Machine Learning in Bioinformatics NAIL107

Part 1: Introduction to the lecture

František Mráz
KSVI MFF UK

Outline

1. Introduction
2. Course logistics
3. Course objectives
4. What is bioinformatics
5. What is machine learning
6. Why machine learning and bioinformatics
7. Types of machine learning
8. Objectives of machine learning, of this course and related fields
9. Definitions and terminology
10. Learning scenarios
11. Sample problem
12. Cross-Validation
13. Biology minimum

1. Introduction

- **Teacher:**
  - František Mráz, KSVI MFF UK,
    Malostranské nám. 25, Praha 1, room 405
  - e-mail: frantisek.mraz@mff.cuni.cz
- **Literature:**
  - Yang, Zheng Rong. *Machine learning approaches to bioinformatics*. Vol. 4. World Scientific, 2010
    - Accessible via library *COMPUTER-SCIENCEnetBASE* at http://www.mff.cuni.cz/fakulta/lib/
  - Alpaydin, Ethem: *Introduction to machine learning, second edition*, 2010, The MIT Press
- **Further Teaching Materials:**
  - Moodle course at
    - http://dl1.cuni.cz/course/view.php?id=2765
    - password for both guest access and sign-in is ***MLiB19***

2. Course Logistics

- **Lecture:** Friday, 10:40-12:10, S9
- **Seminar:** Wednesday, 12:20 – 13:50, SW2
  - **3 homework assignments**
    - Correct in time submitted solution – 10 credits
    - late submission of a task penalized – 1 credit for each started week of the delay
  - **quizzes** (on-line tests)
    - each quiz has a deadline; usually 3 attempts possible
    - results just after submitting
    - no late submission allowed
  - **a term project finished by a presentation** – max. 15 credits
- **Exam:**
  - all possible credits from the seminar will make 40% of the score of the final exam
  - marking scheme: **1**: ⟨100,85), **2**: ⟨85,70), **3**: ⟨70,55),

3. Course Objectives

- Present interesting and challenging biological machine learning problems
  - Data: High dimensional, heterogeneous, non-vectorial
  - Require both standard and non-standard machine learning solutions
- Machine learning methods for solving them

4. What is Bioinformatics

- Various definitions in the literature
  - "the use of computational methods to study biological data" [T.K. Attwood, Parry-Smith, D.J., Introduction to bioinformatics . Essex: Addison Wesley Longman Ltd, 1999]
  - "the application of computational methods to DNA and protein science" [D.W. Mount, Bioinformatics, Sequence and Genome Analysis . New York: Cold Spring Harbor Laboratory Press, 2001]
  - "the field of science in which biology, computer science, and information technology merge into a single discipline" [The National Center for Biotechnology Information (NCBI)]
  - "a multi-discipline, inter-discipline, and cross-discipline science for understanding biological systems, exploring underlying mechanisms of biological complexes, verifying biological hypotheses and providing evidence through *in silico* simulation for further theoretical development" [Z. R. Yang, *Machine learning approaches to bioinformatics*. Vol. 4. World Scientific, 2010]

4. What is Bioinformatics

- Recently, the complete DNA 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 the first step towards the ultimate goals
  1. understanding the functionality of DNA, genes, …
  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*.

5. What is Machine Learning?

- Machine learning = *computer science methods to improve performance criterion using example data or past experience*
  - *Experience:* data-driven task, thus statistics, probability
  - *Computer science:* we need efficient and accurate algorithms, analysis of complexity, theoretical guaranties

- There is no need to "learn" how to calculate salary
- Learning is used when:
  - human expertise does not exist (navigating on Mars),
  - humans are unable to explain their expertise (face recognition),
  - solution changes in time (routing on a computer network),
  - solution needs to be adapted to particular cases (user biometrics).

5. What is Machine Learning?

- **Examples:**
  - OCR – optical character recognition
  - Document classification, spam detection
  - Speech recognition, speech synthesis, speaker recognition
  - Image recognition, face detection
  - Fraud detection (credit card), network intrusion detection
  - Autonomous control of a vehicle (robot, car)
  - Medical diagnosis
  - Recommendation systems, search engines, information extraction

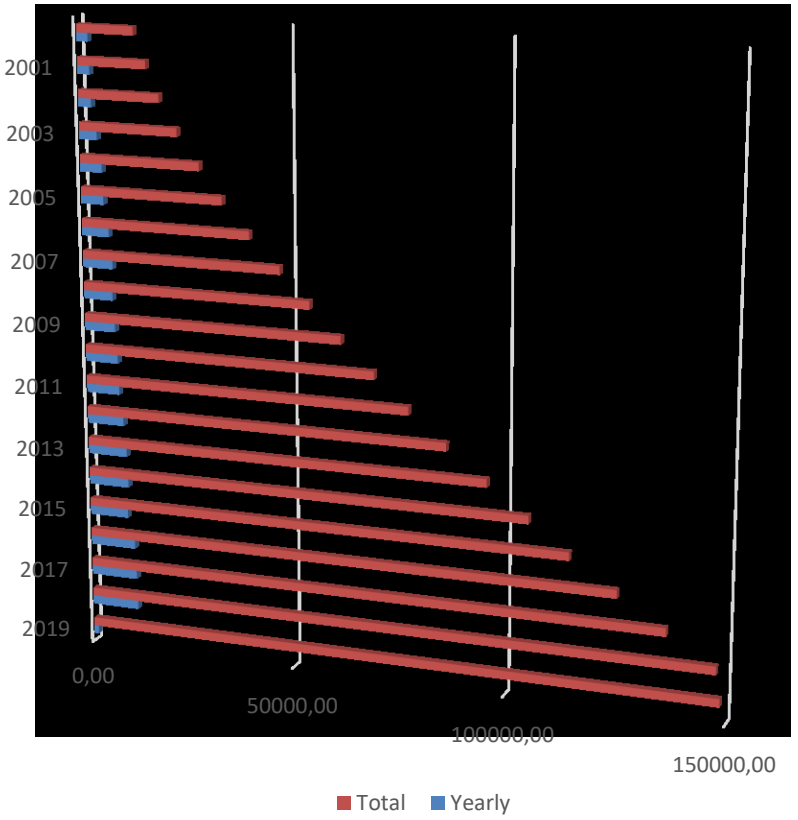
5. What is Machine Learning?

- Many interesting problems in computer science are extremely complex and it is often difficult or even impossible to program directly a solution
- Think how to implement a program able to: recognize a face in a photo, to decide whether an email is a spam, recognize handwriting, categorize a news.
- The same happens in bioinformatics. Consider the problem of recognizing genes in a DNA sequence or inferring the property of a protein from its structure.
- Machine learning offers an alternative methodological approach to deal with these problems.
- By exploiting the knowledge extracted from a sample of data it is possible to design algorithms able to solve this kind of problems.

6. Why Machine Learning and Bioinformatics

- Exponentially growing amount of biological data
- The **Protein Data Bank** (**PDB**) is a repository for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or nuclear magnetic resonance spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible. The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

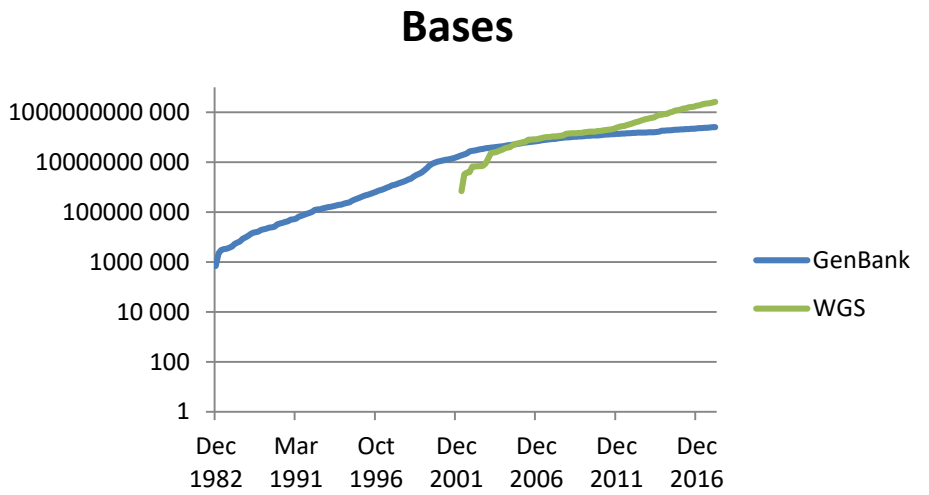
- Protein Data Bank (PDB) growth

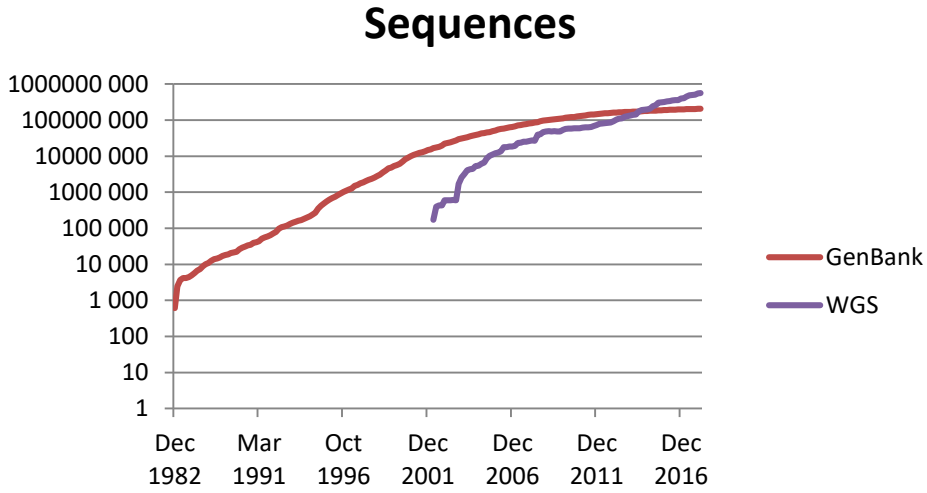


6. Why ML+BI
GenBank Growth Chart

- GenBank stores sequences and related data
- In December 2018
  - 2.85×10^{11} bases in 2.11×10^{8} sequences
  - 3.66×10^{12} bases in 7.73×10^{8} sequence records for **w**hole **g**enome **s**equences (WGS)

- Source: https://www.ncbi.nlm.nih.gov/genbank/statistics/





7. Types of Machine Learning

- **Density estimation:** learning probability distribution according to which data has been sampled (distribution typically selected out of pre-selected family).
- **Dimensionality reduction:** find lower-dimensional manifold preserving some properties of the data (computer vision).
- **Clustering:** partition data into homogenous groups (analysis of very large data sets).
- **Classification:** assign a category to each object (OCR, text classification, speech recognition).
- **Regression:** predict a real value for each object (prices, stock values, economic variables, ratings).
- **Ranking:** order objects according to some criterion (relevant web pages returned by a search engine).

8. Objectives of Machine Learning

- **Algorithms:** design of efficient, accurate, and general learning algorithms to
  - deal with large-scale problems (|data| > 1-10M),
  - make accurate predictions (unseen examples),
  - handle a variety of different learning problems.
- **Theoretical questions:**
  - What can be learned? Under what conditions?
  - How well can it be learned computationally?

8. Objectives of this Course

- **Algorithms:** covers several key learning algorithms.
  - nearest-neighbor algorithms,
  - Hidden Markov models,
  - perceptron, neural networks,
  - support vector machines, kernel methods,
  - boosting, bagging.
- **Applications:** mainly from bioinformatics
  - illustration of the use of algorithms,
  - programming, hands-on experience (seminar).
- **Theory:**
  - complexity analysis and introduction to concepts.

8. Related Fields

- **Statistics**
  - Going from particular observations to general descriptions = *inference*
  - learning = *estimation*
- **Engineering**
  - Classification = *pattern recognition;* often non-parametric and much more empirical
- **Data mining**

  = application of machine learning algorithms to large amounts of data (big data), in the business world

  = knowledge discovery in databases (KDD), in computer science

9. Definitions and Terminology

- **Example:** an object or instance in used data.
- **Features:** the set of attributes, often represented as a vector, associated to an example, e.g., height and weight for gender prediction.
- **Labels:**
  - in classification, category associated to an object, e.g., positive or negative in binary classification;
  - in regression, real-valued numbers.
- **Training data:** data used for training algorithm.
- **Test data:** data exclusively used for testing algorithm.

10. Learning Scenarios

- **supervised learning:**
  - labeled training data
  - Goal: to determine labeling of new data
  - Finite set of labels – *classification*
  - Infinite set of labels (real numbers) – *regression*
- **unsupervised learning:** no labeled data
  - Goal: to group similar data
- **semi-supervised learning:**
  - a small amount of labeled data with a large amount of unlabeled data
  - E.g. assuming that points which are close to each other are more likely to share a label

11. Example – Spam Detection

- **Problem:** classify each e-mail message as SPAM or non-SPAM (binary classification problem).
- **Data:** large collection of SPAM and non-SPAM messages (labeled examples).
- **Features:** define features for all examples (e.g., presence or absence of some sequences of words).
  - critical step (should use prior knowledge).
- **Algorithm:** choose type of algorithm adapted to the problem.
  - typically requires choice of hypothesis
- **Learning stages:**
  - Divide labeled collection into training and test data.
  - Use training data and features to train machine learning algorithm.
  - Predict labels of examples in test data to evaluate algorithm.
  - Algorithms may require choosing a parameter (number of rounds, learning parameter, trade-off parameter) validation set or cross-validation.

12. Cross-Validation

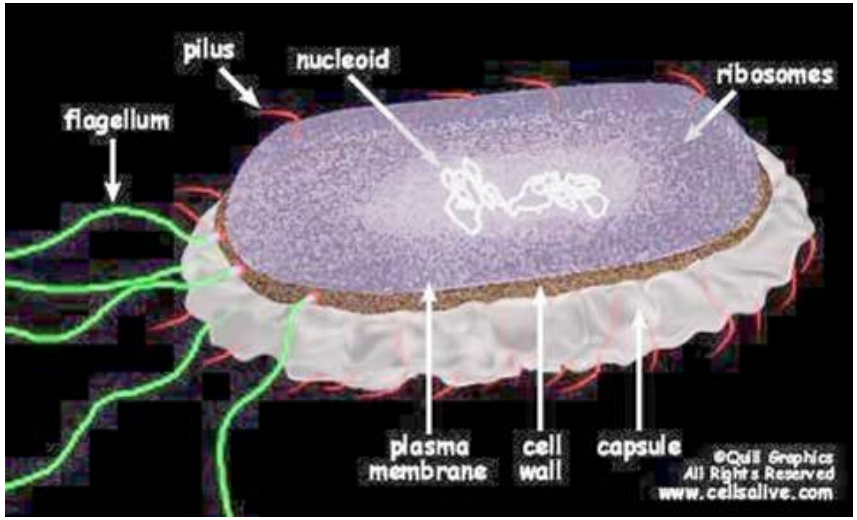
- Partition data into K folds (typically, 5 or 10).
- Train on all but k-th fold $\rightarrow$ hypothesis $h_{\theta,k}, k \in [1, K]$.
- Compute fold cross validation error:

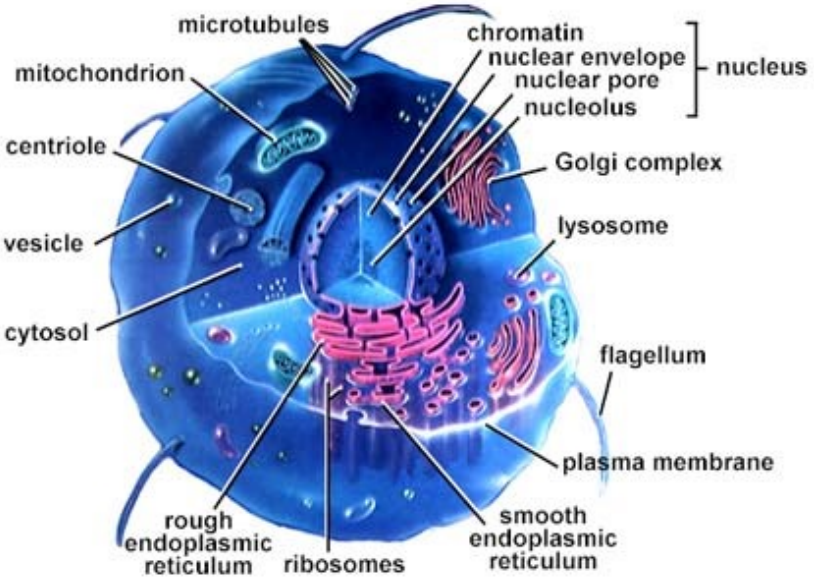
$$\frac{1}{K}\sum_{k=1}^{K} error\left(h_{\theta,k}, \text{fold } k\right)$$

- Where $error\left(h_{\theta,k}, \text{fold } k\right)$ is the error of the hypothesis on fold k
- Choose value of θ minimizing CV error
- When $K = m$ (sample size) leave-one-out cross-validation and error.

BIOLOGY MINIMUM

Prokaryotes vs. Eukaryotes

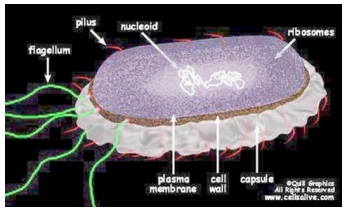




Prokaryotes vs. Eukaryotes

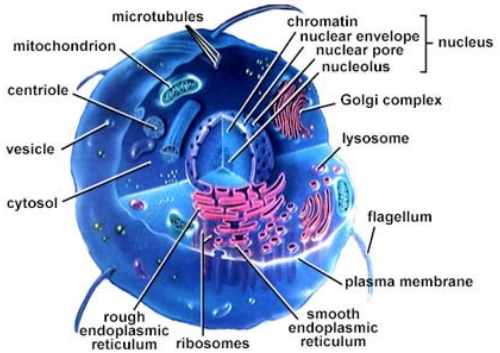
Prokaryotes

- Single cell
- No nucleus
- No organelles
- One piece of circular DNA
- No mRNA post transcriptional modification



Eukaryotes

- Single or multi cell
- Nucleus
- Organelles
- Chromosomes
- Exons/Introns splicing



Prokaryotes vs. Eukaryotes

Prokaryotes

- Eubacterial (blue green algae) and archaebacteria
- only one type of membrane – plasma membrane forms
  - the **boundary** of the proper cell
- The smallest cells known are bacteria
  - Ecoli (Escherichia coli) cell
  - $3x10^6$ protein molecules
  - 1000–2000 polypeptide species.

Eukaryotes

eukaryotic organisms that are not an animal, plant or fungi

- plants, animals, Protista, and fungi
- complex systems of internal membranes forms
  - organelles and compartments
- The volume of the cell is several hundred times larger
  - Hela cell

an immortal cell line used in scientific research

  - $5x10^9$ protein molecules
  - 5000–10,000 polypeptide species

Terminology

- **Genome:** an organism's complete set of DNA
  - a bacteria contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion
  - human genome has 24 distinct chromosomes
  - Each chromosome contains many genes.
- **Gene**
  - basic physical and functional units of heredity
  - specific sequences of DNA bases that encode instructions on how to make **proteins**
- **Proteins**
  - Make up the cellular structure
  - large, complex molecules made up of smaller subunits called **amino acids**

Molecules of Life

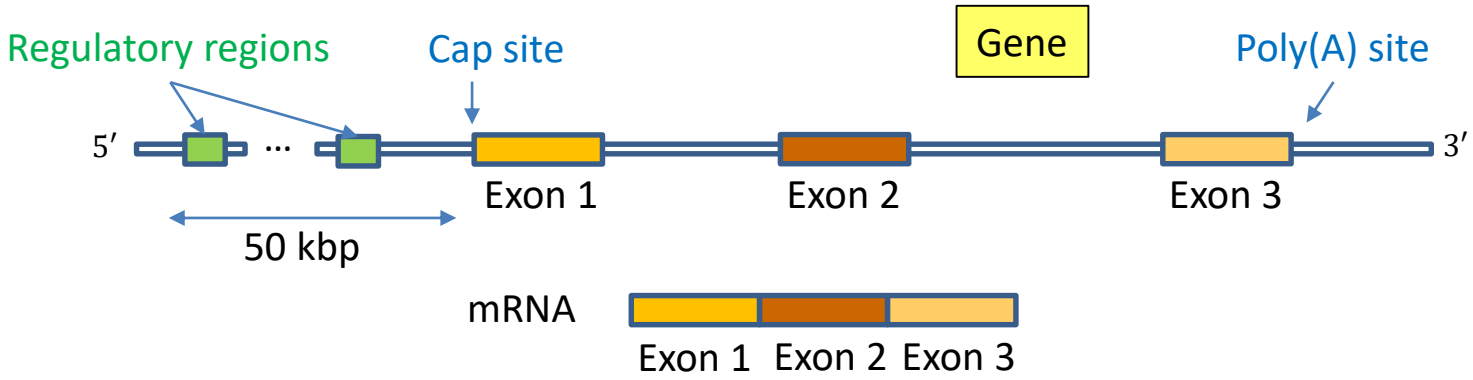
- DNAs
  - Hold information on how cell works
  - Two complementary strands – they are "read" in opposite direction
- RNAs
  - Act to transfer short pieces of information to different parts of a cell
  - Provide templates to synthesize proteins
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g. hair, skin, etc.)

Definition of a Gene

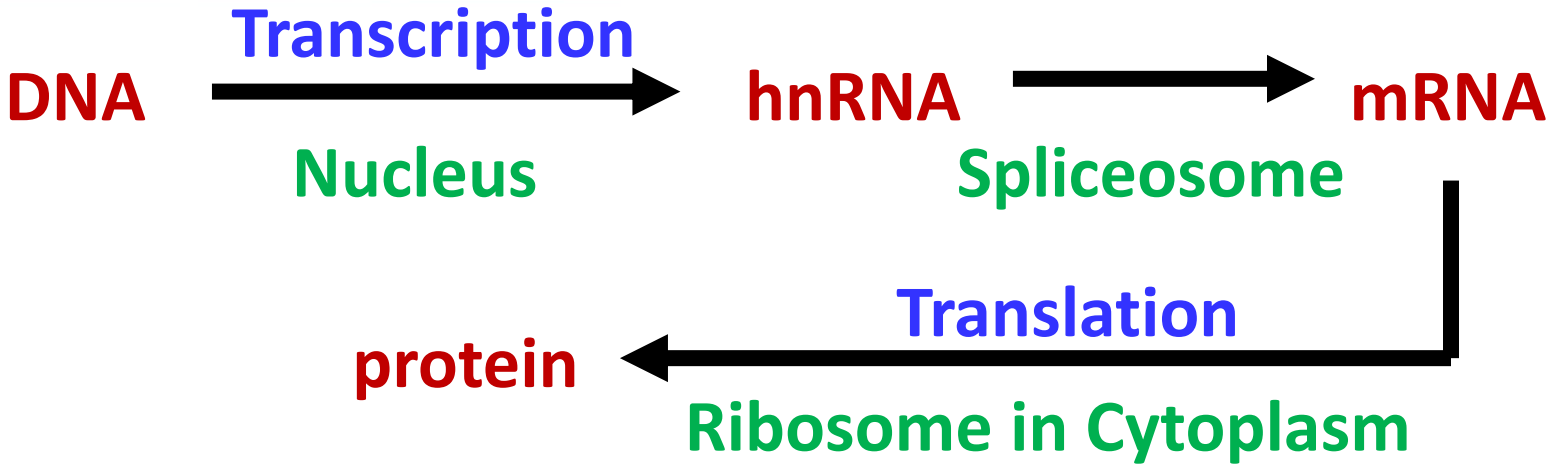
- **Regulatory regions:** up to 50 kb before a gene
- **Exons:** protein coding and untranslated regions UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17,000 bp per exon (mean 145 bp)
- **Introns:** splice acceptor and donor sites, junk DNA
  average 1–50,000 bp per intron
- **Gene size:** Largest – 2.4 Mbp (Dystrophin). Mean – 27 kbp.

bp = base pair(s)

What is the mean number of introns pre gene?



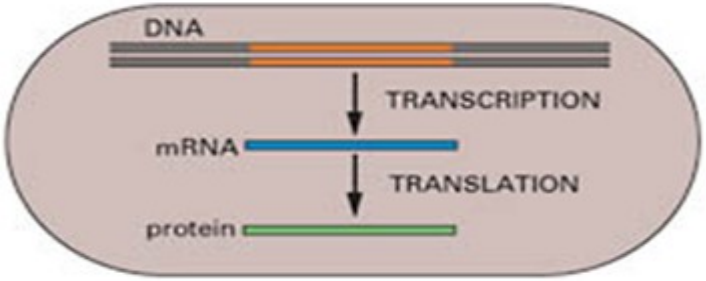
Central Dogma of Biology



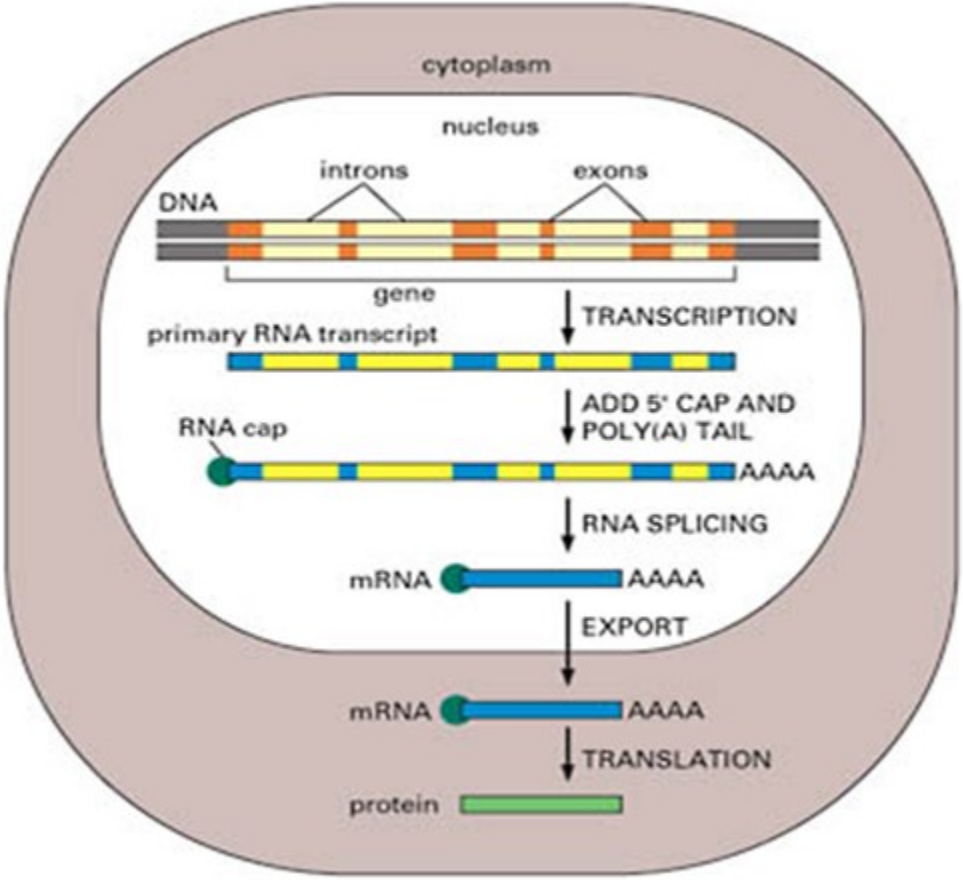
- Base Pairing Rule: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- hnRNA (heterogeneous nuclear RNA): Eukaryotic mRNA primary transcripts whose introns have not yet been excised (pre-mRNA)
- mRNA – this is what is usually being referred to when a bioinformatician says "RNA". This is used to carry a gene's message out of the nucleus

Splicing

PROCARYOTES



EUCARYOTES



Proteins

- Proteins are the workhorses of cells. They act as structural elements, catalyze chemical reactions, regulate cellular activities, and are responsible for cellular structure, producing energy, communication between cells.
- A protein is a linear chain of chemical units called amino acids, of which there are 20 common types.
- The function of a protein is determined by the three-dimensional structure into which it folds.

Amino acids codes:

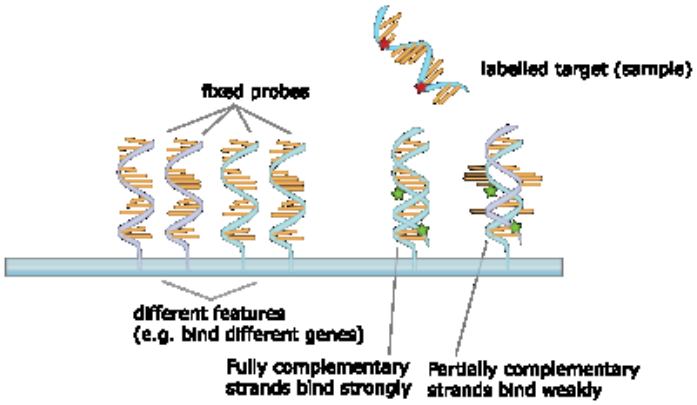
Alanine	**Ala**	**A**	**Glycine**	**Gly**	**G**	**Proline**	**Pro**	**P**
Arginine			Histidine			Serine		
Asparagine	Asn	N	Isoleucine	Ile	I	Threonine	Thr	T
Aspartic acid			Leucine			Tryptophan		
Cysteine	Cys	C	Lysine	Lys	K	Tyrosine	Tyr	Y
Glutamic acid			Methionine			Tyrosine		
Glutamine	Gln	Q	Phenylalanine	Phe	F			

From Genes to Proteins

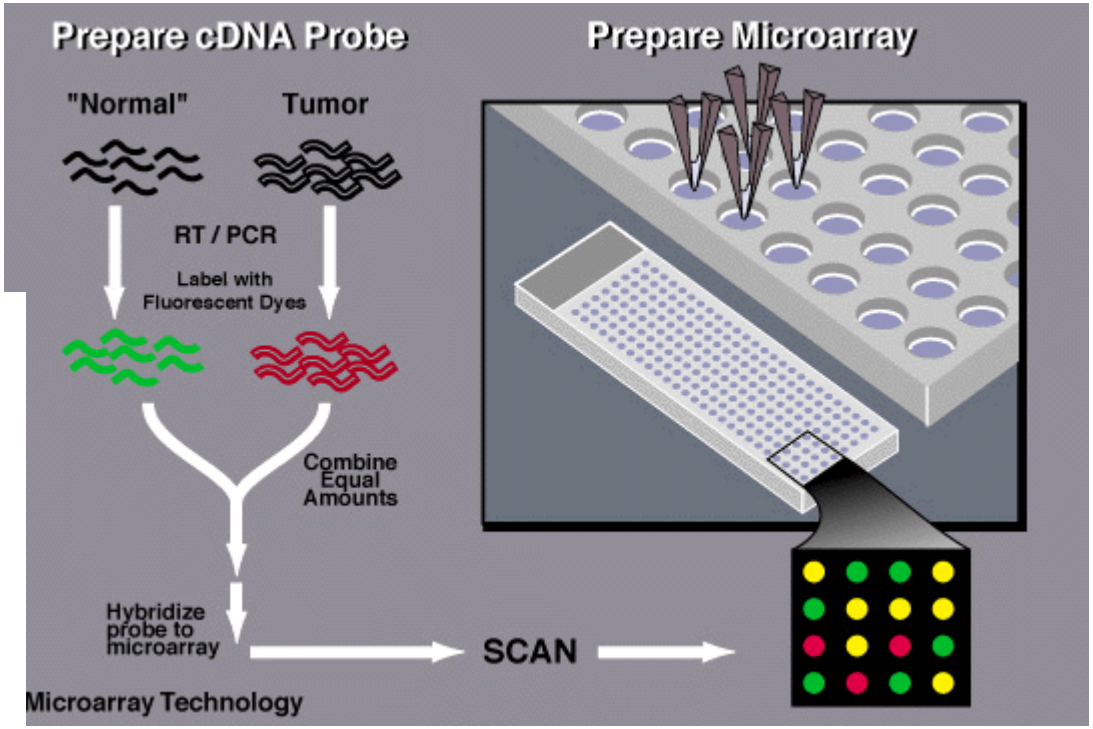
- Translation takes place according to the genetic code, which maps successive triplets (codons) of RNA bases to amino acids.
- With minor exceptions, this many-to-one function from the 64 triplets of bases to the 20 amino acids is the same in all organisms on Earth.
- One of the main problems in science is the protein folding problem of predicting the three-dimensional structure of a protein from its linear sequence of amino acids.
- This problem is far from being solved, although progress has been made by a variety of methods.

Microarray Technology

- all of the cells in a human body contain identical genetic material, but the same genes are not active in every cell
- Tens of thousand of



http://en.wikipedia.org/wiki/DNA_microarray



http://healthinformatics.wikispaces.com/file/view/microarray_technology.gif

REVIEWING SOME BASIC STATS

Reviewing Some Basic Stats

Expected value, sample average

- For a numeric random variable X, the expected value (mean) is

$$\mathrm{E}[X] = \sum_x xP(X = x) \quad \text{or} \quad \int_x xp(x)\,dx$$

- If we take N samples from the same distribution/density, $x_1, \ldots, x_N$, then the sample average $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is an **unbiased estimate** of $\mathrm{E}[X]$.
- That is

$$E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \mathrm{E}[X]$$

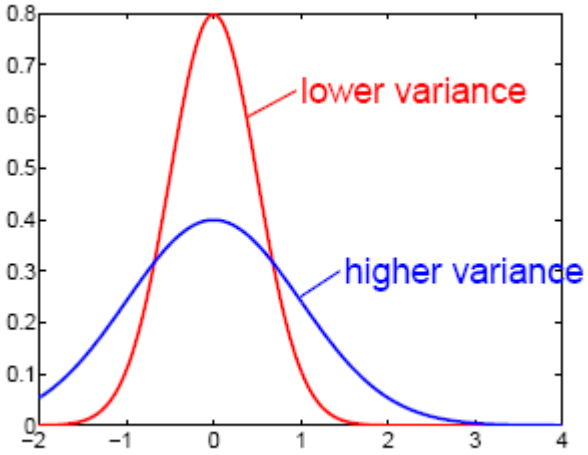
Reviewing Some Basic Stats

Variance

- The variance of X is

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2] =$$
$$= \mathrm{E}[X^2] - 2\mathrm{E}[X]\mathrm{E}[X] + (\mathrm{E}[X])^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

- The variance of X is non-negative and captures how "spread out" X's distribution is.



Reviewing Some Basic Stats

Estimating variance

- The **sample variance** is sometimes

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- It turns out that this underestimates the true variance by a factor of $(N - 1)/N$
- An alternative definition of sample variance – an **unbiased estimator of** $\mathrm{Var}(X)$

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$$

Notation (Mostly)

- $\#object$ – the number of *objects*
- N – the number of *observations*
- d – the number of *variables/attributes*

We assume that $x_{21} = x_{2,1}$

Matrix – Bold, uppercase

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{pmatrix}$$

transposition

Vector – italics, lowercase

$$x_i = (x_{i1}\ x_{i2}\ \cdots x_{id})^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}$$

Vectors are by default columns

- 1 observation = 1 row

Notation (Mostly)

- The value of a variable from all observations

Bold, lowercase

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{pmatrix} \qquad \mathbf{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots \mathbf{x}_p) = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{pmatrix}$$

- If we have a target variable (observation), then y_i is the i-th observation and the observed data consists of $\{(x_1, y_1), \ldots, (x_N, y_N)\}$

Notation (Mostly)

- A vector of length N (the number of observations) will be denoted as

bold, lowercase

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}$$

A scalar	Lowercase, normal	$a \in \mathbb{R}$
A vector of length $= N$	Lowercase, bold	$\mathbf{a} \in \mathbb{R}^N$
A vector of length $k \neq N$	Lowercase, normal	$a \in \mathbb{R}^k$
A matrix	Uppercase, bold	$\mathbf{X} \in \mathbb{R}^{r \times s}$
A random variable	Uppercase, italics	X