# Data Mining in the Bioinformatics Domain

**Shalom Tsur**
SurroMed, Inc.
Palo Alto, CA., USA
tsur@surromed.com

## Abstract

Bioinformatics, the study and application of computational methods to life sciences data, is presently enjoying a surge of interest. The main reason for this welcome publicity is the nearing completion of the sequencing of the human genome and the anticipation that the knowledge derived from this process will have a great impact on modern medicine. The pharmaceutical industry, which expects to utilize the knowledge for new drug design, has a particular interest in bioinformatics.

The structure of data in this domain has its own characteristics which set it apart from data in other domains. While genomic data have a well-known representation as sequences taken from the {A,C,G,T} alphabet, there is no clear model for data representing the expression products of genes: proteins and higher forms of organisms e.g., cells and the multitude of forms they assume in response to environmental challenges.

Data collected at these levels of information can be often thought of as "broad": meaning that for a relatively small number of records representing biological samples, a very large number of attributes, representing measurements or observations is collected per sample. In contrast, typical data used for mining are "long" i.e., consist of a large number of records in which each record is characterized by a relatively small number of attributes.

Mining broad data presents a new and unique challenge. The presentation will elaborate on some of the issues in this domain.



**Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.**

## 1 Introduction

In the biological enterprise, biological samples e.g., blood, are collected from donors or study subjects and are subjected to an array of different measurements. These measurements can be quantitative, to determine the purity or concentration of some substance such as a protein in the sample, or can be qualitative to merely detect the presence of some substance. Measurements of the former type are referred to as assays. The process and conditions under which these samples are processed, the timing and the characterization of the participating subjects, are specified in a study or clinical protocol.

Biology draws a distinction between the *genotype* and *phenotype* of an organism. The genotype is determined by its genetic makeup and is invariant over the organism's life. The phenotype on the other hand, is determined by a set of observable characteristics of the organism that in turn, are determined by its genotype *and* by the environment. Thus, a certain protein is the expressed product of a gene. The measured concentration of this protein in the blood may be the result of a disease burden, taking a certain drug, a diet, exposure etc. The phenotype is thus a set of time varying quantities. Tracing a phenotype over time may provide a longitudinal record of e.g., the evolution of a disease and the response to a therapeutic intervention. By analogy, the code making up a software system (assuming we do not change it) would be its genotype. The dynamic execution behavior of the system, which is dependent on the code, the operating system, the input data and the user-interaction with it, would be its phenotype. It is worth noting that portions of this code may never be executed and hence, will not contribute to the dynamic behavior. Likewise, the biological genome contains large portions of DNA that are considered "junk" and seemingly do not serve any purpose.

From the clinical perspective, subjects interact with physicians who collect their own observations on their patients. A measure of interest in this context is that of a *clinical endpoint*: a set of characteristics that directly measure how well a patient feels, functions or survives. Examples would be a patient's blood pressure, the time required to climb 5 stairs or simply, the response to the question "how do you feel?" Often these measures offer only vague and imprecise information and worse, whenever they become clearer it is often too late: the disease has advanced and at best, can be arrested at the present state, rather than having prevented it at an early stage. There is a strong need therefore to come up with better predictive information that would support the clinical practice.

Lastly, the patient him/herself is the most reliable source of information about his own wellness. Especially since the physician has only a very limited opportunity for interaction and given the constraints he operates under, is unable of forming a complete and reliable picture of the patient's health. Tools for systematically assisting patients in assessing their own health form thus an important complement to the information already collected.

The sources we mentioned here, measurements on biological samples, clinical information and patient's self-assessment define all of the characteristics required for a comprehensive determination of human phenotype. In the future, integrated data warehouses containing these sources will be created and will be used to derive knowledge of interest to a variety of different consumers: the patient herself, the physician/care giver, health insurers, the pharmaceutical industry and lastly, to the academic research community. We will elaborate on some of this knowledge in the sequel. Perhaps, the biggest payoff these integrated data will yield is that of enabling *personalized medicine*—care giving that is based on the individual's genotype and phenotype.

## 2 Biological Markers

According to the NIH Definitions Working Group, a Biological Marker (or Biomarker) is defined as:

> A characteristic that is measured and evaluated as an indication of normal biologic processes, pathogenic processes or pharmacological responses to therapeutic intervention

There exist today a few but well-known examples of biomarkers: elevated levels of Cholesterol (LDL, HDL) are biomarkers for Cardiovascular disease, reduced counts of CD4+ T-cells is a biomarker for HIV, and high PSA (Prostate Specific Antigen) concentration is a biomarker for Prostate Cancer. These characteristics serve only as indicators—they are not necessarily the cause of the disease. In other words, they correlate with the disease but do not form a causal link; eliminating these symptoms does not necessarily influence the course of the disease. The value of this knowledge is therefore in its predictive potential in that the characteristics can be observed a long time before the disease manifests itself to the extent it is observed by the physician in normal clinical practice. Biomarkers serve therefore as an early warning signs, which hopefully enable preventative therapeutic intervention.

Other applications of biomarkers include:

- Determine susceptibility to disease and enable early diagnosis.
- Predict disease severity and outcome
- Predict and monitor response to therapeutic interventions.

We noted that the pharmaceutical industry has a particular interest in this knowledge. The biggest problem facing this industry today is the so called "Clinical Bottleneck:" advances in modern science have created a situation in which potential leads for drugs are generated at a rate that vastly outperforms the ability to evaluate these during clinical trials. The total time from lead identification to completion of trials has therefore tremendously grown, the risk of failure is very high and today, the total cost of a successful launching of a new drug is on the order of $M300–600. Any information that would reduce the time or the risk involved in this process is of great value to the industry and biomarkers are expected to play a critical role in this respect. They could serve to stratify patient populations i.e., classify them into smaller, better-defined sub-populations of patients suffering from some disease. A drug could then be developed for only a particular sub population. This would reduce the risk and cost involved and would ease the FDA licensing requirements that must be met.

## 3 Databases for Biological Information

Databases, or more appropriately data warehouses constructed to support the goals described in the previous sections, accumulate data from a multitude of different sources that are combined to represent the phenotype. For example, in the case of SurroMed Inc., a warehouse is under construction containing measurements obtained using a multitude of bioanalyis techniques: cellular assays to measure populations of cells having certain identifiable antigenic characteristics, immunoassays to measure concentrations of small molecules in the blood, and the results of mass spec measurements to obtain more information about proteins and small organic molecules in the blood. These data are combined with the responses obtained from test subjects to a detailed questionnaire assessing their state of health. A simple data model representing these sources would be:

$$phenotype(Subject, Sample, Time, C_1, \ldots, C_n; \\ S_1, \ldots, S_m; P_1, \ldots, P_k; H_1, \ldots, H_w)$$

where the $C$, $S$ and $P$ components represent the cellular, immunoassay and mass spec measurements respectively, obtained from $Sample$ and the $H$ region represent the health-related information obtained from $Subject$. The data are partially ordered by $Time$ and represent multiple measurements obtained from the same subject over time i.e., they represent the results of longitudinal studies. A multitude of different dependencies and correlations exist between these components, often in ways that are not completely understood. The $H$ region is essentially a long vector of categorical values representing answers to health related questions.

The model represents an array of $N$ samples by $M$ attributes representing measurements or observations. $N$ is on the order of 100's and $M$ is on the order of 1000's. In this model $M \gg N$ and furthermore, as the measurement technology develops, the ratio $M/N$ is expected to increase rapidly. We are thus presented with a *broad data model*. This model is very different from the "typical" data set used in a mining application e.g., a set of credit card transactional records, in which the number of records is very large and the number of attributes is small. Hence $M \ll N$ and we refer to this model as a *long data model*. The model presented here forms the tip of the iceberg in the sense that the model components at this level are the results of a considerable data reduction process at lower levels during which the raw, uninterpreted measurement results were condensed into the the top level parameters. For example, tens of thousands measured cellular events are first clustered into populations and the resulting population statistics are presented at the top level. Performing this data reduction process requires deep domain expertise and the model is a summary of results spanning the biology, chemistry and the medical domains of expertise. The most challenging task is to horizontally interpret this broad model so as to infer from it information of relevance to bio markers.

## 4 Data Mining: Using the Phenotype for Predictive Purposes

We are interested in creating predictive models that would enable us to use a small subset of measured parameters, collected from the $C, S$ and $P$ regions of the model to predict the state of health of a subject. Specifically, assume that we can use the $H$ information (responses to a detailed medical questionnaire, used for self assessment)to partition the subject population into classes. The classes will be determined by an unsupervised clustering method. Denote the vector of health responses of subject $i$ by $\vec{H}_i$. The distance between two response vectors $\vec{H}_i$ and $\vec{H}_j$, obtained from subjects $i$ and $j$ will be denoted by $d_{ij}; d_{ij} = f(\vec{H}_i, \vec{H}_j)$. The objective is to define a distance measure $d$ such that the intra-cluster distance among responses that are "similar" is much smaller than the inter-cluster distance among responses belonging to different clusters. The quality of the clustering clearly depends on the distance measure used. Ideally, the measure maximizes some function (e.g., the average) of the inter-cluster distances and minimizes the intra-cluster distances. Thus, we seek a measure $d$ such that:

$$min\{\frac{1}{N}\sum_{i,j} d_{ij}\} \qquad i,j, \ \ in\ the\ same\ cluster$$

$$max\{\frac{1}{N}\sum_{i,j} d_{ij}\} \quad i,j, \ \ not\ in\ the\ same\ cluster$$

The method, which does not assume any a-priori knowledge about the subjects, has one drawback: there is no objective way to evaluate the quality of the clustering; we cannot determine from the clustered information how similar the state of health of respondents within the same cluster really is. Nor can we label the cluster and associate it with a known state of health. We need therefore an independent method for the verification of the results. The most promising verification method is to link the information with an electronic medical record (EMR) independently obtained about the subject from his/her physician.

Once we have a subject classification we can use it, in a supervised learning mode, to infer a classifier that uses a subset of the measurements ($C, S$ and $P$ regions of the model) as an input vector and which maps this input into one of the subject classes. We seek thus to learn a function $g : [X_1, \ldots, X_k] \rightarrow Y$ where $X_1, \ldots, X_k$ are taken from the measurement regions and $Y$ is a subject class.

The learning methodology of $g$ proceeds by dividing the data in two sets: a training set and a test set. For each of the record of the training set we associate the known subject class. We train the learning algorithm and test the result on the remaining test set. There exist a multitude of different machine learning methods, the Support Vector Machine method [Nell00] appears to be a very promising technique. Different input vectors will produce different classification results. In a broad data model like ours there is a danger of overfitting the data: using large input vectors it is easy to infer a classifier that will produce perfect results for the training points but will perform poorly for any other input data. The big issue is thus to select the smallest input vector to produce high quality classification results. This is a complex combinatorial problem. At this stage the data mining strategy sketched out here is untested and is a subject of ongoing research. It is possible that ultimately, a systematic search for the best input vector in the measurement space may be the only feasible approach to this problem.

## 5 Conclusion

In this short paper we have provided the background and an overview of the emerging domain of Bioinformatics. This area presents a set of new problems that hitherto have not been addressed by the data mining community and it can reasonably be assumed that these problems will become central with the rapid advances of modern biology. Given the space constraints of this paper it is impossible to provide a complete exposition of the field and therefore, only a few challenging problems, of particular interest, were exposed. Nevertheless, it is hoped that this will be sufficient to create more interest in an area that until now was largely hidden from the database community.

## References

[Nell00] Nello Cristiani, John Shawe-Taylor. *Suport Vector Machines*, Cambridge University Press, 2000