EMBnet.journal
Bioinformatics in Action

Romano (2017) EMBnet.journal 23, e884
http://dx.doi.org/10.14806/ej.23.0.884

Reports

Bioinformatics activities at the University Hospital San Martino IST in Genoa

Paolo Romano [1]

[1] IRCCS AOU San Martino IST, Genoa, Italy.

Abstract

This report summarises some of the bioinformatics activities that have been carried out since 1986 at the National Cancer Research Institute of Genoa, now University Hospital IRCCS San Martino IST. Two main interrelated research lines are highlighted: data management for biological resources, and automation of data retrieval and analysis. As developments in Information and Communication Technologies (ICTs) are fundamental to bioinformatics, the NETTAB workshops, which are devoted to the analysis of the impact of new ICTs on bioinformatics research, are also presented.

Introduction

Bioinformatics is still a relatively young discipline. In the '80s, biology and ICT were completely different from today. When the EMBL Nucleotide Sequence Data Library was first established, in 1980, the challenges mainly related to establishing the first databases and developing sequence comparison algorithms.

Many algorithms and software tools, which are now the basis for every molecular data analysis, have since been developed; meanwhile, data availability has been increasing at an impressive speed, thanks to the advent of high-throughput technologies and '*omics*' projects. At the same time, we have moved from local elaboration of data to remote data analysis.

These transformations have required the development and implementation of new tools for remote processing and data sharing. Hence, the focus nowadays has shifted to the integration and analysis of an unprecedented amount of information, aiming to build an interoperable, semantically aware, social and collaboratively-based network environment for bioinformatics.

In this short report, I summarise the main activities of the National Cancer Research Institute of Genoa, now IRCCS AOU San Martino IST, since the '80s, following, and sometimes anticipating, the evolution of bioinformatics tools and databases.

Biological resource data management

The term 'biological resource' is applied to living biological material collected and held in culture collections: bacterial and fungal cultures; animal, human and plant cells; viruses or isolated genetic material. A wealth of information about biological resources has been accumulated in Biological Resource Centres (BRCs) and, although dispersed, a large part of this information is still accessible. Various coordinated efforts have been put into making this information jointly available on-line. Many more improvements can be achieved by adopting innovative ICTs to deepen integration of this information in the bioinformatics network environment.

Automation of data retrieval and analysis

In biology, data integration is limited by the great number of available resources, their size and frequency of updates, their heterogeneity and distribution on different servers. Integration of these data can therefore be achieved only by adopting flexible and extensible tools. XML, Web Services (WSs) and Workflow Management Systems (WMSs) can support the creation and deployment of software able to automate data retrieval and analysis.
A WMS is able to design and create workflows, and to manage their execution. Its main components are i) a graphical interface for composing workflows, entering data and displaying different types of results; ii) an archive to store workflow descriptions, as well as results of executions and related traces; iii) a scheduler able to invoke services when needed; iv) a registry of available services; v) Application Programming Interfaces (APIs) for interoperating with services; and vi) a monitor tool to control workflow execution.

For this to happen, a 'technology-savvy' status must be achieved by providers and users. In this status, databases adhere to standards, and include semantic metadata; software is distributed on the network and can interoperate; and data-analysis procedures can be carried out on the network. A shared methodology for software development should also be adopted by developers and service providers. This could include

Article history
Received: 23 February 2017
Published: 10 April 2017

Romano (2017) EMBnet.journal 23, e884
http://dx.doi.org/10.14806/ej.23.0.884

i) XML schemas for creating shared data models; ii) XML-based languages for data storage and exchange; iii) WSs for software interoperability; iv) ontologies for WS discovery, selection and interoperation; and v) workflows for executing analysis processes.

The Interlab project

The Interlab Project was funded in 1989. One of its goals was to implement data-banks of biological resources. The Cell Line Data Base (CLDB), which collected and made available information on cell lines, the B Line Data Base (BLDB), which stored information relating to HLA typed B lymphoblastoid cell lines, and the Molecular Probe Data Base (MPDB), containing data on oligonucleotides, were built in that context. The databases were first made available on-line through packet-switching data networks (ITAPAC in Italy). Researchers could connect by means of personal computers equipped with standard modems. Dec VT100 terminal emulation was required in order to exploit the interface. Later, new interfaces were built using the Wide Area Information Servers (WAIS) technology, the Gopher system and, with the advent of the World-Wide Web concept, through Web servers. A new hypertext interface was developed for the CLDB – HyperCLDB – to allow effective indexing of its contents by search engines like Google and Yahoo (Romano et al., 2009). HyperCLDB[1] consists of many pages (currently, ~8,750), including detailed descriptions of cell lines and indexes of their features. In each page, hyperlinks are added to connected pages and facilitate navigation.

Common Access to Biological Resources and Information (CABRI)

The Common Access to Biological Resources and Information (CABRI) project was funded by the European Union from 1996 to 1999 (Romano *et al.*, 2005). Among its objectives, it aimed to ease access to information in biological resource catalogues. CABRI[2] is based on the Sequence Retrieval Software (SRS), a search engine designed for integrated queries of molecular biology databases. With SRS, data must reside locally and be stored in 'flat files' (text-only files) with pre-defined, shared syntaxes. Both explicit and implicit links between databases can be defined. At the time, SRS was a good option for making integrated searches of databases with similar contents and interlinks in local environments. CABRI catalogues were implemented in SRS by comparing the data structures of collections' databases, and then defining three shared data-sets for each material: the Minimum Data Set (MDS) includes information needed to uniquely identify a resource; the Recommended Data Set (RDS) includes information useful to achieve an improved description of the characteristics, functions and properties of a resource; and the Full Data Set (FDS) includes all available information related to a resource. Data-input procedures were defined for each item of the MDS and RDS: they provide a textual description of its contents and specify the input process for the corresponding values. CABRI currently includes 28 collections that can be searched either via a simplified interface or the standard SRS interface.

Microbial Resource Research Infrastructure (MIRRI)

The European Microbial Resource Research Infrastructure (MIRRI) project can be considered an evolution of CABRI. One of its main objectives is to provide access to information available in the European collections of microorganisms through a dynamic Information System. The MIRRI-IS should include a repository for BRC catalogues, a tight interconnection with domain information systems, a unique portal for catalogues and associated data, and an interoperable system based on APIs and workflows. Three demonstration systems were developed in the MIRRI preparatory phase: the BacDive demonstrator aims to extend the contents of catalogues with a greater number of well-defined data; the StrainInfo demonstrator is targeted towards a better integration among collection catalogues through the identification of common strains; and the USMI Galaxy demonstrator aims both to support data curation and to integrate catalogues with external resources. A five-year plan for the implementation of the MIRRI-IS has been defined (Romano *et al.*, 2017).

IST Bioinformatics Web Services (IBWS)

A suite of WSs – the IST Bioinformatics Web Services (IBWS) (Zappa *et al.*, 2010) – was developed to make databases available at the IRCCS San Martino IST accessible through standard APIs. IBWS has been developed by using standard tools, and should be easy to invoke by any compliant software, such as Taverna. The main advantage offered by IBWS relates to the possibility of accessing a set of unique archives, which otherwise could only be queried manually, through standard APIs.

Bioinformatics Web Enactment Portal (BioWEP)

The use of WMSs can be difficult for the majority of biologists. Web portals can allow users to enact useful workflows in a friendly environment. The Bioinformatics Web Enactment Portal (Biowep) is a Web application that allows selection and execution of pre-defined, annotated workflows (Romano *et al.*, 2007). It is based on a server-side implementation of the Taverna

[1] http://bioinformatics.hsanmartino.it/hypercldb/
[2] http://www.cabri.org/

enactor. Workflow annotation, which is achieved via an ontology of bioinformatics tasks and data-types, involves registration of the data-types for the main components of the workflow. Users can then select workflows of interest on the basis of their annotation.

NETTAB Workshops series

Continuous monitoring of technological developments, and of their impact on biological research, is needed in order to promote swift adoption of the most promising and innovative bioinformatics tools. This is the objective of the NETTAB Workshops.

NETTAB[3] Workshops are a series of International meetings on "Network Tools and Applications in Biology", held annually in Italy. They aim to introduce participants to the most innovative ICTs, and provide a unique forum for bringing together biologists and bioinformaticians with computer science experts. Workshops include sessions devoted to tools, systems, platforms and early applications of relevant technologies. Keynote lectures and selected presentations are included in the programme, alongside poster sessions and tutorials.

Because of the continuous technological evolution, the workshops focus each year on a different technology or domain. Since 2001, many themes have been discussed, including standardisation for data integration (Genoa, 2001), multi-agent systems (Bologna, 2002), scientific workflows (Naples, 2005), Web Services (Santa Margherita di Pula, 2006), Semantic Web (Pisa, 2007), collaborative research (Catania, 2009), wikis (Naples, 2010), social and mobile applications (Venice, 2013), and reproducibility (Rome, 2016).

[3] http://www.nettab.org/

References

1. Romano P, Kracht M, Manniello MA, Stegehuis G, Fritze D. (2005) The role of informatics in the coordinated management of biological resources collections, *Applied Bioinformatics*. 2005, **4**(3):175-86. http://dx.doi.org/10.2165/00822942-200594030-00002
2. Romano P, Bartocci E, Bertolini G, De Paoli F, Marra D et al. (2007) Biowep: a workflow enactment portal for bioinformatics applications. *BMC Bioinformatics* 2007, **8**(Suppl 1):S19. http://dx.doi.org/10.1186/1471-2105-8-S1-S19
3. Romano P, Manniello A, Aresu O, Armento M, Cesaro M et al. (2009) Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research*, **37**(Database issue):D925-D932. http://dx.doi.org/10.1093/nar/gkn730
4. Romano P, Smith D, Bunk B, Vasilenko A, Glöckner FO. Designing the MIRRI information system. *PeerJ Preprints*. Submitted on February 17, 2017. https://peerj.com/preprints/2815/
5. Zappa A, Miele M, Romano P. (2010) IBWS: IST Bioinformatics Web Services. *Nucleic Acids Research*, **38**(Web Server issue):W712-W718. http://dx.doi.org/10.1093/nar/gkq416