# Introduction

In all areas of biological and medical research, the role of the computer has been dramatically enhanced in the last two decades. The first generation of computational analysis focussed on sequence analysis, where many highly important unsolved problems remain till date. The current and future needs will in particular concern with large level *integration* of extremely diverse sets of data. These novel types of data originate from a variety of high throughput experimental techniques of which many are capable of data production at the levels of entire cells, organs, organisms, or even populations.

These new, efficient experimental techniques include primarily next-gen DNA sequencing, that have led to an exponential growth of linear descriptions of protein, DNA and RNA molecules.

Other new high throughput data producing techniques work as massively parallel versions of traditional experimental methodologies. Genome-wide gene expression measurements

using DNA microrarrays is, in essence, an accumulative simulation of tens of thousands of Northern blots. As a result, computational support in experiment design , processing of results and interpretation of results has become essential these days to analyse such voluminous data.

As genome and other sequencing projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation.

Sequence data needs to be integrated with structure and function data, with gene expression data, with pathways data, with phenotypic and clinical data, and so forth. Basic research within bioinformatics will have to deal with these issues of *system* and *integrative* biology, in the situation where the amount of data is growing exponentially.

Reconstructing the underlying evolutionary history have become an essential component of the research process. This is essential to our understanding of life and evolution, as well as to the discovery of new drugs and therapies.

Thus Bioinformatics has emerged as a strategic discipline at the frontier between biology and computer science, impacting medicine, biotechnology, and society in many ways.

This is due to the inherent complexity of biological systems, brought about by evolutionary tinkering, and to our lack of a comprehensive theory of life's organization at the molecular level. Machine-learning approaches (e.g. neural networks, hidden Markov models, vector support machines, belief networks), on the other hand, are ideally suited for domains characterized by the presence of large amounts of data, "noisy" patterns, and the absence of general theories

The fundamental idea behind these approaches is to *learn the theory automatically from the data*, through a process of inference, modelfitting, or learning from examples.

The concept of information and its quantification is essential for understanding the basic principles of machine-learning approaches in molecular biology.

It is the experience of many people that machine-learning methods are *productive* in the sense that near-optimal methods can be developed quite fast, given that the data are relatively clean. Machine learning is by and large a direct descendant of an older discipline,

statistical model fitting. The major goal in machine learning is to extract useful information from a corpus of data  by building good probabilistic models.The particular twist behind machine learning, however, is to automate this process as much as possible, often by using very flexible models characterized by large numbers of parameters, and to let the machine take care of the rest.

Machine-learning approaches are best suited for areas where there is a large amount of data but little theory. And this is exactly the situation in computational molecular biology. We have discovered a lot in the world of life science but we have to learn much more to interpret the structure, function and philosophy of this world of life. Thus, in computational biology in particular, and more generally in biology one must reason in the presence of a high degree of uncertainty: many facts are missing, and some of the concept of information and its quantification is essential for understanding the basic principles of machine-learning approaches in molecular biology the facts are wrong.

**Data-driven prediction**

The methods employed in machine learning framework should be able to extract essential features from individual examples and to discard unwanted information when present. These methods should be able to distinguish positive cases from negative ones, also in the common situation where a huge excess of negative, nonfunctional sites and regions are present in a genome.

Classification and prediction algorithms are in general computational means for *reducing* the amount of information. The contractive character of these algorithms means that they cannot be inverted; prediction programs cannot be executed backward and thus return the input information. Machine-learning approaches may have some advantages over other methods in having a built-in robustness when presented with uncorrelated data features. Information reduction is a key feature in the *understanding* of almost any kind of system. As described above, a machine-learning algorithm will create a simpler representation of a sequence space that can be much more powerful and useful than the original data containing all details.

**ARTIFICIAL NEURAL NETWORKS**

Artificial neural networks (NNs) were originally developed with the goal of modelling information processing and learning in the brain.

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. In the human brain, a typical neuron collects signals from others through a host of fine structures called *dendrites*. The neuron sends out spikes of electrical activity through a long, thin stand known as an *axon*, which splits into thousands of branches. At the end of each branch, a structure called

a *synapse* converts the activity from the axon into electrical effects that inhibit or excite activity from the axon into electrical effects that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

NNs have become an important tool in the arsenal of machine-learning techniques that can be applied to sequence analysis and pattern recognition problems. At the most basic level, NNs can be viewed as a broad class of parameterized graphical models consisting of networks with interconnected units evolving in time.

**Binomial Classification**

If a classification problem has only two classes, $A$ and $\bar{A}$, it is called a binomial classification. For a given input $d$, the target output $t$ is 0 or 1. The natural probabilistic model is a binomial model. The single output of the network then represents the probability that the input is a member of the class $A$ or $\bar{A}$, that is the expectation of the corresponding indicator function. This can be computed by a sigmoidal transfer function. Therefore, in the case of binomial classification, the output transfer function should be logistic; the likelihood error function is essentially the relative entropy between the predicted distribution and the target distribution. The derivative of $E$ with respect to the total input activity into the output unit, for each example, has the simple expression $-(t - y)$.

**Essential genes**

The genome of an organism characterizes the complete set of genes that it is capable of encoding. However, not all of the genes are transcribed and translated under any defined condition. The robustness that an organism exhibits to environmental perturbations is partly conferred by the genes that are constitutively expressed under all the conditions, and partly by a subset of genes that are induced under the defined conditions.

An essential gene is defined here as a gene necessary for growth to a fertile adult. (Kemphues). Essential genes of an organism constitute its minimal gene set, which is the smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions( Kunin et al, Glass et al). The deletion of only one of these genes is sufficient to confer a lethal phenotype on an organism regardless the presence of remaining genes. Therefore, the functions encoded by essential genes are crucial for survival and could be considered as a foundation of life itself. The identification of essential genes is important not only for the understanding of the minimal requirements for cellular life, but also for practical purposes. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for such drugs (Sarangi A N, et al).

In the era of complete genomes, the total number of genes in a sequenced organism can now be predicted (Claverie), but the function and selective importance of a substantial fraction of genes remains unknown (Hollon). The conditional importance of genes in conferring robustness can be understood in the context of the functional attributes of these genes and their correlations to the defined environmental conditions. However, a priori prediction of such genes for a given condition is yet not possible.

The prediction and discovery of essential genes have been performed by experimental procedures such as single gene knockouts, RNA interference and conditional knockouts (Gustafson), but these techniques require a large investment of time and resources and they are not always feasible.

Considering these experimental constraints, a computational approach capable of accurately predicting essential genes would be of great value. For prediction of essential genes, some

investigators have implemented computational approaches in which most are based on sequence features of genes and proteins with or without homology comparison (Gabriel del Rio et al) . With the accumulation of data derived from experimental small-scale studies and high-throughput techniques, however, it is now possible to construct networks of gene and proteins interaction (De la Rivas J) and then investigate whether the topological properties of these networks would be useful for predicting essential genes.