Intelligent information systems

Machine Learning & Biology

Prof. Jan Komorowski, director of the Linnaeus Centre for Bioinformatics, a computer scientist by education, does interdisciplinary research in molecular biomedicine, especially transcriptomics and proteomics, using machine learning methods

Prof. Jacek Koronacki, director of the Institute of Computer Science (Polish Academy of Sciences), is a specialist in statistical data analysis and machine learning

JAN KOMOROWSKI
Linnaeus Centre for Bioinformatics,
a joint initiative of Uppsala University
and the Swedish University of Agricultural Sciences
Jan.Komorowski@lcb.uu.se

JACEK KORONACKI
Institute of Computer Science, Warsaw
Polish Academy of Sciences
korona@ipipan.waw.pl

The past century, sometimes dubbed the "age of information," saw incredible growth in computers' processing speed and memory capacity. Now that computational power is helping us forge ahead into a new "age of biology"

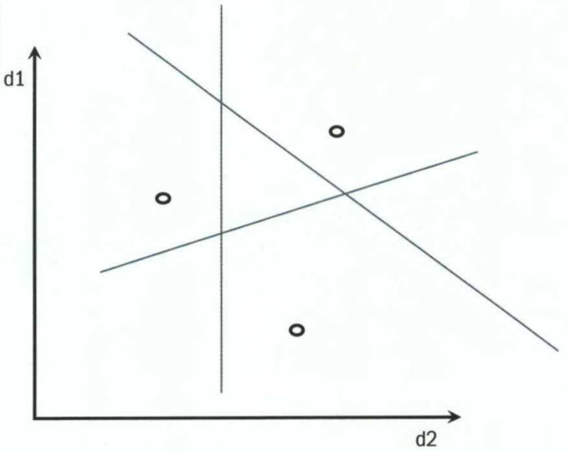
Speaking more broadly, the current century might be described as an era of the life sciences. Biotechnologies now enable massive amounts of data to be collected about living cells and organisms. Whereas the microarrays used in the late 1990s allowed tens of thousands of genes to be observed in parallel, nowadays we have technologies that enable the whole human genome to be sequenced within a week - a single such experiment producing a dataset of around $5 \cdot 10^9$ points, all while costing less than €10,000.

It is no wonder that the life sciences and information sciences finally ceased to be disjoint fields at the end of the past century. Previously, the life sciences had dealt with small quantities of data and a vast majority of research results were described in qualitative language, with a high degree of modality. The models built by biologists are frequently of a qualitative nature, as sets of approximate rules. The more recent appearance of mass quantities of biological data has both enabled and forced the life sciences to turn to completely new approaches. It has quickly become clear how important machine learning methods can and will be for the life sciences, especially methods that generate human-readable models - e.g. taking the form of trees or decision rules. On the one hand such models are fully compatible with quantitative modeling, while on the other hand they still retain the form of qualitative rules.

Learning by example

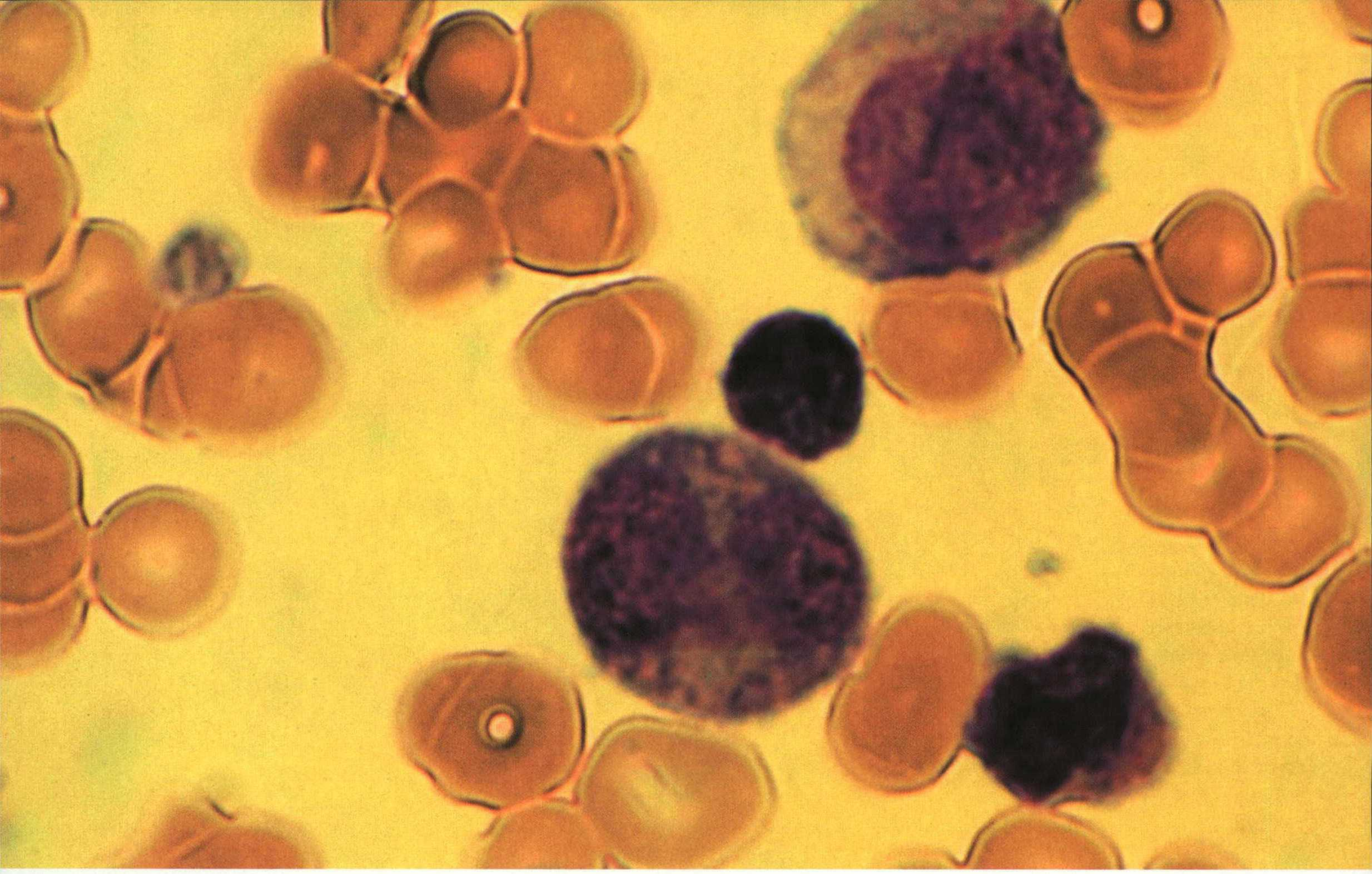
Since biologists now have huge sets of information at their disposal, they also have a great need for "intelligent" and automatic processing methods. Human researchers need the help of computers equipped with algorithms to put that information into order and extract useful knowledge from it. Machine learning algorithms, enabling a computer to "learn" from a set of examples presented to it, can perform such useful knowledge extraction.

Here we have chosen to describe just two examples (illuminating ones, we hope) showing how machine learning methods for classifying objects into given classes can be applied in genomics and proteomics research. For computational purposes, each "object" is represented as a vector, meaning a finite ordered set of "features." Those features take on different values for different objects, coming from fixed ranges of possible values. The classification algorithm learns to classify objects into the given classes based on a "training set" given to it, i.e. a set of objects of known classification.



Jacek Koronacki

Building a classifier able to differentiate two different cell types based on the expression of 7129 different genes would miss the essence of the problem, since in a p-dimensional space (in our case p=7129, illustrated here for the simpler case of p=2) p+1 points can be divided up into two arbitrary subsets by a hyperplane (illustrated here for 3 points) - and our training set contains just 38 points



Photoresearchers/BE&W

How can different types of leukemia be distinguished? Here: a microscope view of a peripheral blood smear from a patient suffering from acute marrow leukemia

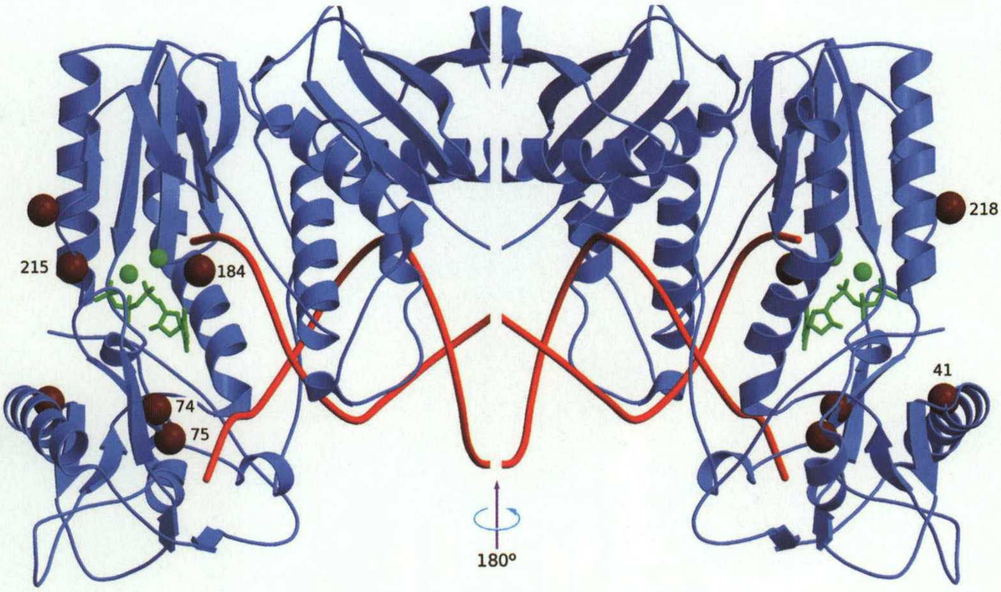
The first application described here models the expression of genes in cells, and thus each feature represents the expression of a specific gene. Each object is in this case a specific cell, for which the expression of selected genes has been measured. In this example the training set consists of 38 objects, each of which is described by a vector reflecting the expression of 7129 different genes. 27 of the training objects are known to belong to a class representing acute lymphoblastic leukemia or ALL, while the remaining 11 objects belong to a class representing acute marrow leukemia or AML (this is a well-known training set, collected and first analyzed by Golub *et al.*). Once the classification algorithm constructs classification rules (also known as decision rules) based on the training set, they can then be applied to the task of automatically classifying new objects of unknown class.

One popular type of classification algorithm is called a classification tree. Here decision rules take the form of an if-then implication, containing a conjunction of conditions on selected features plus a classification conclusion to be drawn if those conditions are met. For example, assuming that the set classes are Φ, Ψ and Ω, one of the rules might take the following form: {*if the* i^{th} *feature of object O has a value greater than a, the* j^{th} *feature has a value no greater than b, and the* k^{th} *feature has a value greater than c, then object O belongs to class* Ω}.

Diagnosing cancer earlier?

Turning back to our example with the leukemia data, we first need to notice that just building a classifier able to distinguish between objects (cells) of the ALL/AML classes will not capture the essence of the problem faced by biologists. Firstly, computationally this task would just boil down to dividing up a set containing only 38 points into two disjoint subsets within a Euclidean space of dimension 7129 - and from the geometrical point of view, at least, that is a trivial task. It can be proven that within any *p*-dimensional space, a hyperplane can divide (p+1) points into two arbitrary and disjoint subsets, provided that those points do not lie within some subspace of the *p*-dimensional space - in other words, in our example involving a space of dimension 7129, we could easily find two disjoint subsets not just for 38 points, but even for as many as 7130 points (see the illustration on p. 28, where *p* is just 2). Solving such a classification task would thus be trivial.

Significantly more interestingly for biologists, we can expect that the correct classification is in fact not determined by all the 7129 genes, but rather by a small fraction thereof. Therefore, before building such a classifier



The second example application of our method involves modeling the drug-resistance of HIV-1 retrovirus in order to identify which amino acids comprising the reverse transcriptase (RT) depicted here are responsible (when mutated) for such resistance. The red balls represent C-alpha amino acid atoms pinpointed by the algorithm as important in the mechanism of resistance to the drug didanosine

Jan Komorowski, Jacek Koronacki

or at least while building it, we would like to find out which genes are specifically linked to a given type of disease. In other words, we are interested not only in classification itself, but also - or perhaps chiefly - in answering the question of which genes are responsible for a given type of disease, or the expressions of which genes signal that a given disease is already in progress in an organism.

A great number of methods based on broadly-construed data analysis have been devised, seeking a way to answer such questions. We have grounds to believe that the method recently proposed by our research team is particularly noteworthy. Our method is conceptually very simple and natural, although it does require a computer to process many simple calculations. A given feature is deemed important in the process of specifying a given class and in differentiating it from other classes (in other words, a feature is deemed to carry information about a class) if that feature "tends" to take part in the process of classifying objects from the training set. This tendency or "inclination" of a feature to take part in the classification process (which we call that feature's "importance") is measured by building many (thousands of) classification trees using different, randomly-selected subvectors of the vector of the features. This method utilizes thousands of classifiers, not in order to build an overall general classifier but rather to identify which features are important for the given classification task. Once the subset of such features is known, we may (but do not necessarily have to) proceed to build a general classifier. Notably, our method not only identifies the important features but also ranks those features in terms of importance. Working with the data of Golub *et al.*, we have obtained results that may be considered extraordinarily interesting.

Earlier attempts at identifying objectively important features - without appealing to biologists' knowledge - usually involved finding those features whose values differ significantly (in the statistical sense) between objects belonging to different classes. Individual features were analyzed, without consideration for their possible interaction with other features. As a result, such methods would only find genes whose expression corresponds to late stages of the cancer types under consideration, expressing especially dramatic changes caused by cancer. Our approach, in turn, seeks to pinpoint features whose values (separately or in tandem with other features) enable objects from different classes to be distinguished. Working with the data of Golub *et al.*, we also managed to detect genes that are weakly but sufficiently differentiated, linked to the onset of or a proclivity for the given type of cancer. Such an outcome suggests that we can diagnose early stages of cancer, or even perhaps predict the risk of its occurrence.

Honing in on drug resistance

Our proposed feature-ranking method allows us to explicitly gauge the interdependencies among features and at the same time to evaluate the contribution of those interdependencies in the process of classifying objects. The second example application we will describe here involves modeling the drug resistance exhibited by the HIV-1 retrovirus, which mutates extraordinarily quickly. Within three days after infection some 10 million viruses in practically all of its mutations occur within the host's cells, some of those mutations resistant to any given medication.

The illustrations on pages 30 and 31 present the strongest interdependencies we discovered in the data on HIV-1 resistance to a drug called didanosine, used to inhibit retrovirus replication (data from the Stanford University HIV Resistance Database). The dataset consisted of 706 objects, each of them described by 560 amino acids and each belonging to one of three classes: resistant to the drug, susceptible to the drug, and exhibiting moderate resistance.

Research on modeling drug resistance has to date focused on building the best possible classifier, while setting aside the problem of explaining the mechanism that underlies such resistance. In our approach, we replaced the names of the amino acids with their physicochemical properties. Using knowledge of biology, we distinguished the seven most important properties, yielding a total of 3920 features. The strongest correlations were found for a total of six positions (amino acids). Of those, five were previously considered by experts to be important in the development of resistance, while one amino acid proved to be new. Once we know which positions are important and what correlations there are between them, we can proceed to develop a three-dimensional reverse transcriptase (RT) structure and analyze the mechanisms of resistance - its acquisition or loss (for simplicity's sake here we have set aside the moderate resistance class).

Within the structural breakdown of HIV RT, certain subdomains figuratively known as the thumb, palm, and fingers are recognized. Mutations do not occur within the thumb subdomain. The first mechanism detected involves active locations in the palm, while the second is related to the movement of the finger subdomains. A mutation in an active location has obvious consequences: the virus has gained resistance to a drug. The fingers can be figuratively described as keeping hold of the DNA so that further nucleotides can be added onto the emerging object. A mutation in such a location leads the drug, which previously prevented the fingers from holding the DNA at the right place, to cease to work and the function of the fingers becomes restored.

Presently we have begun to apply our method to problems whose solutions are not yet known. We therefore hope to achieve one of the objectives of computational biology: to build *in silico* models of phenomena that occur *in vivo* or *in vitro*, yielding promising hypotheses that can then be conclusively tested in the lab.

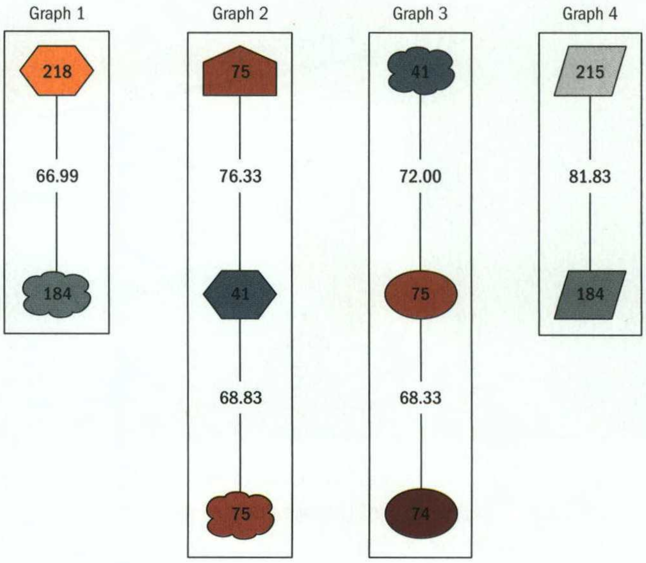
The team implementing this project benefited from consultations with Dr. Krzysztof Ginalski and Dr. Witold Rudnicki from Warsaw University's Interdisciplinary Center for Mathematical and Computational Modeling. ■

Further reading:

Dramiński M., Rada-Iglesias A., Enroth S., Wadelius C., Koronacki J., Komorowski J. (2008). Monte Carlo Feature Selection for Supervised Classification. *Bioinformatics, 24 (1)*, 110-117.

Dramiński M., Kierczak M., Koronacki J., Komorowski J. (2009). *Monte Carlo Feature Selection and Interdependency Discovery in Supervised Classification.* [In:] Koronacki J., Raś Z., Wierzchoń S., Kacprzyk J. (Eds.). *Advances in Machine Learning*, Springer (in press).

This diagram illustrates the strongest interdependencies we discovered in the data on resistance to the drug didanosine, used to inhibit HIV retrovirus replication. The shape indicates the biochemical property, the color shows the number of the position, the linking bars indicate the interdependence, and the numbers on the linking bars note the strength of the interdependency



Jacek Koronacki