Machine Learning and Data Mining in Bioinformatics

George Tzanis, Christos Berberidis, and Ioannis Vlahavas
Department of Informatics, Aristotle University of Thessaloniki, Greece

INTRODUCTION

Machine learning is one of the oldest subfields of artificial intelligence and is concerned with the design and development of computational systems that can adapt themselves and learn. The most common machine learning algorithms can be either supervised or unsupervised. Supervised learning algorithms generate a function that maps inputs to desired outputs, based on a set of examples with known output (labeled examples). Unsupervised learning algorithms find patterns and relationships over a given set of inputs (unlabeled examples). Other categories of machine learning are semi-supervised learning, where an algorithm uses both labeled and unlabeled examples, and reinforcement learning, where an algorithm learns a policy of how to act given an observation of the world.

Data mining is a more recently emerged field than machine learning is. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently. The scope of data mining is the knowledge discovery from large data amounts with the help of computers. It is an interdisciplinary area of research, that has its roots in databases, machine learning, and statistics and has contributions from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. The main difference between machine learning and data mining is that machine learning algorithms focus on their effectiveness, whereas data mining algorithms focus on their efficiency and scalability.

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies as well as the introduction of new ones that made possible the conduction of many large scale experiments. An important example is the Human Genome Project, that was founded in 1990 by the U.S. Department of Energy and the U.S. National Institutes of Health (NIH) and was completed in 2003. A representative example of the rapid biological data accumulation is the exponential growth of GenBank (Figure 1), the U.S. NIH genetic sequence database (www.ncbi.nlm.nih.gov). The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data. This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology. Luscombe et al. (2001) identify the aims of bioinformatics as follows:

- The organization of data in a way that allows researchers to access existing information and to submit new entries as they are produced.
- The development of tools that help in the analysis of data.
- The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

There is a strong interest in methods of knowledge discovery and data mining to generate models of biological systems. In order to build knowledge discovery systems that contribute to our understanding of biological systems, biological research requires efficient and scalable data mining systems.

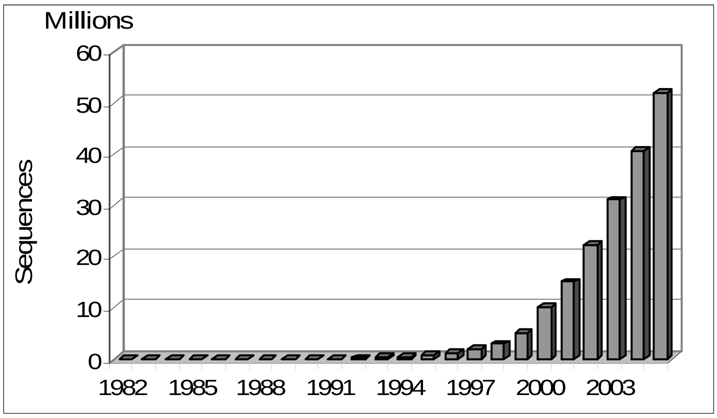

Figure 1: Growth of GenBank (1982-2005)

BACKGROUND

One of the basic characteristics of life is diversity, which can be noticed by the great differences among living creatures. Despite this diversity, the molecular details underlying living organisms are almost universal. Every living organism depends on the activities of a complex family of molecules called *proteins*. Proteins are the main structural and functional units of an organism's *cell*. A typical example of proteins is enzymes, which catalyze (accelerate) chemical reactions. There are four levels of protein structural arrangement (conformation) as listed in Table 1. The statement about unity among organisms is strengthened by the observation that similar protein sets, having similar functions, are found in very different organisms. Another common characteristic of all organisms is the presence of a second family of molecules, the *nucleic acids*. Their role is to carry the information that "codes" life. The force that created both the unity and the diversity of living things is evolution (Hunter, 2004).

Table 1: The four levels of protein conformation

Conformation Level	**Description**
Primary structure	The sequence of amino acids, forming a chain called polypeptide.
Secondary structure	The structure that forms a polypeptide after folding.
Tertiary structure	The stable 3D structure that forms a polypeptide.
Quaternary structure	The 3D structure formed by the conjugation of two or more polypeptides.

Proteins and nucleic acids are both called *macromolecules*, due to their large size compared to other molecules. Important efforts towards understanding life are made by studying the structure and function of biological macromolecules. The branch of biology concerned in this study is *molecular biology*.

Both proteins and nucleic acids are linear *polymers* of smaller molecules called *monomers*. The term sequence is used to refer to the order of monomers that constitute a macromolecule. A sequence can be represented as a string of different symbols, one for each monomer. There are twenty protein monomers called *amino acids*. There exist two nucleic acids, *deoxyribonucleic acid* (*DNA*) and *ribonucleic acid* (*RNA*). The DNA and RNA monomers are *nucleotides*. There are five different nucleotides depending on the nitrogen base they contain, namely *adenine* (*A*), *cytosine* (*C*), *guanine* (*G*), *thymine* (*T*), and *uracil* (*U*). DNA does not contain U, whereas RNA contains U instead of T. DNA is the genetic material of almost every living organism. RNA is the genetic material for some viruses such as HIV, but has also many functions inside a cell and plays an important role in protein synthesis (Table 2).

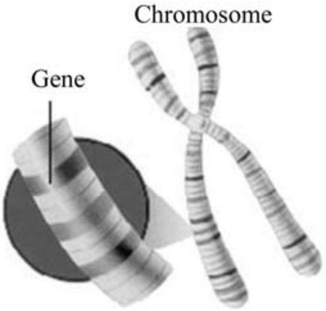

Figure 2: Gene-chromosome relationship

Organisms are classified in *eukaryotes* and *prokaryotes*. Eukaryotic cells contain a nucleus, whereas prokaryotic cells lack this structure. Eukaryotes include many organisms like animals, plants, fungi, and protists, whereas prokaryotes include bacteria and archaea. Although some prokaryotes are multicellular organisms, most of them contain a single cell.

The genetic material of an organism is organized in long double stranded DNA molecules called *chromosomes*. An organism may contain one or more chromosomes. A *gene* is a DNA sequence located in a particular chromosome and encodes the information for the synthesis of a protein or RNA molecule. The relationship between a gene and a chromosome is depicted in Figure 2. All the genetic material of a particular organism constitutes its *genome*.

Table 2: Some of the basic types of RNA

RNA Type	**Description**
Messenger RNA (mRNA)	Carries information from DNA to protein.
Ribosomal RNA (rRNA)	The main constituent of ribosomes, the cellular components where the protein synthesis takes place.
Transfer RNA (tRNA)	Transfers amino acids to ribosomes during protein synthesis.
Small nuclear RNA (snRNA)	It is found in the nucleus and is important in a number of processes including RNA splicing.

The central dogma of molecular biology, as coined and re-stated by Francis Crick (1958; 1970), describes the flow of the biological sequence information (Figure 3). The general transfers that appear in most organisms are described by the filled arrows. In particular, DNA is transcribed into RNA and then RNA is translated into protein. The circular arrow around DNA denotes its replication ability. Moreover, there are some more specific transfers. In retroviruses RNA is reverse transcribed into DNA. Also, some viruses can replicate their RNA. Finally, in the laboratory it is possible to directly translate DNA into a protein. These more special transfers are denoted by the unfilled arrows of Figure 3.

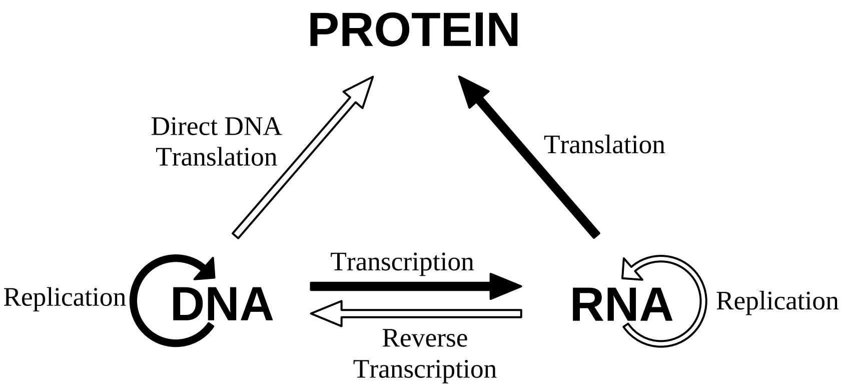

Figure 3: The flow of biological sequence information

BIOLOGICAL DATA ANALYSIS

Data mining deals with the discovery of useful knowledge from databases. It is the main step in the process known as Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996), although the two terms are often used interchangeably. Other steps of the KDD process are the selection, cleaning, and transformation of the data and the visualization and evaluation of the extracted knowledge. Some of the most popular tasks are classification, regression, clustering, association analysis, and sequence analysis. Depending on the nature of the data as well as the desired knowledge there is a large number of algorithms for each task. All of these algorithms try to fit a model to the data. Such a model can be either *predictive* or *descriptive*. A predictive model makes a prediction about data using known examples, while a descriptive model identifies patterns or relationships in data. Table 3 presents the most common data mining tasks.

Table 3: Common data mining tasks

Predictive	**Descriptive**
Classification. Maps data into predefined classes. **Regression.** Maps data into a real valued prediction variable.	**Association Analysis.** The production of rules that describe relationships among data. **Sequence Analysis.** Same as association, but sequence of events is also considered. **Clustering.** Groups similar input patterns together.

The algorithms for the predictive data mining tasks presented in Table 3 correspond to the supervised machine learning algorithms. Respectively, the algorithms for the descriptive data mining tasks correspond to the unsupervised machine learning algorithms. Depending on the purpose of the analysis one can select either a machine learning (interested in effectiveness) or a data mining (interested in efficiency and scalability or dealing with noisy data) algorithm.

Biological Sequence Analysis

As many genome sequencing projects have been completing, there is the need to annotate those genomes. The observed paradigm shift from static structural genomics to dynamic functional genomics (Houle et al., 2000) and the assignment of functional information to known sequences is deemed particularly important. Gene prediction is concerned with the identification of stretches of DNA that are biologically functional. It is the next step after the sequencing and is particularly important for the annotation and understanding of a genome. Gene prediction is not a straightforward task, especially for eukaryotic genomes, that are more complex. For example, eukaryotic genes consist of coding parts (*exons*) that are separated by intervening non-coding sequences called *introns*. Introns are removed from the transcribed RNA by the process of *splicing* (Figure 4). One of the most important biological problems that demand the construction of predictive data mining or machine learning models is the recognition of the splice sites, namely the boundaries between adjacent exons and introns. The problem becomes even more challenging, if one considers the possibility of alternative splicing, namely the production of different mature mRNA molecules, depending on the number of the exons that are finally concatenated. The phenomenon of alternative splicing is one of the most interesting findings of the last years and guided to the conclusion that a gene may code more than one protein products. Splicing does not take place in prokaryotes, since their genes lack introns. Another reason that gene prediction is easier in prokaryotic genomes is the presence of known conserved patterns around various signals, like promoters, transcription start sites, and translation initiation sites of prokaryotic sequences.

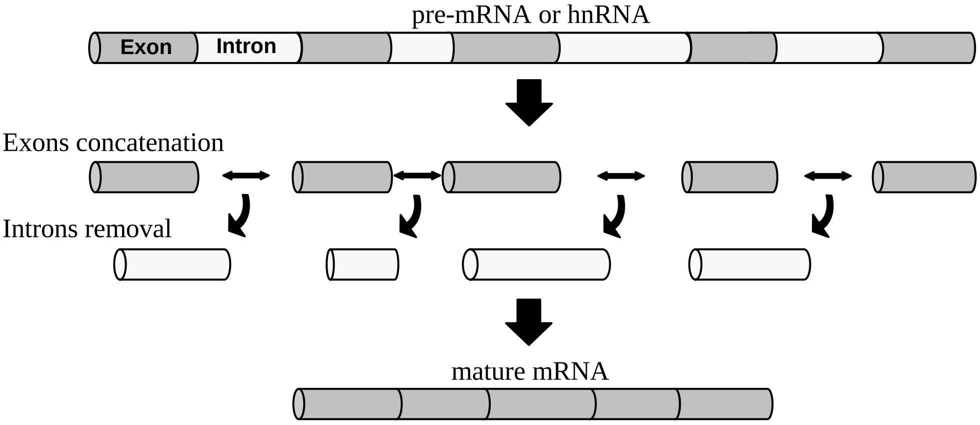

Figure 4: The process of RNA splicing

The most important sequence analysis tasks that exploit machine learning and data mining algorithms are the following:

- Prediction of regulatory regions (i.e. *promoters* and *enhancers*), which are segments of DNA where regulatory proteins bind preferentially and thus control gene expression and consequently protein expression.
- Prediction of the transcription start site, where the process of transcription starts.
- Prediction of the translation initiation site, where the process of translation initiates.
- Prediction of the splice sites for determining exons and introns
- Prediction of polyadenylation sites where a polyA (multiple adenines) tail is added at the mRNA sequence. Alternative polyadenylation makes the problem more challenging.
- Prediction of coding region in Expressed Sequence Tags (ESTs) that are short and partial sequences of transcribed spliced mRNAs.
- Comparison of a sequence with a database of known sequences for finding possible homologies (e.g. close evolutionary relationships) and grouping of structurally related sequences.

Many machine learning and data mining techniques have been utilized to deal with the above problems. Usually most of these techniques have to be modified and adapted so that can be efficiently and effectively applied to this kind of problems. The most common include neural networks, Bayesian classifiers, decision trees, and support vector machines, (Ma & Wang, 1999; Hirsh & Noordewier, 1994; Zien et al., 2000), as well as multiple classifier systems (Tzanis et al., 2007).

Gene Expression Analysis

Each organism contains a number of genes that code the synthesis of an mRNA or protein molecule. Every cell in an organism -with only few exceptions- has the same set of chromosomes and genes. However, two cells may have very different properties and functions. This is due to the differences in abundance of proteins. The abundance of a protein is partly determined by the levels of mRNA which in turn are determined by the expression levels of the corresponding gene. The process of conversion of the information encoded in a gene, first into mRNA and then to a protein structures and functions of a cell is called *gene expression*. A tool for analyzing *gene expression* is *microarray* or *gene chip*. A microarray experiment measures the relative mRNA levels of typically thousands of genes, providing the ability to compare the expression levels of different biological samples. These samples may correlate with different time points taken during a biological process or with different tissue types such as normal cells and cancer cells (Aas, 2001). Another method for measuring gene expression is Serial Analysis of Gene Expression (SAGE), which allows the quantitative profiling of a large number of mRNA transcripts (Velculescu et al., 1995). Although this method is more expensive than microarrays, it has the advantage that the experimenter does not have to select the mRNA sequences that will be studied.

The greatest challenge posed by gene expression data is that they contain a small number of samples (less than a hundred), and a very large number of features (genes), that is typically in thousands. Many feature selection approaches have been utilized for reducing the dimensionality of the data by selecting a small number of genes (see Xing et al., 2001). Moreover, a large number of genes are usually irrelevant and uninformative for the classification. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied.

Clustering is the far most used method in gene expression analysis. Clustering methods can be used to cluster genes with similar behavior or samples with similar gene expressions together. A category of clustering algorithms, called subspace clustering or bi-clustering algorithms, are used to simultaneously cluster genes and samples. Hierarchical clustering is another frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment (Yeung et al., 2003; Smolkin & Ghosh, 2003). External criteria have also been used for validating clusters quality based on predefined partitions of the data (Tzanis & Vlahavas, 2007).

Data Mining in Structural Bioinformatics

Structural bioinformatics is the subfield of bioinformatics which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules, especially for proteins. The application of machine learning and data mining in structural bioinformatics is quite challenging, since structural data are not linear. Moreover, the search space for most structural problems is continuous, namely infinite and demands highly efficient and heuristic algorithms.

Many machine learning and data mining algorithms have been utilized for the prediction of various protein properties such as active sites, modification sites, localization, stability, globularity, shape, protein domains, secondary structure and interactions (Whishart, 2002).

The most popular methods for this task are neural networks, nearest neighbor classifiers and hierarchical clustering algorithms.

Machine learning and data mining methods are also applied for protein secondary structure prediction. This problem has been studied for over than 35 years and many techniques have been developed. Initially, statistical approaches were adopted to deal with this problem. Later, more accurate techniques based on information theory, Bayes theory, nearest neighbors, and neural networks were developed. Combined methods such as integrated multiple sequence alignments with neural network or nearest neighbor approaches improve prediction accuracy.

Other important problems of structural bioinformatics that utilize machine learning and data mining methods are the RNA secondary structure prediction, the inference of a protein's function from its structure, the identification of protein-protein interactions and the efficient design of drugs, based on structural knowledge of their target.

Bioinformatics Text Mining

Text mining in molecular biology, defined as the automatic extraction of information about genes, proteins and their functional relationships from text documents (Krallinger and Valencia, 2005), has emerged as a hybrid discipline on the edges of the fields of information science, bioinformatics and computational linguistics.

It is critically important for biology researchers to have access to the most up to date information on their field of research. Current research practice involves on-line search for gene related information utilizing the latest technologies in Information Retrieval, Semantic Web and Text Mining. A new term lately used by bioinformatics experts to describe the text body where they can extract information such as ontology, interaction and function between biological entities is the *textome*. Generally, it can include all parseable and computable scientific text body.

The rapid progress in biomedical research has led to a dramatic increase in the amount of available information, in terms of published articles, journals, books and conference proceedings. Today, PubMed alone provides access to more than 14 million citations from MEDLINE and additional life sciences journals. In total, more than 4,800 journals are currently indexed by PubMed. Although PubMed is by far the richest database of abstracts, citations and full text articles, there is a plethora of such sources of scientific publications on Biology such as NCBI BookShelf for e-books and a large number of on-line resources. The researchers' need to exploit this enormous volume of available information, along with the avail of high performance and efficiency data mining, natural language processing and information retrieval tools has given birth to a new field of research and application, called Bioinformatics Text Mining (BTM). Other terms for BTM are Bio(logical) Text Mining and Biomedical Text Mining.

From a data miner's point of view, biomedical literature has certain characteristics that require special attention, such as heavy use of domain-specific terminology, polysemic words (word sense disambiguation), low frequency words (data sparseness), creation of new names and terms and different writing styles.

Several studies have categorized the tasks of BTM from different points of view. Cohen and Hersh (2005) provide a high-level categorization, identifying the main tasks to be the following:

- Named Entity Recognition (NER).
- Text classification.
- Synonym and abbreviation extraction.
- Relationship extraction.
- Hypothesis generation.

FUTURE TRENDS

Because of the special characteristics of biological data, the variety of new problems and the extremely high importance of bioinformatics research, a large number of critical issues is still open and demands active and collaborative research by the academia as well as the industry. Moreover, new technologies that permit the faster and cheaper conduction of large scale experiments led to a constantly increasing number of new questions on new data. Examples of hot problems in bioinformatics are the accurate prediction of protein structure and gene expression analysis. The recent discovery of the role of a small RNA molecule, called *microRNA* or *miRNA*, has posed new questions. MicroRNA interferes with gene expression by either splitting an mRNA into tiny useless pieces, or preventing it from translating into proteins. In both cases the gene coding for this mRNA is not expressed. Among the most challenging applications of bioinformatics is molecular medicine. Disease prognosis, early disease diagnosis, and therapy response are the current trends. The analysis of a patient's genetic profile makes possible the personalized medicine, which guides clinicians to select the most appropriate medication. Machine learning and data mining have been developing and improving in order to deal efficiently and effectively with the problems posed by the data explosion in biology. As Houle et al. (2000) mention, these improvements will enable the prediction of protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways and signaling cascades. The final objective of such analysis will be the illumination of the way conveying from genotype to phenotype and the specification of the molecular and cellular details that govern life.

CONCLUSION

The recent technological advances, have led to an exponential growth of biological data. New questions on these data have been generated. Scientists often have to use exploratory methods instead of confirming already suspected hypotheses. Machine learning and data mining are two relative research areas that aim to provide the analysts with novel, effective and efficient computational tools to overcome the obstacles and constraints posed by the traditional statistical methods. Feature selection, normalization of the data, visualization of the results and evaluation of the produced knowledge are equally important steps in the knowledge discovery process. The mission of bioinformatics as a new and critical research domain is to provide the tools and use them to extract accurate and reliable information in order to gain new biological insights.

REFERENCES

Aas, K. (2001). Microarray Data Mining: A Survey. NR Note, SAMBA, Norwegian Computing Center.

Cohen, A.M. & Hersh, W.R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71.

Crick, F.H.C. (1958). On protein synthesis. *Symposium of the Society for Experimental Biology XII*, 139-163.

Crick, F.H.C. (1970). Central Dogma of Molecular Biology. Nature, 227, 561-563.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. AAAI Press/MIT Press, Menlo Park, California, USA.

Hirsh, H. & Noordewier, M. (1994). Using Background Knowledge to Improve Inductive Learning of DNA Sequences. Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications, 351-357.

Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. & Lundahl, S. (2004. March 10). Database Mining in the Human Genome Initiative. Whitepaper, Bio-databases.com, Amita Corporation. Available: http://www.biodatabases.com/ whitepaper.html

Hunter, L. (2004). Life and Its Molecules: A Brief Introduction. AI Magazine, 25(1), 9-22.

Krallinger, M. & Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. Genome Biology, 6(224).

Luscombe, N.M., Greenbaum, D. & Gerstein, M. (2001). What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods of Information in Medicine, 40(4), 346-358.

Ma, Q. & Wang, J.T.L. (1999). Biological Data Mining Using Bayesian Neural Networks: A Case Study. International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 8(4), 433-451.

Smolkin, M. & Ghosh, D. (2003). Cluster Stability Scores for Microarray Data in Cancer Studies. BMC Bioinformatics, 4, 36.

Tzanis, G., Berberidis, C., & Vlahavas, I. (2007). MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction. Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Lyon, France.

Tzanis, G. & Vlahavas, I. (2007). Mining High Quality Clusters of SAGE Data. Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics, Vienna, Austria.

Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. (1995). Serial Analysis of Gene Expression, Science, 270(5235), 484-487.

Whishart, D.S. (2002). Tools for Protein Technologies. In Sensen, C.W. (Ed.), Biotechnology, (Vol 5b) Genomics and Bioinformatics, 325-344, Wiley-VCH.

Xing, E.P., Jordan, M.I., & Karp, R.M. (2001). Feature selection for high-dimensional genomic microarray data. Proceedings of the 18th International Conference on Machine Learning, 601-608.

Yeung, Y.K., Medvedovic, M. & Bumgarner, R.E. (2003). Clustering Gene-Expression Data with Repeated Measurements. Genome Biology, 4(5), R34.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. & Muller, R.-K. (2000). Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. Bioinformatics, 16(9), 799-807.

KEY TERMS

Genotype – The exact genetic makeup of an organism.
Gene Mapping – The process of creating a genetic map assigning DNA fragments (genes) to chromosomes.
Gene regulation – The cellular control of the time that a gene will be activated and the amount of gene product that will be produced.
Phenotype – The physical appearance characteristics of an organism.

Sequence Alignment – The process to test for similarities between a sequence of an unknown target protein, and a single (or a family of) known protein(s).
Sequencing – The process of determining the order of nucleotides in a DNA or RNA molecule or the order of amino acids in a protein.
Transcript - A sequence of messenger RNA.