

Méthodes d'apprentissage automatique pour la bioinformatique

BIOL-F-524

Gianluca Bontempi

Département d'Informatique
Boulevard de Triomphe - CP 212
<http://www.ulb.ac.be/di>

Machine learning for bioinformatics

Objectives: Present both conventional and recent techniques for creating models from bioinformatics data.

Parts in the course:

- Introduction to data analysis for bioinformatics (2 hours)
- Classification techniques (3 hours)
- Classification algorithms (5 hours)
- Feature selection (2 hours)

Prerequisites

- Notion of probability (conditional probability, Bayes theorem)
- Notions of estimation (bias, variance)
- Principal component analysis
- Notions of programming

The molecular biology dogma

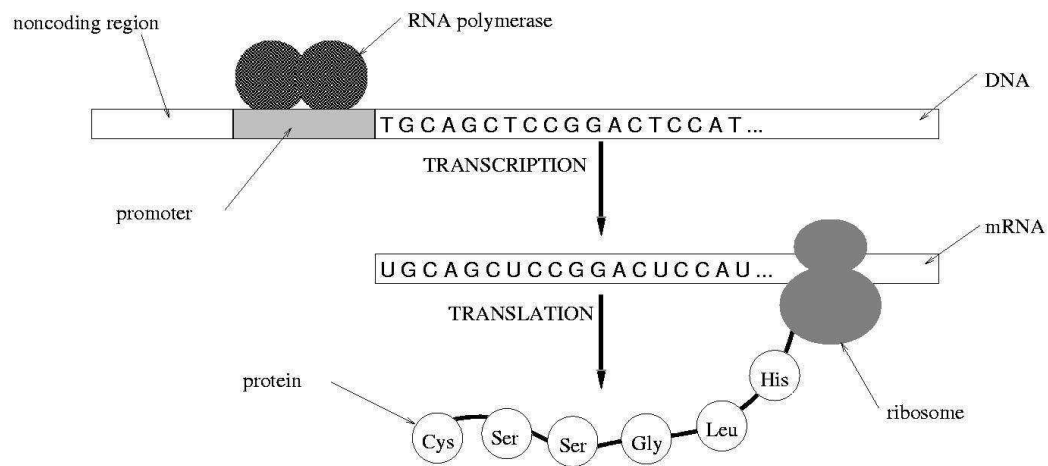


Figure 1: **The process of gene expression.** DNA and RNA are composed of linear chains of nucleotides. *Transcription* involves synthesizing messenger RNA (mRNA) using DNA as a template. The enzyme *RNA polymerase* is the molecule that transcribes DNA into RNA. A site where RNA polymerase binds to DNA to begin transcription is called a *promoter*. Messenger RNA is *translated* to protein by a molecule called a ribosome. Proteins are linear chains of amino acids; each amino acid is encoded by a string of three consecutive nucleotides. *Noncoding regions* are stretches of DNA that do not encode proteins.

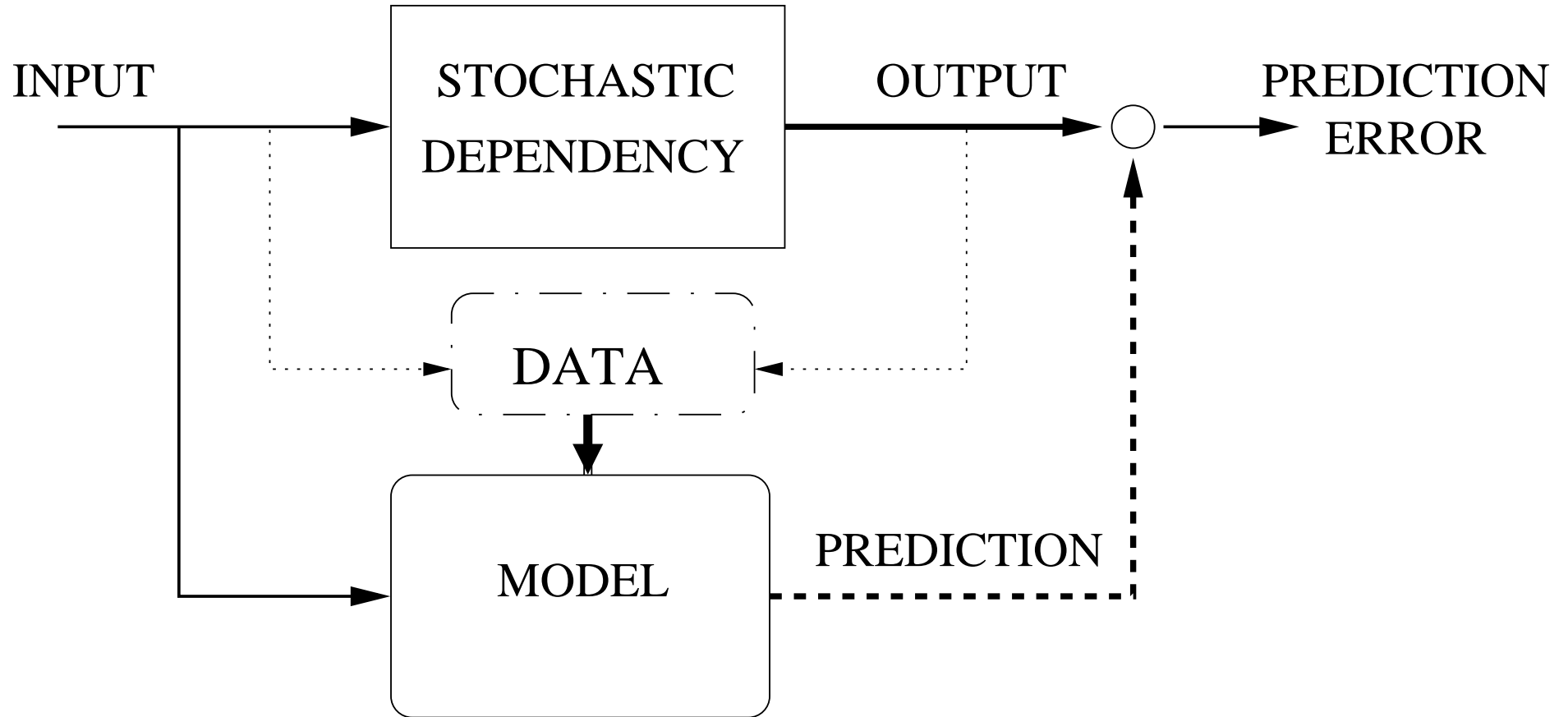
Data analysis in bioinformatics

- Within last years the complete sequence has been determined for a number of genomes from humans and other organisms.
- Determining the nucleotide sequence of a DNA molecule, however, is only a first step towards the ultimate goals of
 1. understanding the functionality
 2. knowing the locations of all the genes and regulatory sites of the molecule.
- The result of sequencing efforts and the availability of new measurement tools (e.g. microarrays) makes a great volume of data available for analysis.
- This has created the need for (semi) automated methods to analyze massive datasets. Data analysis methods are expected to support biologists in discovering patterns, understanding correlations, reducing complexity, predicting events. This is often referred to as *knowledge discovery*.

Biological applications of prediction methods

- Prediction of protein sub-cellular localization.
- Prediction of translation initiation sites.
- Prediction of protein secondary structure.
- Prediction of protein-protein interactions from primary structure.
- Gene recognition
- Gene classification from micro-array expression data
- Tissue classification from micro-array expression data
- Regulatory element detection using correlation with expressions

Prediction by supervised learning



Prediction by supervised learning

- The prediction problem is also known as the **supervised learning** problem, because of the presence of the outcome variable which guides the learning process.
- Collecting a set of training data is analogous to the situation where a teacher suggests the correct answer for each input configuration.
- What is peculiar of supervised learning is
 1. the lack of (parametric) knowledge of the (possibly nonlinear) phenomenon underlying the data,
 2. the availability of only a finite and noisy dataset to estimate the model,
 3. the priority given to an accurate prediction wrt interpretability of the model.

Prediction by supervised learning

The use of supervised learning techniques for prediction requires the definition of several aspects:

- the output variable to be predicted,
- the set of input features expected to provide useful information for predicting the output,
- the training set,
- the family of models used to approximate the unknown input/output relationships,
- the criteria to assess the quality of the prediction model.

The following examples will show how different choices have been taken in existing applications of supervised learning techniques to bioinformatics problems.

Protein localization

- One of the main tasks of proteomics is the assignment of functionalities to sequenced proteins. The assignment of a function for a given protein has proved to be especially difficult where no clear homology to proteins of known function exists.
- One field of proteomics that has recently received a lot of attention recently is *protein localization*.
- Protein expression analysis can indicate **whether** proteins are expressed, but it is also important to know **where** proteins are expressed, and where they go over time.
- Knowing the sub-cellular location that a protein resides in may give important insights as to its possible function.
- Even when the basic function of a protein is known, knowing its location in the cell may give insights as to which pathway an enzyme is part of.

Protein localization (II)

- There is an increasing shift away from general protein expression analysis and toward mapping proteins distribution, relative abundance, tissue specificity, and movement.
- By tracking these parameters (in healthy versus diseased tissue and in control versus treated tissue), researchers can gain a greater understanding of these proteins functions and determine which are likely to be the best drug targets.

Prediction of protein localization

- Sub-cellular localization is a key functional characteristic of proteins. [3]
- A fully automatic and reliable prediction system for protein sub-cellular localization would be very useful.
- Two types of prediction methods have been developed:
 1. based on the recognition of protein N-terminal sorting signals
 2. based on amino-acid composition.
- In both cases protein sub-cellular localization is seen as a multi-class classification problem.

Prediction of localization by signal

- All new proteins in the cell have a tag (signal peptide) on them, telling whether the protein is to be sent out of the cell or to a special part in the cell. By comparing tags from known proteins we can find out where an unknown protein will be located.
- Supervised learning methods (neural networks) have been used for eukaryotic species to discriminate between proteins destined for the mitochondrion (mTP), the chloroplast (cTP), the secretory pathway (SP), and other localization [4] on the basis of N-terminal sorting sequence information.
- The reliability of these predictive methods is strongly dependent on the quality of the protein N-terminal assignment. The methods are inaccurate when the signals are missing or only partially included.

Prediction of localization by composition

- It was shown that intra-cellular and extracellular proteins differ significantly in their amino acid composition.
- As consequence, alternative prediction approaches (neural networks in [10] and support vector machines in [6]) focus on the study of the correlation of amino acid composition with different sub-cellular localizations.
- Twenty input features, one for the fraction of each amino acid are used.
- A method based on the amino acid composition are expected to be comparatively stable to wrong sequence assignment.
- Note that the output to be predicted (i.e. the localization) remains the same but the set of input features is changed !

Predicting protein-protein interactions

- A goal of proteomics is to elucidate the structure, interactions and functions of all proteins within cells and organisms
- The interaction between proteins is fundamental to a broad spectrum of biological functions (e.g. regulation of metabolic pathways, immunologic recognition, DNA replication, protein synthesis).
- Whether or not two proteins will bind to form a stable complex that is prerequisite to biological function is dependent on the three-dimensional conformations of the proteins. At the same time the sequence specifies the conformation.
- Computational techniques could represent an alternative to conventional proteomics methods known to be tedious, labor intensive and potentially inaccurate.

Predicting protein-protein interactions (II)

- The authors of [1] studied if the knowledge of the amino acid sequence alone might be sufficient to estimate the propensity for two proteins to interact.
- Given a database of known protein-protein interaction pairs, a machine learning system is trained to recognize interactions on the basis of the primary structure and associated physiochemical properties.
- Protein interaction data were obtained from the Database of Interacting Proteins.
- For each amino-acid sequence feature vectors were assembled from encoded representations of tabulated residue properties (e.g. charge, hydrophobicity, surface tension).
- Input patterns are obtained by concatenating the vectors of features of the interacting proteins. Negative examples were obtained by randomizing amino acid sequences.

Gene classification

- Genome researchers are shifting their focus from *structural* genomics to *functional* genomics.
- Structural genomics is the initial phase of genome analysis, whose goal is to construct high resolution genetic and physical maps as well as complete sequence information of the chromosomes.
- Functional genomics is the second phase, aiming at studying the functionality of genes of a single organism as well as studying and correlating the functionality of genes across many different organisms.
- The traditional approach to functional genomics consists in using sequence data to determine the function of genes and/or the corresponding proteins. The idea is that genes with sufficient similar sequences also perform similar functions.

Gene classification (II)

- However, sometimes sequence comparisons can be uninformative and misleading as well as there a lot of species for which we do not have complete sequence information.
- Recently methods have been developed for monitoring genome-wide mRNA expressions: oligonucleotide chips, SAGE (serial analysis of gene expression) and microarrays.
- These tools allows to observe expression levels of the entire genome under many different induced conditions.

Gene classification (III)

- Knowing when and under what conditions a gene or a set of genes is expressed often provides strong clues as to their biological role and function.
- One way of using the data produced by microarray experiments to determine the function of unknown genes is to use clustering algorithms to group together genes that have similar expression profiles. Based on the distribution of known and unknown genes in such clusters, some information about the function of previously unknown genes can be inferred.
- An alternative is provided by supervised learning methods. The key advantage of supervised over unsupervised methods is that the predictive precision of these methods can be quantified.
- The authors of [2] used several classification algorithms to predict if a gene has a particular function based on expression profiles.

Gene classification (III)

- Authors of [2] used class definitions made by the MIPS Yeast Genome Database.
- Six functional classes were considered: tricarboxylic acid cycle (TCA), respiration, cytoplasmic ribosomes, proteasome, histones and helix-turn-helix proteins.
- A learning machine was trained to predict the correct class on the basis of the microarray expression data.

	E1	E2	E3	En	Y
G1								c_1
G2								c_1
G3								c_3
...								c_2
...								c_3
...								c_4
...								c_2
...								c_1
GN								c_3
G_q								?

Patient classification (breast cancer)

- Breast cancer is one of the most common malignant tumors affecting women.
- Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome.
- Cancer classification has been based primarily on morphological appearance of the tumor, but with serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. The strongest predictors for metastasis fail to classify accurately breast tumors according to their clinical behavior.
- Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes.

Breast cancer classification (II)

- Chemiotherapy or hormonal therapy reduces the risk of distant metastasis by approximately one-third; however 70-80% of patients receiving this treatment would have survived without it. Also, these therapies frequently have toxic side effects.
- Diagnosis of cancer must be accurate in order for the patient to receive the correct treatment and so have the best chance of survival.
- The cellular and molecular heterogeneity of breast tumors and the large number of genes potentially involved in controlling cell growth, death and differentiation emphasize the importance of studying multiple genetic alterations in concert.
- The development of microarray technology provides the opportunity of correlating genome-wide expressions with the response of tumor cells to chemotherapy.

Breast cancer classification (III)

- Systematic investigation of expression patterns of thousands of genes in tumors using DNA microarrays and their correlation to specific features of phenotypic variation might provide the basis for an improved taxonomy of cancer.
- It is expected that variations in gene expression patterns in different tumors could provide a “molecular portrait” of each tumor, and that the tumors could be classified into subtypes based solely on the difference of expression patterns.
- The authors of [12] applied classification techniques to identify a gene expression signature strongly predictive of a short interval to distant metastases in patients without tumor cells in local lymph nodes at diagnosis.

	G1	G2	G3	Gn	Y
P1								c_1
P2								c_1
P3								c_3
...								c_2
...								c_3
...								c_4
...								c_2
...								c_1
PN								c_3
P_q								?

Input/output problems

All the previous examples are characterized by

1. An outcome measurement, also called **output**, usually quantitative (like the gene expression) or categorical (like metastasis or not).
2. a set of **features** or **inputs**, also quantitative or categorical, that we wish to use to predict the output.

If we suppose that we have available a set of input/output data, also called **training set**, we could use statistical methods to build a **prediction model** or **learner** which could enable us to predict the outcome for **new unseen objects**.

Regression and classification

According to the type of output we can distinguish between two types of prediction tasks:

Regression when we predict quantitative outputs, e.g. real or integer numbers

Classification (or pattern recognition) where we predict qualitative or categorical outputs which assume values in a finite set of classes (e.g. black, white and red) where there is no explicit ordering. Qualitative variables are also referred to as **factors**.

Notation

- In order to clarify the distinction between random variables and their values, we will use the **boldface notation for denoting a random variable** (e.g. \mathbf{z}) and the normal face notation for the eventually observed value (e.g. $z = 11$).
- The notation $P_{\mathbf{z}}(z) = \text{Prob} \{ \mathbf{z} = z \}$ denotes the probability that the discrete random variable \mathbf{z} take the value z .
- The notation $F_{\mathbf{z}}(z)$ denotes the probability

$$\text{Prob} \{ \mathbf{z} \leq z \}$$

that the continuous random variable \mathbf{z} take the value $\leq z$. The suffix indicates that the probability relates to the random variable \mathbf{z} . This is necessary since we often discuss probabilities associated with several random variables simultaneously.

- Example: \mathbf{z} could be the age of a student before asking and $z = 22$ could be his value after the observation.

Notation

- N is used to denote the number of observations.
- n is used to denote the number of variables.
- $\{z_1, \dots, z_N\} \leftarrow F_{\mathbf{z}}(\cdot)$ means that the random sample of size N $\{z_1, \dots, z_N\}$ has been i.i.d. generated from the distribution $F_{\mathbf{z}}(\cdot)$.
- D_N is used to denote the *training set* or more generally the set of observations available for estimation and/or prediction.

References

- [1] J. R. Bock and D. A. Gough. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.
- [2] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. In *Proc. Natl. Acad. Sci.*, volume 97, pages 262–267, 2000.
- [3] F. Eisenhaber and P. Bork. Wanted: subcellular localization of proteins based on sequence. *Trans. Cell. Biol.*, 8:169–170, 1998.
- [4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.*, 300:1005–1016, 2000.
- [5] R. Farber, A. Lapedes, and K. Sirotkin. Determination of eucaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology*, 226:471–479, 1992.

- [6] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.
- [7] A. Lapedes, C. Barnes, C. Burks, R. Farber, and K. Sirok-tin. Application of neural networks and other machine learning algorithms to dna sequence analysis. In G. Bell and T. Marr, editors, *Computers and DNA*, volume VII, pages 157–182. Addison Wesley, 1989.
- [8] A.G. Pedersen and H. Nielsen. Neural network prediction of translation initiation sites in eukaryotes: perspectives for est and genome analysis. In T. Gasterland, editor, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, volume 5, pages 226–233. AAAI Press, 1997.
- [9] N. Qjan and T. J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884, 1988.
- [10] A. Reinhardt and T. Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acid Research*, 26(9):2230–2236, 1998.
- [11] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the perceptron algorithm to distinguish

translational initiation sites in *e. coli*. *Nucleic Acid Research*, 10(9):2297–3011, 1982.

- [12] L. J. van't Veer, H. Dai, and M. J. van de Vijver. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
- [13] A. Zien, G. RAtsch, S. Mika, B. Scholkops, T. Lengauer, and K.-R. Muller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.