We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

10

Knowledge Based Expert Systems in Bioinformatics

Mohamed Radhouene Aniba and Julie D. Thompson
Laboratoire de Biologie et Génomique Intégrative, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)
France

1. Introduction

The recent revolution in genomics and bioinformatics has taken the world by storm. From company boardrooms to political summits, the issues surrounding the human genome, including the analysis of genetic variation, access to genetic information and the privacy of the individual have fuelled public debate and extended way beyond the scientific and technical literature. During the past few years, bioinformatics has become one of the most highly visible fields of modern science. Yet, this 'new' field has a long history, starting with the triumphs of molecular genetics and cell biology of the last century, where bioinformatics was used for the computational treatment and processing of molecular and genetic data.
Despite its widespread use, no single standard definition exists to describe bioinformatics. From the biologist's point of view, it is generally considered to be the use of computational methods and tools to handle large amounts of data and the application of information science principles and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. On the other hand, a computational scientist will generally define bioinformatics as a direct application area of existing algorithms and tools and the use of mathematical and computational approaches to address theoretical and experimental questions in biology.
In July 2000, the NIH (National Institute of Health) released a working definition of bioinformatics as the research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyze, or visualize such data, where as computational biology was defined as the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems.

2. « Omics » science and bioinformatics

In the last decade, the high throughput genome sequencing techniques and other large scale experimental protocols, have led not only to an exponential increase in the amount of biological data, but also to the diversification of molecular biology data, leading to a new biological research paradigm, one that is data-rich and data-driven. Many of the emerging areas of large-scale biology are designated by adding the suffix '-omics' to existing terms.

Source: Expert Systems, Book edited by: Petrică Vizureanu,
ISBN 978-953-307-032-2, pp. 238, January 2010, INTECH, Croatia, downloaded from SCIYO.COM

The most widely known omics sciences are genomics (the quantitative study of genes, regulatory and non-coding sequences), transcriptomics (RNA and gene expression), proteomics (proteins and their expression) and metabolomics (metabolites and metabolic networks). The importance to the life-science community as a whole of such large-scale approaches is reflected in the huge number of citations to many of the key papers in these areas, the human and mouse genome papers being the most obvious examples.
In the context of this huge flood of data, complex data management and integration systems are now being introduced to collect, store and curate all this heterogeneous information in ways that will allow its efficient retrieval and exploitation. These developments are opening up new possibilities for large scale bioinformatics projects, aimed at understanding how genetic data is translated into molecules, networks and pathways, all the way to physiology and even ecological systems.

3. Current challenges

The development of high-throughput biotechnologies and the subsequent omics studies is paving the way to exciting new routes of scientific exploration. Instead of being restricted to the analysis of a handful of genes or proteins per experiment, whole genomes and proteomes (the complete set of proteins encoded by an organism's genome) can be analyzed today. This allows biologists, with the help of bioinformaticians, to explore more complicated processes than were possible before (Carroll et al, 2006; Lein et al, 2007; Souchelnytskyi, 2005; Spellman et al, 1998; van Steensel, 2005). Nevertheless, the new data poses as many problems as it does opportunities. An obvious example is the human genome project where, once the technological limitations were lifted and the DNA sequence was obtained, the bottleneck rapidly shifted to its annotation, i.e. the identification of the genes encoded by the genome and their functions. As a consequence, in a similar way to the new biotechnologies, large bioinformatics projects have been initiated, with numerous research groups collaborating to tackle complex issues, such as the detailed annotation of the complete human genome (The ENCODE Project Consortium, 2004).
Thus, new layers of bioinformatics annotations and predictions are laid on top of the experimental omics data and are made available progressively through public web-accessible data stores, such as Ensembl (www.ensembl.org) or the UCSC Genome Browser (genome.ucsc.edu). The fact that the data is broadcast via the web leads to new issues of data management, maintenance and usage. Easy access is a crucial factor that will allow biologists to use the data as a rich source of information for *in silico* data integration experiments. Nevertheless, accessing these heterogeneous data sets across different databases is technically quite difficult, because one must find a way to extract information from a variety of search interfaces, web pages and APIs. To complicate matters, some databases periodically change their export formats, effectively breaking the tools that allow access to their data. At the same time, most omics databases do not yet provide computer-readable metadata and, when they do, it is not in a standard format. Hence, expert domain-specific knowledge is needed to understand what the data actually represents, before it can be used efficiently.
To further complicate matters, today's bioinformatics analyses require a combination of experimental, theoretical and computational approaches and data must be integrated from a large variety of different data resources, including genomic sequences, 3D structures, cellular localisations, phenotypes and other types of biologically pertinent data. However,

several problems related to the 'omics' data have been highlighted. For example, data emerging from 'omic' approaches are noisy (data can be missing due to false negatives, and data can be misleading due to false positives) and it has been proposed that some of the limitations can be overcome by comparing and integrating data obtained from two or more distinct approaches (Ge et al, 2003). In this context, a major challenge for bioinformaticians in the post-genomic era is clearly the collection, validation and analysis of this mass of experimental and predicted data, in order to identify pertinent biological patterns and to extract the hidden knowledge (Roos, 2001).

Important research efforts are now underway to address the problems of data collection and storage. One approach has been data warehousing, where all the relevant databases are stored locally in a unified format and mined through a uniform interface. SRS (Etzold et al, 1996) and Entrez (Sculer et al, 1996) are probably the most widely used database query and navigation systems for the life science community. Alternatively, broadcast systems implement software to access mixed databases that are dispersed over the internet and provide a query facility to access the data. Examples include IBM's DiscoveryLink (Haas et al, 2001), BioMOBY (Wilkinson et al, 2003). More recently, semantic web based methods have been introduced that are designed to add meaning to the raw data by using formal descriptions of the concepts, terms, and relationships encoded within the data. Many of these technologies are reviewed in more detail in (Romano, 2008).

Today's information-rich environment has also led to the growth of numerous software tools, designed to analyse and visualize the data. The tools can be merged using pipelines, or more recently workflow management systems (WMS), to provide powerful computational platforms for performing *in silico* experiments (Halling-Brown et al, 2008). However, the complexity and diversity of the available analysis tools mean that we now require automatic processing by 'intelligent' computer systems, capable of automatically selecting the most appropriate tools for a given task. In particular, one major insight gained from early work in intelligent problem solving and decision making was the importance of domain-specific knowledge.

Thus the field of bioinformatics has reached the end of its first phase, where it was mainly inspired by computer science and computational statistics. The motivation behind this chapter is to characterize the principles that may underlie the second phase of bioinformatics, incorporating artificial intelligence techniques. In this context, knowledge-based expert systems represent an ideal tool for an emerging research field, known as 'integrative systems biology'.

4. Expert systems in bioinformatics

Expert systems have been used in bioinformatics for many years, although some past and current projects have incorporated similar technologies without actually employing the term 'expert system'. The domains covered range from fundamental computational biology studies to medical research, forensic sciences and environmental protection research.

In this section we will describe some examples of expert system applications, highlighting the importance of knowledge based architectures and their impact on advanced research.

4.1 Medical diagnostics

One of the earliest direct applications of expert systems in a biological discipline was in medicine. This is a very data-rich domain and knowledge based expert systems (KBS)

quickly became an essential tool for diagnostics and personalized treatments. KBS are widely used in domains where knowledge is more prevalent than data and that require heuristics and reasoning logic to derive new knowledge. The knowledge in a KBS is stored in a knowledge base that is separate from the control and inference programs and can be represented by various formalisms, such as frames, Bayesian networks, production rules, etc. In the medical field, a combination of domain knowledge and data are used for the detection, diagnosis, (interpretation) and treatment of diseases. Depending on the problem, the balance between data and knowledge varies and appropriate systems are identified and deployed, including knowledge biased computing models such as rule-based reasoning (RBR), model-based reasoning (MBR) and case-based reasoning (CBR).

In RBR, the knowledge is represented by symbolic rules (Ligeza, 2006) and inference in the system is performed by a process of chaining through rules recursively, either by reverse or forward reasoning (Patterson, 1990). In MBR, the knowledge base is represented as a set of models (satisfying assignments, examples) of the world rather than a logical formula describing it. In CBR, the knowledge is stored as a list of cases, where a case consists of a problem, its solution, and some information about how the solution was obtained. New problems are then solved by reusing the past cases.

Each of these approaches has advantages and disadvantages and as a result, many systems in the medical domain use a combined approach. One example is the BOLERO expert system (Lopez et al, 1997), which uses CBR to improve a RBR medical diagnosis based on the information available about the patient. The MIKAS system (Khan et al, 1997) also integrates CBR and RBR, with the goal of automatically providing a diet recommendation that is strongly tailored to the individual requirements and food preferences of a patient. Other systems incorporate a mixture of MBR and CBR, such as PROTOS (Porter et al, 1986), which uses knowledge acquisition for heuristic classifications in medical audiology, or CASEY (Koton, 1988) which uses CBR to improve the computational efficiency of a causal model of heart disease. Finally, T-IDDM (Montani et al, 2003) is a multi-modal reasoning system providing an accurate decision support tool for the diagnosis of type 1 diabetes, based on a mixture of RBR, MBR and CBR.

In cases where the translation of implicit knowledge into explicit rules is problematic, alternatives to the reasoning systems described above are intelligent computing systems (ICS) or statistical inference, such as that based on Bayes' theorem, which assigns a probability to each advised output (equivalent to a disease in the medical domain). Statistics-based approaches have been exploited in expert systems for the diagnosis and management of pacemaker-related problems (Bernstein et al, 1995) or other medical predictions (Dragulescu et al, 1995). While this type of expert system is suitable for reciprocally exclusive diseases and independent symptoms, they fail when some symptoms have the same cause (are connected) or when a patient can suffer from more than one disease. In this context, artificial neural networks (ANN) have been developed. ANNs have been widely utilized and are an accepted method for the diagnosis of data intensive applications. (Grossi et al, 2007) showed that ANN can improve the classification accuracy and survival prediction of a number of gastrointestinal diseases. ANN has also been used in many other fields, including radiology, urology, laboratory medicine and cardiology (Itchhaporia et al, 1996). Finally, ICS, like KBS, often rely on combined approaches or hybrid expert systems (HES). For example, (Brasil et al, 2001) developed a HES for the diagnosis of epileptic crises, where some of the problems inherent to ANNs were resolved using Genetic Algorithms.

4.2 DNA sequence analysis: Forensic science

An interesting application of expert systems is the analysis of DNA sequences in forensic science. We have all heard about forensics thanks to Hollywood movies and TV series, where a scientist simply inserts a tube into a machine and immediately recognizes the suspect. A lot of us think that this is somewhat exaggerated and unrealistic. Well this is not the case. Forensics is a very developed science and due to its importance, governments devote large budgets to research in this domain.

The processing of forensic DNA samples and the interpretation of DNA profile data is complex and requires important resources both in terms of equipment and in highly trained personnel. But the growth of robotic equipment to automate the extraction of DNA from forensic samples, to quantify and amplify the samples, together with multi-capillary electrophoresis instrumentation has shifted the emphasis to the data analysis stage. Traditionally, the analysis and interpretation of DNA profile data was performed manually by at least two independent highly trained, experienced human scientists. However, this is a time-consuming process and in recent years, DNA profiling interpretation has been automated by replacing the human workers with bioinformatics software and notably, expert systems.

The analysis begins with a sample of an individual's DNA, which is then processed to create a sample DNA profile. This stage is performed by two software packages: GeneScan and Genotyper (Applied Biosystems, Foster City, CA, USA), and includes a manual review step. The subsequent interpretation of the DNA profile involves the comparison of the sample DNA profile with another sample or a database to determine whether there is a genetic match. The interpretation is not a simple process since extremely high accuracy and consistency of forensic evidence is clearly necessary. Therefore, a hierarchy of decision rules has been produced to deal with a certain number of biological and technological artefacts and noise in the data. Once the artefacts have been removed, all remaining peaks are assumed to be "real" and can be assigned either to the individual in the case of single donor samples or can be treated by another set of complex decision rules, in the case of mixed donor samples.

A number of knowledge based expert systems have been developed in order to automate this part of the DNA analysis as much as possible, thus reducing the amount of time taken to analyse an important number of DNA profiles, e.g. GeneMapper ID (Applied Biosystems, Foster City, CA, USA), FaSTR DNA (Power et al, 2008) or FSS-i3 (The Forensic Science Service DNA Expert System Suite FSS-i3). In each of these expert systems, rules are activated when a DNA profile is not from a unique source or when the quality of the profile is substandard. The expert system then requests the analyst to manually re-examine the data and accept or reject the assignment made. The combined use of such automated systems has been shown to provide independent "expert" analyses, which increase consistency and save analysis time compared to manual processing.

4.3 DNA sequence analysis: Genome annotation

Since the completion of the human genome in 2003, a large number of genomes of other organisms have been sequenced. These genome sequences consist of strings of As, Ts, Cs, and Gs representing the base pairs that make up the DNA, and have lengths in the order of millions of characters. Without marking the locations of biologically important parts of the sequence such as the genes and their regulatory elements, this string of characters has little usefulness. A major challenge in the "post-genomic era" is therefore the localisation of the genes in each genome, their organization, structure and function. Two areas of genomic biology are dedicated to this task, namely structural and functional annotation. Structural

annotation refers to the task of identifying genes, their location on the genome sequence, their exon/intron structure and the prediction of the molecular sequences (RNA and proteins) that they encode. Functional annotation aims then to predict the biological function of the gene products: RNA and protein molecules.

Structural annotation methods can be classified into: (1) *ab initio* methods, based on codon usage to discriminate between coding and non-coding regions, and pattern searches for regulatory elements, (2) methods that use evolutionary conservation to infer gene localization and structure, (3) hybrid methods that combine these two approaches and usually present the best compromise in terms of sensibility and specificity in gene detection.

Computational methods for functional annotation are mainly split into two types: (1) similarity based approaches that infer a function based on the comparison of a given sequence with a sequence of known function and (2) phylogenomic inference approaches, based on evolutionary history and relationships between biological sequences. Phylogenomic methods avoid many of the false inference problems associated with the simpler similarity-based methods, although they require a high degree of biological expertise, are time consuming, complex, and are difficult to automate.

Both structural and functional annotations usually require the composite chaining of different algorithms, software and methods and expert biologists are often needed to make important decisions, modify the dataset and to compare intermediate results, which is labor intensive and can be error prone. In order to handle the large amounts of data produced by genome sequencing projects, automation of these pipelines is an absolute necessity. Diverse attempts have been made to develop annotation platforms automating some of these pipelines, particularly in the domain of structural annotation (e.g. the Ensembl pipeline (Hubbard et al, 2002). With regard to functional annotation, a number of platforms are available that can automate either the similarity-based approaches or the more complex phylogenomic approaches.

FIGENIX (Figure 1) (Gouret et al, 2005) is one example of an automated annotation platform featuring an expert system that models the biologists' expertise and is able to compare results from different methods, and to evaluate the significance of predictions. FIGENIX incorporates a number of different pipelines for structural and functional annotation, in particular, a structural annotation pipeline, which is a hybrid method combining *ab initio* and similarity-based approaches, and a functional annotation pipeline fully automating a complex phylogenomic inference method.

4.4 Protein sequence analysis

The complete sequencing of genomes for a number of organisms and their subsequent annotation has led to the definition of thousands of new proteins of unknown biological function. In order to investigate the biological activity of the proteins, the first step is to determine its primary structure, i.e. the ordered sequences of amino acids making up the protein. Knowledge of the amino acid sequence then allows predictions to be made about protein structure and the relationships between different proteins. Expert systems have been used to solve a number of different problems in such protein analyses.

The precise determination of amino acid sequences in proteins is a very important analytical task in biochemistry, and the most common type of instrumentation used for this employs Edman degradation (Edman, 1956). In this method, N-terminal amino acid residues are repeatedly labelled and cleaved from the protein. The amino acids are then identified as their phenylthiohydantoin (PTH) derivatives by high performance liquid chromatography

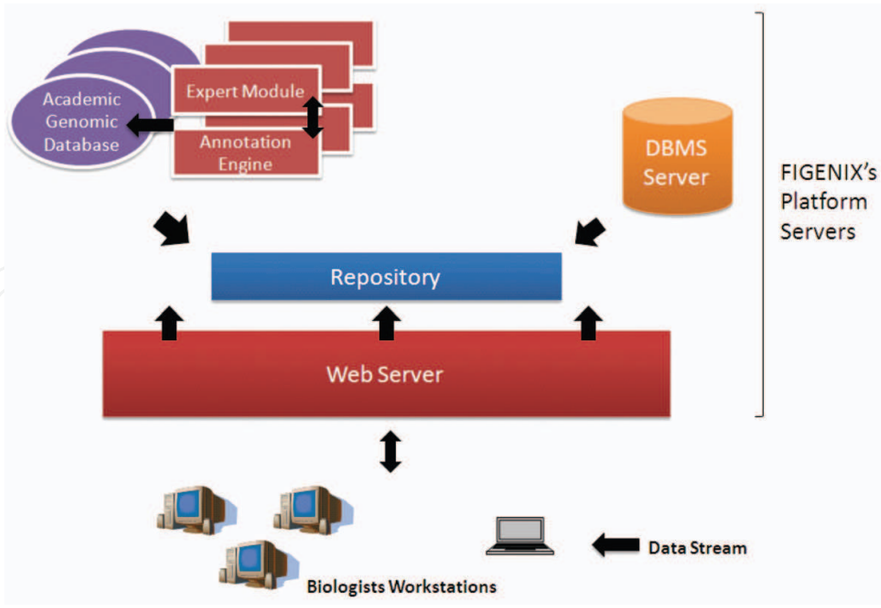

Fig. 1. Figenix architecture : "Annotation Engine" or "Expert System" servers can be cloned and distributed on several CPU.

(HPLC: a form of column chromatography used frequently in biochemistry and analytical chemistry to separate, identify, and quantify compounds). Interpretation of the HPLC data is often problematic due to unstable baselines (points of reference used to identify peak in the chromatograph), drifts in elution times of PTH-amino acids, and additional contamination effects which are a consequence of partial cleavage during Edman degradations and/or side reactions that occur during the cycles. Such problems may cause peak misidentification by onboard software programs on commercial sequencers, and visual interpretation of the HPLC data by a human expert is often required for successful interpretation of the chromatographic profile.

In this context (Hu et al, 1996) developed an expert system that uses heuristic rules built by human experts in protein sequencing. The system is applied to the chromatographic data of PTH-amino acids acquired from an automated sequencer. The peak intensities in the current cycle are compared with those in the previous cycle, while the calibration and succeeding cycles are used as ancillary recognition criteria when necessary. The retention time for each chromatographic peak in each cycle is then corrected by the corresponding peak in the calibration cycle at the same run. Such technologies have now been adapted by several biotechnology manufacturers and have been incorporated into their laboratory equipment to increase precision, making them more useful for experimentalists.

As another case study of an expert system application, (Praveen et al, 2005) reported an expert system for rapid recognition of metallothionein (MT) proteins. MT proteins are responsible for regulating the intracellular supply of biologically essential zinc and copper ions and play a role in protecting cells from the deleterious effects of high concentration metal ions. MT is generally induced when the organism experiences certain stress conditions and therefore, recognition of MT from tissues or from animal models is very important.

In order to develop an expert system for this task, the physical and chemical characteristics of MT proteins were derived based on a set of experiments conducted using animal models. The derived characteristics were broken into a set of rules, including (1) proteins with low

molecular weight versus high molecular weight, (2) proteins with metal content versus no metal content, (3) the presence or absence of aromatic amino acids and (4) sulphur content versus no sulphur content. The derived rules (consisting of a series of attribute and value pairs, followed by a single conclusion that contains the class and the corresponding class value) were produced using the ID3 algorithm (Quinlan, 1986), and a minimum number of rules were selected using human expertise in order to maximize true positive recognition. The rules were then formulated as an IF – THEN – ELSE algorithm and translated into VISUAL BASIC language statements, providing an efficient software solution.

Another important application of expert systems in protein analysis is mass spectrometry (MS). MS is an analytical technique for the identification of the composition of a sample or molecule. It is also used for the determination of the chemical structures of molecules, such as peptides (parts of protein sequences) and other chemical compounds. The method begins with degradation of the proteins by an enzyme having high specificity and the resulting peptides are subject to analysis by MS. A computer algorithm is then used to compare the masses established for the resulting peptides with theoretical masses calculated for every sequence in a protein or DNA sequence database and the studied protein is identified based on the results of this comparison.

ProFound (Zhang et al, 2000) is an expert system which employs a Bayesian algorithm to identify proteins from protein databases using mass spectrometric peptide mapping data. Bayesian probability theory has been widely used to make scientific inference from incomplete information in various fields in bioinformatics, including sequence alignment and NMR spectral analysis. When the system under study is modelled properly, the Bayesian approach is believed to be among the most coherent, consistent, and efficient statistical methods. The ProFound system ranks protein candidates by taking into account individual properties of each protein in the database as well as other data relevant to the peptide mapping experiment, including data from multiple digestions, the amino acid content of individual peptides, and protein components in mixtures. The program consistently identifies the correct protein(s) even when the data quality is relatively low or when the sample consists of a simple mixture of proteins.

4.5 Comparative genomics and evolutionary studies

The annotation of a single genome provides important clues to the functions of the encoded genes, and how they work together in the complex networks that perform the essential processes of living organisms. However, in the famous words of Theodore Dobzhansky: "nothing makes sense in biology except in the light of evolution" and therefore, the field of comparative genomics (the comparison of genome sequences from two or more organisms) and the reconstruction of ancestral genomes are now playing essential roles in understanding the evolutionary processes that have shaped the biological complexity of living organisms.

An expert system, called CASSIOPE (Rascol et al, 2009), has been developed specifically to address the problem of ancestral genome reconstruction. CASSIOPE compares the organizational structure of genomic regions that have been conserved in a large number of informative species. Hypotheses can then be formulated to account for such conserved genomic regions: (1) the regions are due to chance and are not biologically important, (2) the regions result from a standard ancestral region through inheritance or (3) the regions are due to evolutionary convergence with possible selective pressure. The objective then is to select the most likely explanation for these conserved regions. CASSIOPE is able to reject the null hypothesis (conserved regions are due to chance) in favor of one of the two alternatives,

although it cannot differentiate between them (ancestral regions or convergence). The system aims to automatically reproduce the chain of analyses and decisions performed by human experts and incorporates fundamental evolutionary biology-based concepts. First, orthologs (genes in different species that evolved from a common ancestral gene by speciation keeping the same functions) and paralogs (genes related by duplication within a genome, thus different functions) are detected via phylogenetic analysis. Second, the probability that the genomic regions from different species are inherited from a common ancestor is estimated based on the neutrality theory.

The process developed in CASSIOPE (Figure 2) involves various tasks, such as phylogenetic reconstruction or consulting web databases, and each task is independent from the others.

CASSIOPE deploys the following core tasks: (1) data processing: a modular system with various agents (virtual machines that work on specific tasks) deployed in conjunction with an expert system that communicates with every agent and takes rule-based decisions to answer initial biological questions, (2) data comparison: database searches can be performed for newly sequenced regions or genomes, (3) detection of orthologous genes by robust phylogenetic reconstruction, (4) statistical scoring to assess the significance of conserved regions and (5) a reverse-search feature making it possible to extend the initial searches.

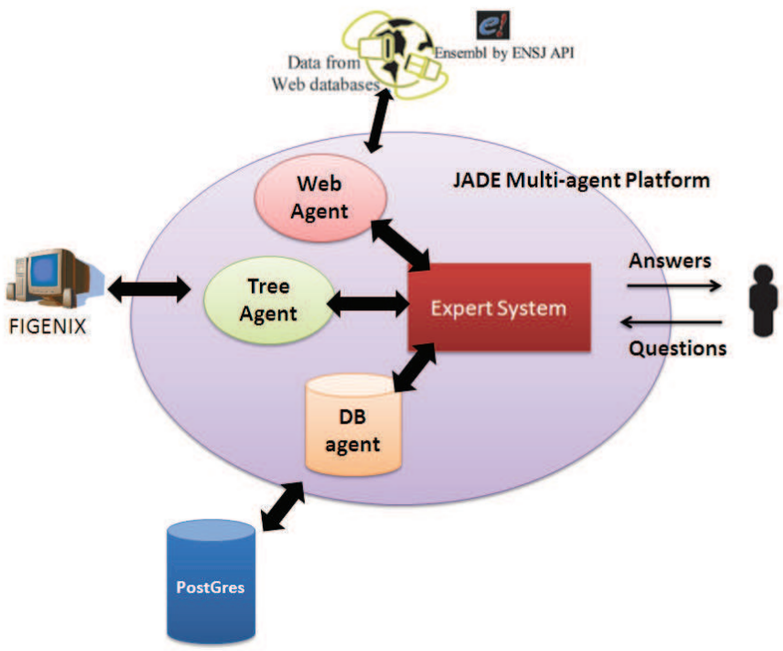

Fig. 2. CASSIOPE multi-agent system

The whole process is controlled by the expert system, which uses the other agents to obtain the data required. The expert system communicates with the different agents to answer questions, such as which genomic regions are significantly conserved? It receives queries and tries to find the necessary, pertinent data in the databases. The knowledge in the system is formulated in the form of an advanced rule set allowing decision-taking on the information received. For example, if data is partial, unavailable or outdated (> 1 month), the system deduces the agent interrogations it has to perform. The rule sets of the expert system can be updated, removed or added, just as a human scientist would.

5. Conclusion

Bioinformatics is a cross-disciplinary research field that is concerned with the efficient management and analysis of heterogeneous data and the extraction of valuable knowledge. Due to the exponential growth of the biological databases, especially in the so-called 'post-genomic era' since the human genome project, the task of extracting the hidden patterns underlying the data has become more and more difficult. The situation is further complicated by the complexity of biological data (numerous different types, sources, quality, etc.) and the creation of tools that can exploit the accumulated human expertise in this field is now crucial.

Thanks to recent developments in IT, the storage and the maintenance of the petabytes of data have been widely addressed and efficient solutions are being developed. Future challenges will concentrate less on how we can store the huge amount of data, and more on how we can explore the hidden knowledge. Success in this domain will have profound effects on fundamental research and our knowledge of the mechanisms involved in the complex networks that make up living beings. Bioinformatics is also playing an increasing important role in a wide range of applications, including medical diagnostics and treatment, pharmaceutics and drug discovery, agriculture and other biotechnologies.

The relatively simple task of annotating genes involves consulting various molecular, functional, disease and other databases to verify complete or partial annotations. Online data resources such as those furnished by the National Center for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute and many others, have become invaluable in helping annotators attribute putative functions to proteins based on computational results. The way in which biological information is stored, i.e. in distinct, heterogeneous data sources, means that data integration is an essential prerequisite to its annotation. Information regarding functional properties of genes for example is distributed in various online databases which were developed independently and do not inherently interoperate.

Today, new technologies are changing the bioinformatics data landscape and are providing new sources of raw data, including gene expression profiles, 3D structures, genetic networks, cellular images, DNA mutations involved in genetic diseases and many more. The current flood of data provides unique opportunities for systems-level studies. At the same time, it also poses as many new challenges. As a consequence, the field of systems biology has emerged, focusing on the study of complex biological systems, that exploits the new genomic data and the recent developments in bioinformatics. Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations (Ideker et al, 2001). Intelligent systems, such as the expert systems described in this chapter, will play an essential role in these studies, by integrating human expertise and computational and mathematical tools to highlight the knowledge underlying the data.

6. References

Bernstein, A.D.; Chiang, C.J. & Parsonnet, V. (1995). Diagnosis and management of pacemaker-related problems using and interactive expert system. IEEE 17th *Annual Conference on Engineering in Medicine and Biology Society*, 1, (701–702).

Brasil, L.M.; De Azevedo, F.M. & Barreto, J.M. (2001). Hybrid expert system for decision supporting in the medical area: complexity and cognitive computing International. *Journal of Medical Informatics*, 63, (19–30).

Carroll, J.S.; Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K. C., Hall, G.F., Wang, Q., Bekiranov,S., Sementchenko, V., Fox, E.A., Silver, P.A., Gingeras, T.R., Liu, X.S. & Brown, M. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.*, 38, (1289–1297).

Consortium, T.E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306, 5696, (636-640).

Dragulescu, D. & Albu, A. (2007). Expert system for medical predictions. 4th International Symposium on Applied Computational Intelligence and Informatics, 13–18.

Edman, P. (1956). On the mechanism of the phenyl isothiocyanate degradation of peptides. *Ada Chim. Scand.*, 10, (761-768).

Etzold, T.; Ulyanov, A. & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Meth Enzymol*, 266, (114–128).

Ge, H.; Walhout, A.J.M. & Vidal, M. (2003). Integrating `omic' information: a bridge between genomics and systems biology. *Trends in Genetics*, 19, (551-560).

Gouret, P.; Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P.& Danchin, E.G. (2005). FIGENIX: Intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics.*, 6, 198.

Grossi, E.; Mancini, A. & Buscema, M. (2007). International experience on the use of artificial neural networks in gastroenterology. *Digestive and Liver Disease*, 39, (278–285).

Halling-Brown, M. & Shepherd, A.J. (2008). Constructing computational pipelines. *Methods Mol Biol*, 453, (451-470).

Hass, L.; Schwartz, P., Kodali, P., Kotlar, E., Rice, J.& Swope, W. (2001). Discovery Link: a system for integrating life sciences data. *IBM Syst. J.*, 40, (489-511).

Hu, L.; Saulinskas, E.F., Johnson, P. & Harrington, P.B. (1996). Development of an expert system for amino acid sequence identification. *Comput Appl Biosci.,* 12, 4, (311-318).

Hubbard, T.; Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Res.*, 30, 1, (38-41)

Ideker, T.; Galitski, T. & Hood, L. (2001). A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2, (343-72).

Itchhaporia, D.; Snow, P.B., Almassy, R.J. & Oetgen, W.J. (1996). Artificial neural networks: current status in cardiovascular medicine. *Journal of the American College of Cardiology*, 28, 2, (515–521).

Khan, A.S. & Hoffmann, A. (2003). Building a case-based diet recommendation system without a knowledge engineer. *Artificial Intelligence in Medicine* 27, (155–179)

Koton, P. (1988). Reasoning about evidence in causal explanations. P*roceedings of the Seventh National Conference on Artificial Intelligence, AAI Press, Menlo Park, CA*, (256–263).

Lein, E.S.; Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T.M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H.W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R.,

Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramee, A.R., Larsen, K.D., Lau, C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Sivisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpf, K.R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Varnam, L.R., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Yuan, X.F., Zhang, B., Zwingman, T.A.& Jones, A.R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445, (168-176).

Ligeza, A. (2006). Logical Foundations for Rule-Based Systems Studies in Computational Intelligence; Springer, Berlin. 11.

Lopez, B. & Plaza, E. (1997). Case-based learning of plans and medical diagnosis goal states. *Artificial Intelligence in Medicine*, 9, (29–60).

Montani, S.; Magni, P., Bellazzi, R., Larizza, C., Roudsari, A.V. & Carson, E.R. (2003). Integrating model-based decision support in a multi-modal reasoning system for managing type 1 diabetic patients. *Artificial Intelligence in Medicine*, 29, (131–151).

Patterson, D.W. (1990). Introduction to Artificial Intelligence and Expert System, *Prentice-Hall, Inc., Englewood Cliffs, NJ, USA* .

Porter, B.W. & Bareiss, E.R. (1986). PROTOS: an experiment in knowledge acquisition for heuristic classification tasks. *Proceedings of the First International Meeting on Advances in Learning (IMAL), Les Arcs, France*, (159–174).

Power, T.; McCabe, B. & Harbison, S.A. (2008). FaSTR DNA: a new expert system for forensic DNA analysis. *Forensic Sci Int Genet*, 2, 3, (159-165).

Praveen, B.; Vincent, S., Murty, U.S., Krishna, A.R. & Jamil, K. (2005). A rapid identification system for metallothionein proteins using expert system. *Bioinformation*, 1, 1, (14-15).

Quinlan, J.R. (1986). Induction of Decision Trees. *Readings in Machine Learning*, 1, 1, (81-106).

Rascol, V.L.; Levasseur, A., Chabrol, O., Grusea, S., Gouret, P., Danchin, E.G. & Pontarotti, P. (2009). CASSIOPE: An expert system for conserved regions searches. *BMC Bioinformatics*, 10, 284.

Romano, P. (2008). Automation of in-silico data analysis processes through workflow management systems. *Brief Bioinform*, 9, (57-68).

Roos, D.S. (2001). Computational biology. Bioinformatics--trying to swim in a sea of data. *Science*, 291, (1260-1261).

Schuler, G.D.; Epstein, J.A., Ohkawa, H. & Kans, J.A. (1996). Entrez: molecular biology database and retrieval system. *Methods Enzymol*, 266, (141-162).

Souchelnytskyi, S. (2005). Bridging proteomics and systems biology: what are theroads to be traveled? . *Proteomics*, 5, (4123–4137).

Spellman, P.T.; Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, 9, (3273–3297).

van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet*., 37, (18–24).

Wilkinson, M.D. & Links, M. (2002). BioMOBY: an open source biological web services proposal. *Brief Bioinform.*, 3, (331-341).

Zhang, W. & Chait, B.T. (2000). ProFound: an expert system for protein identification using mass spectrometric peptide mapping infocrmation. *Anal Chem.*, 72, 11, (2482-2489).

Expert Systems

Edited by Petrica Vizureanu

ISBN 978-953-307-032-2
Hard cover, 238 pages
Publisher InTech
Published online 01, January, 2010
Published in print edition January, 2010

Expert systems represent a branch of artificial intelligence aiming to take the experience of human specialists and transfer it to a computer system. The knowledge is stored in the computer, which by an execution system (inference engine) is reasoning and derives specific conclusions for the problem. The purpose of expert systems is to help and support user's reasoning but not by replacing human judgement. In fact, expert systems offer to the inexperienced user a solution when human experts are not available. This book has 18 chapters and explains that the expert systems are products of artificial intelligence, branch of computer science that seeks to develop intelligent programs. What is remarkable for expert systems is the applicability area and solving of different issues in many fields of architecture, archeology, commerce, trade, education, medicine to engineering systems, production of goods and control/diagnosis problems in many industrial branches.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mohamed Radhouene Aniba and Julie D. Thompson (2010). Knowledge Based Expert Systems in Bioinformatics, Expert Systems, Petrica Vizureanu (Ed.), ISBN: 978-953-307-032-2, InTech, Available from: http://www.intechopen.com/books/expert-systems/knowledge-based-expert-systems-in-bioinformatics

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen