This article appeared in the October 2003 issue of the "Bulletin of the Swiss Statistical Society" (46; 10-17), and is reprinted by permission from its editor.

Challenges in Bioinformatics for Statistical Data Miners

Dr. Diego Kuonen

Statoo Consulting, PSE-B, 1015 Lausanne 15, Switzerland

kuonen@statoo.com

Abstract

Starting with possible definitions of statistical data mining and bioinformatics, this article will give a general side-by-side view of both fields, emphasising on the statistical data mining part and its techniques, illustrate possible synergies and discuss how statistical data miners may collaborate in bioinformatics' challenges in order to unlock the secrets of the cell.

What is statistics and why is it needed?

Statistics is the science of "learning from data". It includes everything from planning for the collection of data and subsequent data management to end-of-the-line activities, such as drawing inferences from numerical facts called data and presentation of results. Statistics is concerned with one of the most basic of human needs: the need to find out more about the world and how it operates in face of variation and uncertainty. Because of the increasing use of statistics, it has become very important to understand and practise statistical thinking. Or, in the words of H. G. Wells: "*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write*".

But, why is statistics needed? Knowledge is what we know. Information is the communication of knowledge. Data are known to be crude information and not knowledge by itself. The way leading from data to knowledge is as follows: from data to information (data become information when they become relevant to the decision problem); from information to facts (information becomes facts when the data can support it); and finally, from facts to knowledge (facts become knowledge when they are used in the successful completion of the decision process). That is why we need statistics and statistical thinking. Statistics arose from the need to place knowledge on a systematic evidence base.

What is data mining?

Data mining has been defined in almost as many ways as there are authors who have written about it. Because it sits at the interface between statistics, computer science, artificial intelligence, machine learning, database management and data visualisation (to name some of the fields), the definition changes with the perspective of the user. Here is a not so random sample of a few:

- "*Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.*" (M. J. A. Berry and G. S. Linoff)
- "*Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.*" (U. Fayyad, S. Chaudhuri and P. Bradley)
- "*Data mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets.*" ("Insightful Miner 3.0 User Guide")

In summary, we think of data mining as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns or models in data to make crucial decisions. "Valid" means that the patterns hold in general, "novel" that we did not know the pattern beforehand, and "understandable" means that we can interpret and comprehend the patterns. Hence, like statistics, data mining is not only the modelling and prediction steps, nor a product that can be bought, but a whole problem solving cycle/process that must be mastered, and is (almost) always a team effort. Indeed, data mining is the core component of the so-called "knowledge discovery in databases" (KDD) process, which is illustrated in Figure 1. For learning from data to take place, data from many sources ("databases") must first be gathered together and organised in a consistent and useful way ("data warehousing"). Data need to be cleaned and preprocessed ("data cleaning"). Quality decisions and quality mining results come from

quality data. Data are always dirty and are not ready for data mining in the real world. But, before the data can be analysed, understood, and turned into information, a task-relevant data target needs to be created ("data selection"). The main part of the KDD process is data mining, which is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. The choice of a particular combination of techniques to apply in a particular situation depends on both the nature of the data mining task and the nature of the available data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernible – or, patterns that are so obvious that no-one has noticed them before. This is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value.

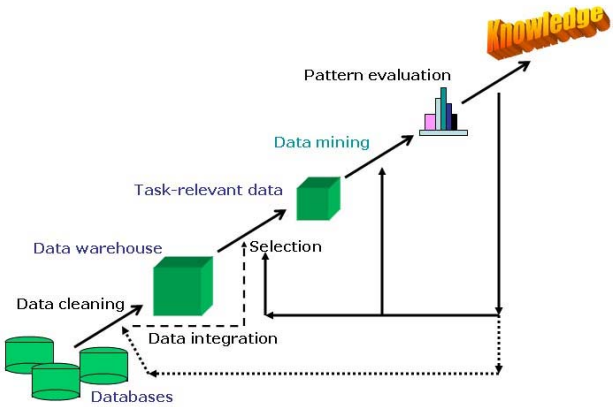

Figure 1 *The KDD process and its phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required.*

Recall that we defined statistics as the science of learning from data, and remember that the main way leading from data to knowledge is: from data to information and from information to knowledge. Data are what we can capture and store, and become information when they become relevant to the decision problem. Information relates items of data and becomes knowledge when it is used in the successful completion of the decision process. Hence knowledge relates items of information. As we see, the main problem is to know how to get from data to knowledge, or, as J. Naisbitt said: "*We are drowning in information but starved for knowledge*". The remedy to this problem is statistical data mining.

Data mining tasks

Let us briefly define the main tasks well-suited for data mining, all of which involve extracting meaningful new information from the data. The six main activities of data mining are: *classification* (examining the feature of a newly presented object and assigning it to one of a predefined set of classes); *estimation* (given some input data, coming up with a value for some unknown continuous variable); *prediction* (the same as classification and estimation except that the records are classified according to some future behaviour or estimated future value); *affinity grouping* or *association rules* (determine which things go together, also known as dependency modelling); *clustering* (segmenting a population into a number of subgroups or clusters); and *description and visualisation* (exploratory or visual data mining). Learning from data comes in two flavours: directed ("supervised") and undirected ("unsupervised") learning. The first three tasks – classification, estimation and prediction – are all examples of supervised learning. In supervised learning ("class prediction") the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The next three tasks – affinity grouping or association rules, clustering, and description and visualisation – are examples of unsupervised learning. In unsupervised learning ("class discovery") no variable is singled out as the target; the goal is to establish some relationship among all the variables. Unsupervised learning attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. This is similar to looking for needles in haystacks.

From a statistical perspective, many data mining tools could be described as flexible models and methods for exploratory data analysis. In other words many data mining tools are nothing else than

multivariate data analysis methods. Or, in the words of I. H. Witten and E. Franke: "*What's the difference between machine learning and statistics? Cynics, looking wryly at the explosion of commercial interest (and hype) in this area, equate data mining to statistics plus marketing*".

The development of new (statistical) data mining and knowledge discovery tools is a subject of intensive research. One motivation behind the development of these tools is their potential application in modern biology. For more than a decade scientists have laboured to obtain the entire genomic sequence for several organisms. The first complete genome published of a free-living organism, the bacterium *Haemophilus influenzae*, was published by "The Institute of Genomic Research". And a well-known database is the 30'000 to 35'000 human genes and the complete sequence of more than 3 billion chemical base pairs being compiled by the "Human Genome Project". The science of bioinformatics – the discipline that generates computational tools, databases, and methods to support genomic research – has grown up as a result of these efforts.

What is bioinformatics?

The genome is the total amount of genetic information that an organism possesses. Genes in turn are made from DNA, which contains the complete genetic information that defines the structure and function of an organism. DNA stores information in the form of the base nucleotide sequence, which is a string of 4 letters (Adenine, Cytosine, Guanine and Thymine), *e.g.*

```
TTCAGCCGATATCCTGGTCAGATTCTCTAAGTCGGCTATAGGACCAGTCTAAGAGA
```

for about 3 billion letters is the human genome. Genes are segments of DNA that contain the "recipe" to make proteins and proteins are the crucial molecules that do most of a cell's work. Indeed, the "central dogma of molecular biology" states that, in a cell, information flows from the nuclear DNA to RNA to protein synthesis. Hence proteins are formed using the genetic code of the DNA, and a protein sequence can be represented by a string of 20 letters, each standing for an amino acid. Stored digitally, in computers worldwide, are trillions of sequences, *i.e.* trillions of pieces of information – information, which needs to be turned into knowledge.

Identifying and interpreting interesting patterns hidden within the immense list of bases that constitute a genome is a critical goal in molecular biology research. A number of genes in a new genome will be unlike anything previously known. The ability of an algorithm not only to find known sequences but also to recognise new sequences is crucial. This is only one of the topics of bioinformatics. More generally:

- "*Bioinformatics is the science of managing and analysing biological data using advanced computing techniques. Especially important in analysing genomic research data.*" (Glossary of genetic terms from the "DOE Human Genome Program")
- "*Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and more generally, asking biological questions with a computer.*" (J.-M. Claverie and C. Notredame)
- "*Bioinformatics is the study of genetic and other biological information using computer and statistical techniques.*" (L. Helmuth)

We think of bioinformatics as the science of storing, extracting, organising, analysing, interpreting and utilising information from biological sequences and molecules. For example, the past few years have witnessed an extraordinary surge of interest in sequence and transcriptome analysis; technologies which offer the first great hope for providing a systematic way to explore the information contained in the genome. Bioinformatics merges such new techniques with computer science technology and advanced statistical (data mining) methods to organise, analyse and interpret data. Mining for data and learning from it has evolved in much the same way. Older methods executed by statisticians took a long time to yield constructive information. Now, current software and techniques help make bioinformatics (and data mining) a more accessible process in the information age that we live in. What characterises the state of bioinformatics is the flood of data that exists today or that is anticipated in the future; data that need to be mined to get reliable information. The mountain of information that is, for example, the draft sequence of the human genome may be impressive, but without interpretation that is all what it remains – a mass of

data. For learning from data to take place, the KDD process should be applied in order to get from data to information and from information to knowledge – to unlock the secrets of the cell.

Bioinformatics tasks

One of the most basic operations in bioinformatics involves searching for similarities, or homologies, between a (newly) sequenced piece of DNA and previously sequenced DNA segments from various organisms. Finding near-matches allows researchers to predict the type of protein the (new) sequence encodes. This not only yields for drug targets early in drug development but also weeds out many targets that would have turned out to be dead ends. Let us investigate, for example, whether there are any proteins similar to the mouse *nucleolin* in the protein database SWISS-PROT. The solution can be found by using a bioinformatics tools like the "Basic Local Alignment and Search Tool" (BLAST). BLAST is part of a suite of DNA- and protein-sequence search tools accessible online in various customized versions from many databases. The core part of BLAST is a beautiful and powerful example of the application of probability theory and statistics (compromising aspects of random walk theory, of renewal theory and asymptotic distribution theory) within bioinformatics. The protein sequence of the mouse *nucleolin* – the task-relevant data – is

```
>sp|P09405|NUCL_MOUSE Nucleolin (Protein C23) - Mus musculus (Mouse).
VKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTPA
KKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVPT
PGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPAK
AAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVAE
EEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKVE
GSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAED
LEKALELTGLKVFGNEIKLEKPKGRDSKKVRAARTLLAKNLSFNITEDELKEVFEDAMEI
RLVSQDGKSKGIAYIEFKSEADAEKNLEEKQGAEIDGRSVSLYYTGEKGQRQERTGKTST
WSGESKTLVLSNLSYSATKETLEEVFEKATFIKVPQNPHGKPKGYAFIEFASFEDAKEAL
NSCNKMEIEGRTIRLELQGSNSRSQPSKTLFVKGLSEDTTEETLKESFEGSVRARIVTDR
ETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKGEGGFGGRGGGRGGFG
GRGGGRGGRGGFGGRGRGGFGGRGGFRGGRGGGGDFKPQGKKTKFE
```

(*Source: www.expasy.org/cgi-bin/get-sprot-fasta?P09405*)

Figure 2 is an example of an output returned by the BLAST server of the "National Center for Biotechnology Information" (NCBI; see *www.ncbi.nlm.nih.gov/blast/*). The figure shows where the query is similar to other sequences. Together with a so-called "hit list" (not shown) one sees whether the sequence looks like something already in the database and whether or not one can trust it as a good "hit".

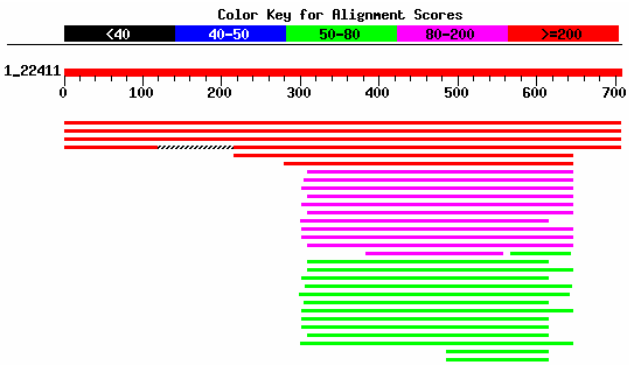

Figure 2 *Graphical output of the BLAST at www.ncbi.nlm.nih.gov/blast/. The query sequence is on the top and each bar represents the portion of another sequence similar to the query sequence and the region of the sequence where this similarity occurs. Red bars indicate the most similar searches, pink pars indicate matches that are a bit less good, and green bars indicate matches that are not impressive at all. The blue and the black bars indicate matches that have the worst scores.*

Already with such pair-wise alignments, the interpretation is a bit of an art. On the other hand, for "multiple sequence alignments" the scores that tell how reliable the database search is do not (yet) exist. In addition, multiple sequence alignment methods are not perfect as exact algorithms exhaust current computational resources. Hence "imperfect" heuristic algorithms are used. The people who make them know this, and the users who apply them should also consider this fact. Figure 3 is an example of running the so-called ClustalW algorithm together with Jalview (*www.es.embnet.org/Services/MolBio/jalview/*) to

view its output, *i.e.* showing the multiple sequence alignment. Note that multiple alignment and the so-called phylogenetic tree construction (see Figure 4 for an example) are inter-related problems. Indeed, multiple alignments are the basis for the construction of phylogenetic trees. On the other hand, multiple alignment aims at aligning a whole set of sequences to determine which subsequences are conserved and this works best when a phylogenetic tree of related proteins is available. The amount of resources for making sequence alignments online is almost overhelming, including, for example, possibilities to produce multiple alignment using the Gibbs sampler, genetic algorithms or simulated annealing. *etc.*.

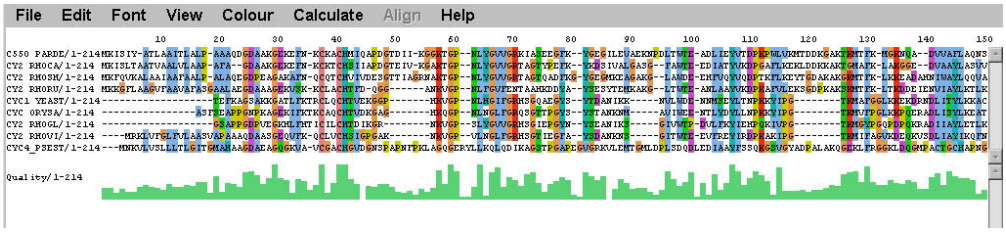

Figure 3 *Graphical illustration of a multiple sequence alignment.*

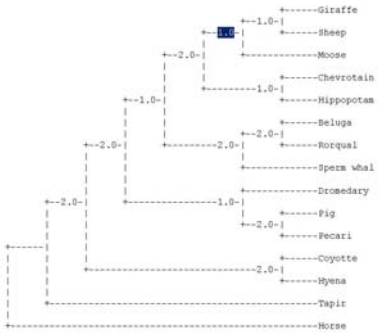

Figure 4 *An example of a phylogenetic tree using the "Neighbor Joining" method for tree construction. The tree was produced by means of Phylip (bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html). The number above each branch indicates the stability, and its reliability has been checked using the bootstrap.*

In summary, sequence analysis seeks to tease out information based on a sequence itself, or on the similarity of one sequence to another ("pair-wise alignment"), or even on patterns among groups of sequences ("clustering"). Another example of unsupervised learning from sequence analysis data is affinity grouping ("categorical sequence mining"), where one wants to discover sequences of events that commonly occur together, *e.g.* in a set of DNA sequences `ACGTC` is followed by `GTCA` after a gap of 9, with a probability of 30%. On the other hand, supervised data mining techniques can be applied as well. For example, the goal of "DNA sequence classification" is to distinguish "junk" segments from coding segments and this can be done using supervised learning.

Sequence data are not the only digital biological information available to researchers. Another data type beginning to fill countless disk drives is that resulting from gene expression analysis. Genomic sequence itself reveals only the possibilities of genetic manifestation. Within any given cell, only a small fraction of genes are "expressed", that is, being actively translated into proteins through intermediate RNA molecules. The patterns of gene expression distinguish a skin cell from a liver cell, and a cancerous cell from a normal cell. In the past several years, a new technology, called DNA microarrays (or gene chips), has attracted tremendous interests among biologists. This technology promises to globally monitor gene expression on a single chip so that researchers can have a better picture of the interactions among hundreds to thousands of genes simultaneously. Using sets of short artificially constructed sequences ("probes") – each designed to stick to ("hybridise") a specific molecule of DNA – researchers can measure the relative quantities of RNA molecules in a particular tissue sample and thus infer what genes are active. The design of the probes is itself a very computationally intensive task, but the rate at which expression analysis studies can generate data pose yet another computational challenge. The most common method for expression analysis utilises small glass slides onto which has been deposited a grid of many hundreds of probes. In this case, the basic data is now two-dimensional images comprising spots of varying

intensities representing the relative expression in a particular tissue sample of the genes being assayed. An example of such an image is given in Figure 5.

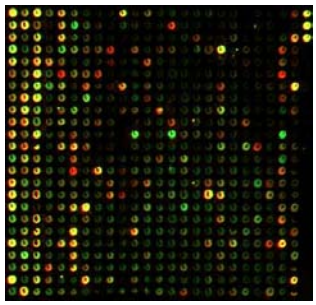

Figure 5 *Microarray experiments produce gene expression images like the one pictured here. These images must be converted to numbers before analysis can proceed.*

These image files require significantly more storage than one-dimensional sequence data. To extract quantitative cleaned-up expression data from the images complex image analysis techniques need to be applied. Once the data is derived from the images, the computational problem can become one of unsupervised statistical data mining: looking for patterns of expression across thousands of genes (high dimensional data) from any number of samples (normally only containing a very limited number). Unsupervised techniques used include hierarchical clustering, *k*-means and self-organizing maps in order to identify new subgroups or classes, and association analysis. The results of unsupervised learning can be pictured in several ways, *e.g.* by linking them to a so-called "heat map": a plot of the expression matrix of genes (row) and samples (columns). An example is given in Figure 6, where the genes and the samples have been arranged in orderings derived from hierarchical clustering**,** which has been applied independently to the genes (rows) and to the samples (columns). The two-way rearrangement of Figure 6 produces an informative picture of genes and samples. Indeed, there are two possibilities of clustering in microarray analysis. First, in order to know if there are groups of genes with a similar expression pattern across different experimental conditions (*i.e.* samples). The assembly of such groups might provide insights about novel genes based on the observation that genes involved in the same functional pathway tend to have similar expression patterns. Second, to discover classes in the samples, *e.g.* tissue type, disease *vs.* healthy. Hence the two dendrograms in Figure 6 themselves are very useful, as biologists can, for example, interpret the gene clusters in terms of biological processes.

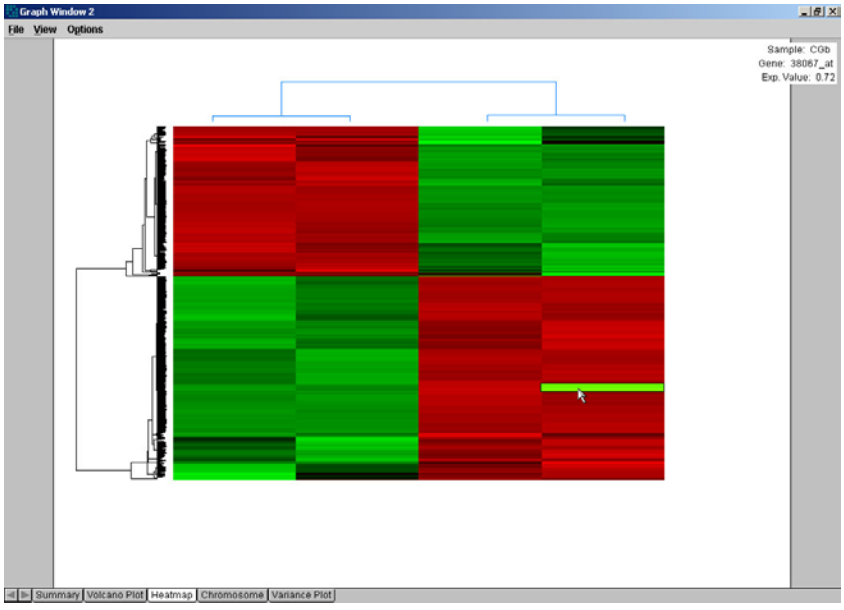

Figure 6 *An interactive heat map using Insightful's S+ArrayAnalyzer. It is linked to dendrograms resulting from average linkage hierarchical clustering, which has been applied independently to the rows (genes) and columns (samples). Pixels coloured red signify positive expression values and those coloured green signify negative expression. The brighter the colour the larger the intensity in absolute value.*

The first goal of microarray analysis is class (disease or tissue type) discovery, which can be performed using unsupervised techniques. Once such groups are known, if they were not known beforehand, supervised data mining techniques can be applied. For example, to classify entities into known classes, *e.g.* diseases and therapies, one could apply classification techniques like discriminant analysis, nearest neighbour methods, artificial neural networks, Bayesian networks, support-vector machines, decisions trees, boosting, bagging and independent component analysis. New techniques are still being developed.

Still another important type of data in bioinformatics are three-dimensional structural descriptions of proteins and other biologically important molecules; see Figure 7 for an example. While there are computer-based efforts to determine protein structure from basic sequence information, the full "protein-folding" problem is considered a grand challenge and thus lies in the domain of advanced supercomputing research. Three-dimensional protein structure carries much more biologically relevant information than the one-dimensional sequence of amino acids, and there are several efforts dedicated to increasing the throughput of structure determination. As more structure data become available, more computational resources will be focused on manipulating molecules and simulating molecular interactions as part of drug lead identification.

Figure 7 *Three-dimensional structure of the ras protein (image credit: "U.S. Department of Energy Human Genome Program", www.ornl.gov/hgmis).*

Bioinformatics and visualisation

Last but not least, let us briefly mention also exploratory or visual data mining. Visualisation is used in many areas within bioinformatics, with varying success: for some topics good tools already exist, while for others they do not, often both because of the screen space problems common to many visualisation problems and because sufficient thought has not gone into how best to visualise the data to aid comprehension. In many area of bioinformatics it is important to be able to view information at several levels of detail and shift between them readily, which can be challenging for software.

Some bioinformatics visualisations are also computationally challenging. A general complaint is that visualisation tools are not sufficiently interactive to allow effective exploration of data. The visualisation of biological data is also hampered by the wide range of data types and exponentially increasing volume of data available, and by the lack of interoperability of existing tools. To overcome this problem clearly requires the development of policies on data sharing and standards. One development within the overall bioinformatics knowledge extraction field, visualisation being only a side aspect of it, that may be of wider utility is the "Distributed Annotation System" (DAS; see *www.biodas.org*). The latter allows for a distributed set of annotations to a database, without the modification of the original database itself.

There is an interesting aside to this: the question of whether we may rapidly be approaching a limit to how much significant biological information we can turn into knowledge. As ever more biological phenomena are being described, our concepts of biological knowledge and understanding will change, and we will need to recruit ever more computer and (statistical) data mining tools to organize such knowledge and to extract and present the relevant information to us in a comprehensible way. The development of concepts and models that integrate such complex knowledge and allow its visualisation to make it

accessible is the grand future challenge of bioinformatics. When problems such as these are solved, and as more knowledge is mined out of DNA sequences, this will bring a revolution in understanding of human health, genetics and the functioning of living organisms.

Conclusion and challenges for statistical data miners

Statistical data mining approaches seem ideally suited for bioinformatics, since it is data-rich, but lacks a comprehensive theory of life's organization at the molecular level. Or, in the words of I. E. Alcamo: "*Keeping up with the directions and applications of DNA is a never-ending job*".

However, data mining in bioinformatics is hampered by many facets of biological databases, including their size, their number, their diversity and the lack of a standard ontology to aid the querying of them, as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels and domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanisms appropriate to all. The integration of biological databases is also lacking, so it can be very difficult to query more than one database at once. For example, the massive amount of microarray data collected so far has been generated on multiple platforms and is stored in different formats, levels of detail and locations. This makes it difficult for any research group to re-analyse or verify the data, or compare the results with their own. Finally, the possible financial value of, and the ethical considerations connected with, some biological data means that the data mining of biological databases is not always as easy to perform as is the case in some other areas.

From a statistical data miner's perspective most bioinformaticians tend to be ignorant of statistical data mining, are too impatient for solutions (pressure to publish), expect the statistical data miners to know the solutions long before they have any data, will only use the latest algorithms, ignore software unless it is easy to use and are always updating the data files. On the other hand, from a bioinformatician's perspective statistical data miners do not understand the biological questions, take too long to come up with answers, moan about sample size and replication, speak a different scientific language, speak different programming languages and use dreadful software. Hence most bioinformaticians and statistical data miners continue to sarcastically criticise each other. However, as we have illustrated, the field of bioinformatics, like statistical data mining, concerns itself with "learning from data" or "turning data into information and information into knowledge". It is important to note that bioinformatics can learn from statistical data mining - that, to a large extent, statistical data mining is fundamental to what bioinformatics is really trying to achieve. There is the opportunity for an immensely rewarding synergy between bioinformaticians and data miners.

Data mining and bioinformatics are fast expanding research frontiers. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective analysis. The active interactions and collaborations between these two fields have just started and a lot of exciting results will appear in the near future. Bioinformatics and data mining will inevitably grow toward each other because bioinformatics will not become knowledge discovery without statistical data mining and thinking. A maturity challenge for statistical data miners and bioinformaticians is to widen their focus until true collaboration and the unlocking of the secrets of the cell become reality.

(*Image credit: "U.S. Department of Energy Human Genome Program", www.ornl.gov/hgmis*)

Note

This article was presented during an invited lecture on "Bioinformatics: Data Mining and Graphical Methods" at the seminar of the ROeS (the Austro-Swiss region of the "International Biometric Society"), which was held in St. Gallen, Switzerland, from September 29 to October 2, 2003. The article is reprinted by permission from its organisers.

References and resources

Alcamo, I. E. (2000). *DNA Technology: The Awesome Skill*. Academic Press.

Baldi, P. & Brunak, S. (2001). *Bioinformatics - The Machine Learning Approach (2nd Ed.)*. Cambridge, MA: MIT Press.

Bayat, A. (2002). Science, medicine, and the future - bioinformatics. *British Medical Journal*, 324, 1018-1022.

Berry, M. J. A. & Linoff, G. S. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. New York: Wiley.

Claverie, J.-M. & Notredame, C. (2003). *Bioinformatics for Dummies*. New York: Wiley.

Demidov, V. V. (2003). Golden jubilee of the DNA double helix. *Trends in Biotechnology*, 21, 139-140.

Dutton, G. (2002). Gene expression data mining. *The Scientist*, 16, 50.

Ewens, W. & Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction*. New York: Springer.

Han, J. (2002). How can data mining help bio-data analysis? In: M. J. Zaki, J. T. L. Wang and H. T. T. Toivonen (Eds.), *Proceedings of the "2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics"*, 1-2 (*www.cs.rpi.edu/~zaki/BIOKDD02/*).

Han, J. & Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hand, D. J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining.* Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Helmuth, L. (2001). A genome glossary. *Science*, 291, 1197.

Holmes, S. P. (1999). Phylogenies: an overview. In: M. E. Halloran and S. Geisser (Eds.), *Statistics in Genetics, The IMA Volumes in Mathematics and its Applications, 112*, 81-118.

Kanehisa, M. (2000). *Post-Genome Informatics*. Oxford: Oxford University Press.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis (2nd Ed.)*. New York: Springer.

Liu, J. S. (2002). Bioinformatics: microarrays and beyond. *Amstat News*, 303, 59-67.

Mann, B., Williams, R., Atkinson, M., Brodlie, K, Storkey, A. & Williams, C. (2003). Scientific data mining, integration, and visualisation. *Report of the workshop held at the e-Science Institute, Edinburgh, 24-25 October 2002* (*umbriel.dcs.gla.ac.uk/NeSC/general/talks/sdmiv/report.pdf).*

Nguyen, D. V., Arpat, A. B., Wang, N. & Carroll, R. J. (2002). DNA microarray experiments: biological and technological aspects. *Biometrics*, 58, 701-717.

Sebastiani, P., Gussoni, E., Kohane, I. S. & Ramoni, M. F. (2003). Statistical challenges in functional genomics (with discussion). *Statistical Science*, 18, 33-70.

Stein, L. (2002). Creating a bioinformatics nation. *Nature*, 417, 119-120.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. London: Chapman & Hall.

Witten, I. H. & Franke, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

"Association for Computing Machinery Special Interest Group on Knowledge Discovery in Data and Data Mining" (ACM SIGKDD): *www.acm.org/sigkdd/*

"Bioinformatics World" magazine: *www.bioinformaticsworld.info*

"Cracking the Code of Life – Journey into DNA": *www.pbs.org/wgbh/nova/genome/dna.html*

"DOE Human Genome Program": *www.DOEgenomes.org*

"ExPASy Molecular Biology Server": *www.expasy.org*

"Functional Genomics: Beyond the Genome – Introduction to Bioinformatics and Genomics": *www.library.csi.cuny.edu/~davis/Bioinformatics/Bioinfo_Genomics.html*

Homepage of H. Jiawei (data mining and bioinformatics): *www-faculty.cs.uiuc.edu/~hanj*

Homepage of A. Robinson (visualisation, data mining and bioinformatics): *industry.ebi.ac.uk/~alan*

Homepage of M. J. Zaki (data mining and bioinformatics): *www.cs.rpi.edu/~zaki*

"Introduction to Elements of Biology – Cells, Molecules, Genes, Functional Genomics, Microarrays": *www.ebi.ac.uk/microarray/biology_intro.htm*

Statoo Consulting's bioinformatics page: *www.statoo.com/en/bioinformatics/*

Statoo Consulting's free data mining and statistics email newsletter: *lists.statoo.com*

