# Computational Biology



**Andrew LaBrunda**
*GTA, Guam*

**Michelle LaBrunda**
*Cabrini Medical Center, USA*

## INTRODUCTION

It is impossible to pinpoint the exact moment at which computational biology became a discipline of its own, but one could say that it was in 1997 when the society of computational biology was formed. Regardless of its exact birthday, the research community has rapidly adopted computational biology and its applications are being vigorously explored.

The study and application of medicine is a dynamic challenge. Changes in medicine usually take place as a result of new knowledge acquired through observation and experimentation. When a tamping rod 1-inch thick went through Phineas Gage's head in 1848, his survival gave the medical field an unusual opportunity to observe behavior of a person missing their prefrontal cortex. This observation lead to the short-lived psychosurgical procedure known as a lobotomy, which attempted to change a person's behavior by separating two portions of a person's brain (Pols, 2001). Countless observations, experiments and mistakes represent how almost all medical knowledge has been acquired.

The relatively new field of computational biology offers a nontraditional approach to contribute to the medical body of knowledge. Computational biology is a new field combining biology, computer science, and mathematics to solve problems that are unworkable with traditional biological techniques. It includes traditional areas such as systems biology, molecular biology, biochemistry, biophysics, statistics, and computer science, as well as recently developed disciplines including bioinformatics and computational genomics. Algorithms, which are able to closely model biological behavior, validate the medical understanding of the observed processes and can be used to model scenarios that might not be able to be physically reproduced.

The goal of computational biology is to use mathematics and computer science to model biological systems on the molecular level. Instead of taking on large complex systems, computational biology is starting small, literally. Modeling problems in molecular biology and biochemistry is a far less daunting task. At a microscopic level, patient's characteristics drop out of the equation and all information behavior affecting is known. This creates a deterministic model which, given the same input, will always produce the same output. Some of the major subdisciplines of computational biology are computational genomics, systems biology, protein structure prediction, and evolutionary biology, all of which model microscopic structures.

## COMPUTATIONAL GENOMICS

An organism's heredity is stored as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Each of the storage methods contains a linear chain of finite elements called bases. In the case of DNA the chain is composed of four bases, adenine, cytosine, guanine, and thymine, and RNA is composed of four bases, adenine, cytosine, guanine, and uracil. This information can be representing in a computer through a series of linked-lists or arrays. The order of bases in DNA/RNA is 99.9% the same between members of the same species and genetic sequencing is the way of determining the order of the bases. Certain regions of DNA, called genes, are used by the body to create proteins. These proteins are used to construct and maintain the organism. Genes can be thought of as blueprint instructions for how to make each unique person. There are long stretches of DNA between the genes the function of which is not well understood. The sum of all the genetic information about an organism is called a genome. When a computer is applied to deciphering an organism's genome, this is known is computational genomics. Computational genomics are used to better understand and compare sequences, identify related organisms, and measure biodiversity.

There are two major challenges in genomic studies. The first is genetic sequencing, (determining the order of the bases that make a strand of genetic material) and the second is localizing the genes within the genome. Sequence comparison is probably the most useful computational tool for molecular biologists. Repositories containing hundreds of genomes have been established and are available for public access. A biologist now has the ability to compare a unique sequence of DNA with the already known genetic sequences from this massive repository. Prior to the application of computers to the problems genomics scientist had to manually attempt to align sequences using ill suited tools such as word processors.

By identifying the genome for several related organisms it is possible to compare the related species and see how genetic mutation could allow one organism to evolve while leaving the other species unchanged. This type of analysis provides scientists with a more accurate descriptive tool to identify differences between organisms rather than relying on physical taxonomy. It also gives scientist an opportunity to see how a species' genetic information changes over time. If a large number of species have a genetic mutation which increases their successfulness, this observation may indicate an environmental change. As an example, if three unrelated species in different parts of the world genetically mutate to grow longer or thicker hair, one could hypothesize that an ice age was beginning.

One of the most important modern molecular genetic advances is the sequencing of specific genes and computational biology is becoming of increasing importance in sequencing studies. In the future, medications may be tailor made to the needs of each individual based on their specific genetic makeup. One of the hurdles that must be crossed to reach this point is a cost effective and accurate computer sequencing technique (Mitchell & Mitchell, 2007). Standardizing the reporting system of gene sequencing would minimize error and give researches a common template from which data can be extracted.

Now, computers have the ability to perform searches using pair-wise or fuzzy logic algorithms. Fuzzy logic allows for the identification of nonintuitive relationships. Identification of such relationships can provide a great deal of insight. The sequence of a gene not only encodes genetic information, but also give clues as to the function of the gene (Gibas & Jambeck, 2001, pp 13-14). Studies are actively being done using fuzzy systems to model genetic processes (Ishibuchi & Nojima, 2006).

## NEURAL NETWORKING AND FUZZY LOGIC

A fuzzy system is a control system utilizing a nondiscrete mathematical system to form loosely coupled decision making data structures. Fuzzy logic has many different applications. The most widely accepted solution for pattern matching is that of Neural Networking (NN). Neural networks are simplistic software models of brain function at a cellular level. The nervous system is composed of neurons, which receive input through dendrites and transmit output through axons, all of which communicate through synapses (small spaces between nervous system cells across which chemical messages are transmitted chemically). Hypothetical software models of neural processes can be applied to emulate some of the pattern matching abilities of the brain. For example, a series of lotteries forming words can be fed into a NN system. The NN first "learns" by observing the probability that a sequence of letters will occur will occur. As the NN processes additional input, nonlinear relationships emerge where new data uses a similar set of nodes, representing neurons, during processing. Fuzzy systems accept inputs between 0 and 1 representing a continuum of trueness. This is different from classical discrete computational systems, which only allow inputs of false (0) and true (1). Fuzzy neural-based systems try to maximize accuracy while minimizing complexity.

## PROTEIN FOLDING

Proteins are molecules which comprise many of the structural and functional components of living things. They are made by stringing together amino acids. All of the proteins in the human body are made from combination of various quantities of only 20 different amino acids. As proteins are created (by ribosomes) they fold into three-dimensional structures and it is the 3-dimensional interaction of the amino acids that is important. Sometimes a protein may be composed of multiple three dimensional structures interacting to perform a specific function in the body. It is the three-dimensional structure of a protein that determines its function. Many diseases occur when just one amino acid is replaced for another, such as in sickle cell anemia.

Understanding the structure of proteins allows understanding of their function and may someday provide a new therapeutic target for people with disease caused by defects in protein structure.

Research is currently being done using computational biology to understand protein structure in the energy-producing portion of cells called mitochondria (Gabaldon, 2006). Mitochondria are unique in that they are all inherited maternally. Understanding mitochondrial protein structure helps us understand how normal cells function and allows for understanding of diseases caused by mitochondrial dysfunction. Classically, mitochondrial diseases are considered to be uncommon and include diseases such as Leber's hereditary optic neuropathy and Kearns-Sayre syndromes, but more recent evidence has linked mitochondrial function to common diseases such as type 2 diabetes, Parkinson's disease, atherosclerotic heart disease, stroke, Alzheimer's disease, and cancer (Baloyannis, 2006; Folmes & Lopaschuk, 2007). Understanding mitochondrial protein structure will likely provide insight into disease processes for a gamut of pathological processes.

Although no computer model is currently able to accurately replicate the processes of protein folding, the accuracy of models has been improving. Currently, a popular method of protein modeling is to look at the hydrophobic (water repelling) and hydrophilic (water loving) properties of each individual amino acid and modeling the protein structure on a two or three-dimensional grid lattice. The