairoch and Boeckmann, 1992), MIPS (Mewes et al. 1997), PDB (Berman et al. 2002); pattern sequences, that represent the pattern associated with the alignments of the sequences, as in InterPro (Apweiler et al. 2001), PROSITE (Falquet et al. 2002), GELBANK (Babnigg and Giometti, 2004), collection of proteins, nucleotide sequences as in NCBI (Pruitt et al. 2006); genomic information, as transcriptome (Cahoy et al. 2008) and pathways, in MPW (Selkov et al. 1998), PUMA2 (Maltsev et al. 2006), LIGAND (Goto et al. 2002).

A further classification distinguishes databases according to the source of data: primary databases contain biological information that are directly derived from experiments, as in EMBL (Stoesser et al. 1999) and GenBank (Benson et al. 1998), while secondary databases collect data derived from other storage sources, as repositories, different databases, analysis, i.e. in Swiss-Prot (Bairoch and Boeckmann, 1992) and ENZYME (Bairoch, 2000).

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/database-systems-biology/77962

Related Content

Impact of Digitalization on the Efficiency of Supply Chain Management in the Digital Economy
Anna Mikhaylova, Tatyana Sakulyeva, Tamara Shcherbina, Natalia Levoshichand Yuri Truntsevsky (2021). *International Journal of Enterprise Information Systems (pp. 34-46).*
www.irma-international.org/article/impact-of-digitalization-on-the-efficiency-of-supply-chain-management-in-the-digital-economy/282016

The Influence of Organisational Size, Internal IT Capabilities, and Competitive and Vendor Pressures on ERP Adoption in SMEs
Cliff Cartmanand Angel Salazar (2011). *International Journal of Enterprise Information Systems (pp. 68-92).*
www.irma-international.org/article/influence-organisational-size-internal-capabilities/58047

Improving Stakeholder Communications and IT Engagement: A Case Study Perspective
G. Verley (2007). *Handbook of Enterprise Systems Architecture in Practice (pp. 160-171).*
www.irma-international.org/chapter/improving-stakeholder-communications-engagement/19423

Pricing Outcomes in Dual-Channel Monopoly and Partial Duopoly
Farooq M. Sheikh, M. Ruhul Aminand Nafeez Amin (2007). *Modelling and Analysis of Enterprise Information Systems (pp. 134-149).*
www.irma-international.org/chapter/pricing-outcomes-dual-channel-monopoly/26846

A Systemic View on Enterprise Architecture Management: State-of-the-Art and Outline of a Building Block-Based Approach to Design Organization-Specific Enterprise Architecture Management Functions
Sabine Buckland Christian M. Schweda (2014). *A Systemic Perspective to Managing Complexity with Enterprise Architecture (pp. 237-254).*
www.irma-international.org/chapter/a-systemic-view-on-enterprise-architecture-management/80913