Integrated computer environments for exploratory genomics

François RECHENMANN
INRIA Rhône-Alpes
http://www.inrialpes.fr/helix/

Nearly every paper or talk on bioinformatics begins with an apocalyptic description of the steady growth of the volume of genomic and post-genomic data. And it is true that technological progresses over the last 15 years, or so, have opened the way to the systematic investigation of biological processes at the molecular and cellular levels.

Since the first sequence of the genome of a living organism in 1995, the bacterium *H. influenzae*, dozens of bacterial genomes have been published, together with several eukaryotic genomes.

DNA chips and other similar devices allow to measure the expression level of several thousands genes simultaneously. Similarly, advanced spectrometry techniques lead to high throughput protein detection and identification.

The DNA sequencing and gene expression measuring techniques are now routinely used and produce high volumes of data that accumulate in private and public databases, the exponential growth of which is repeatedly underlined. Storing data is however the least problem set to biologists by these methodological upheavals. Analysing the data to turn them into new biological knowledge is the very challenge.

Identifying genes and their components within a DNA sequence is already a problem which, despite intense research efforts, has not yet found a reliable algorithmic solution, at least for eukaryotic genomes. Understanding the functions of the products of these genes is a problem for which the solution lies farther away. The classical approach is presently to scan the databases and look for genes whose sequence is similar to the sequence of the gene under study. The idea is then to attribute this gene the functions that are attached to the genes which have been retrieved on the basis of their sequence similarity. The drawbacks and the limitations of this approach are obvious.

Multiplying the ways genes may be related one to the others is probably a promising lead to follow in this quest for gene functions. A gene may thus be declared to be related to another one on the basis of their sequence similarity, but also on the basis of some physical or chemical properties shared by the proteins they code for, or because they code for enzymes which are involved in the same metabolic pathway, and so on. Clearly, this way of reasoning about gene functions requires the availability of heterogeneous data from multiple sources and puts therefore the emphasis on a fundamental problem set to bioinformatics: how to organise the genomic and post-genomic data in order to make more easily appear correlations, set intersections and any regularity which might produce fruitful hints?

In the present state of data organisation, a biologist who wants to relate heterogeneous data has to browse among the various databases that are accessible *via* Internet[1]. The diversity of these databases sets several problems: i) their structures and formats are different and transformation programmes (called *scripts*) have to be specifically written; ii) the semantics of a database conceptual scheme is usually not made explicit, so that one cannot be sure that a term such as "gene" in a database covers the same concept in another one. To overcome these syntaxic and semantic interoperability problems, which are severe obstacles in the data analysis processes, two classes of solutions are being experimented in order to develop and to maintain a consistent data set according to a pre-defined problematics. The first one consists in copying and importing in a database all the data that are relevant to the problematics; the second one advocates the development of a virtual database that provides a unified up-to-date view over a set of heterogeneous databases.

Accessing the data is only one part of the problem. The biologist needs also to apply on these data the adequate analysis methods. Research in bioinformatics has produced a widening set of powerful algorithms and most of the programmes are freely available *via* Internet. The first issue is therefore methodological: How to select the methods that suit the data and the objective of the analysis? How to set the correct values to the parameters? How to evaluate the relevance of the results? The second issue is technical: the data formats expected by a programme may not be the formats in which the data at hand are presently encoded. This is another aspect of the interoperability problem and the biologist who plans to apply a series of methods on a dataset has to spend quite a lot of time in writing scripts to transform one format into another.

Turning genomic and post-genomic data into biological knowledge is a highly exploratory process which relies on the cognitive abilities of the biologists and the processing power of bioinformatics methods. Whereas data and methods are fully available, the interoperability problems slow down this inductive process and result in a waste of useful time.

The solution is to design integrated bioinformatics environments in which all the data are represented in a consistent way, using high-level data and knowledge models, and all the methods are explicitly described through their input and output so that the system can automatically solve any mismatch between the expected and the actual data formats.

GenoStar is such an integrated environment for exploratory genomics. It was developed by a consortium of four members: INRIA, Institut Pasteur, Hybrigenics and GENOME express. The consortium has been partly supported by the French Ministry for Research (*Programme Génomique* and *Programme GenHomme*).

GenoStar is made up of several application modules which share data and knowledge management facilities. All the data that are manipulated by the application modules, and all the results they produce, are explicitly represented in an entity-relationship model. Within a module, the methods are organized into strategies, the execution of which addresses complex analysis problems.

[1] Every year, the January issue of the journal *Nucleic Acids Research* gives an excellent overview of the diversity of databases on the web: http://www3.oup.co.uk/nar/database/c/.

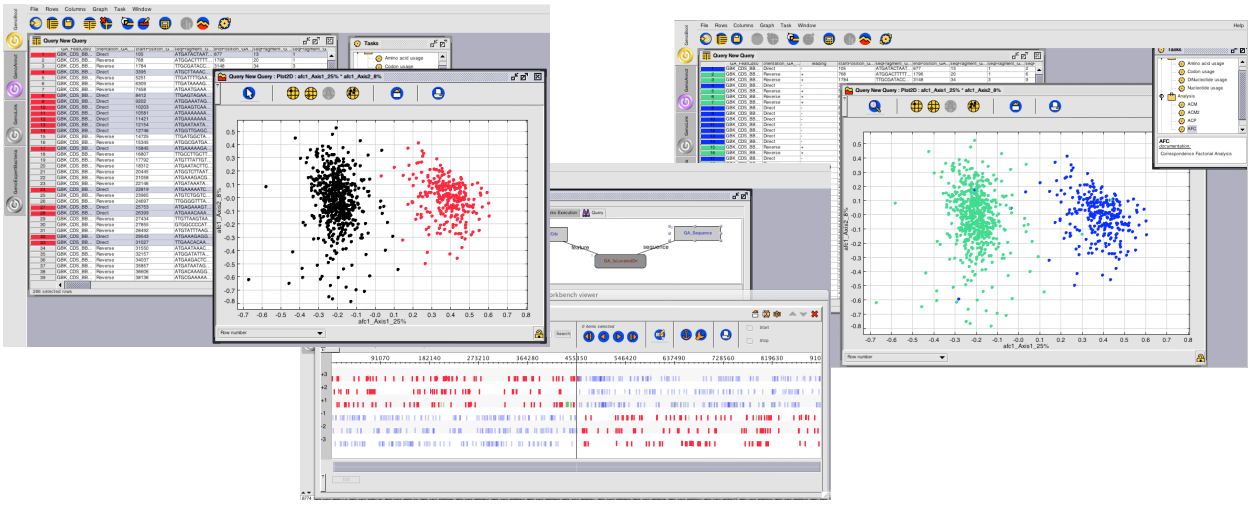
The present version of GenoStar is made up of three modules: GenoAnnot, GenoLink and GenoBool, which can easily exchange data.

GenoAnnot relies on several sequence analysis methods to perform the syntaxic annotation of bacterial genomes. It produces predictions on the position of genes and of other pertinent features. Conflicting and convergent predictions can be displayed with the help of a map viewer.

By allowing biologists to browse through a network of biological entities and bioinformatics objects, GenoLink help them to characterize the function of genes. The links of a network represent different relationships between the entities and the set of their types is easily extendable. A powerful request engine allows to explore the network and to draw meaningful inferences.

GenoBool offers several data analysis methods which can be applied to heterogeneous sets of data after they have been adequately coded. GenoBool thus allows the user to discover new relationships between properties of biological entities. Once the pertinence of a relationship has been asserted, it may lead to the introduction of a new kind of link in GenoLink.

However, the efficiency of GenoStar for genomic data analysis goes far beyond the individual capabilities of these three modules. Its very architecture offers indeed several original features which help the biologist in performing complex and exploratory analysis tasks. Since the modules and the methods they offer are completely interoperable, it is easy to send the results obtained in one module to another one which will compute an alternative view on the data and thus give a further clue for their interpretation. The following figure gives an excellent illustration of these capabilities.



On the basis of the way their sequences make use of the genetic code, the set of genes of the bacterium *B. burdorferi* has been partitioned into two clusters with the help of a classification method of the GenoBool module (left screen shot). Coming back to the GenoAnnot module, which produced the prediction of the gene locations and therefore their sequences, the user will see (centre) that the genes of a same cluster tend to be located either on the leading or on the lagging strand. He may then formulate an hypothesis, validate it through an explicit request on the set of genes (right screen shot), and search for a biological explanation to this surprising correlation between a physico-chemical property and the codon usage (this correlation has been discovered in 1998 by James O. McInerney: "Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*" PNAS 1998; 95: 10698-10703).