

Biological systems and data: A machine learning perspective

A brief introduction

The emergence of complex systems

- Evolution (a theory for the development of biological systems)
 - Evolutionary *computation* (algorithms inspired by theory of evolution and realised in computational terms)
 - Evolutionary computation simulates the evolution of computational models using *life*-inspired constraints
- Data (fossils/evidence of biological complexity)
 - Data-driven analysis of genomic data (large-scale analysis for “patterns”, utilising statistical and machine learning algorithms)
 - Enables the development of computational models using static but real observations (snapshots of nature)

Learning objectives

- Know the basic structure of the cell (our biological system)
- Be familiar with basic processes for gene expression and protein synthesis
- Be familiar with biological sequence data and some data resources: Genbank, Swissprot and PDB
- Be able to exemplify and explain the application of machine learning to problems like gene expression analysis and determining protein localization
- Be familiar with K-means, Hierarchical cluster analysis, Support vector machines and Bayesian inference when applied to the aforementioned problems

Biological systems?

Cells

The human body has roughly about 1,000,000,000,000 cells (1 trillion). There are around 200 types which are basically the same as those found in snakes, birds and other mammals.

Above is a eukaryotic cell (from the Nobel prize web site). Prokaryotic cells have no nucleus, hence the genetic material is stored and sometimes handled differently. All cells (with a few exceptions) contain the same genetic information.

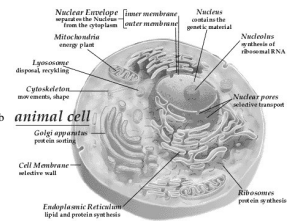


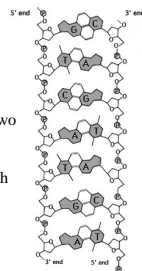
Illustration: Urban Frank

DNA

Most genomes (e.g. cellular) are made of **DNA**.

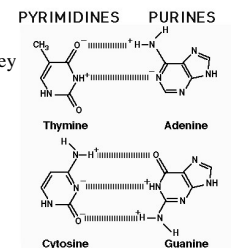
The DNA contains a **blueprint of the organism**.

The DNA is a large molecule made up by two complementary strands of so-called **nucleotides** (there are four bases, often referred to as A, C, G and T, A pairs up with T, C pairs up with G).



Nucleotides

The nucleotides pair up as they do because of the chemical characteristics



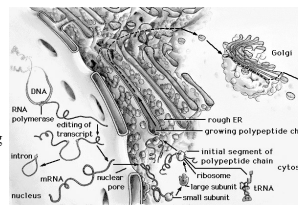
Genes

The genome of a human being (**Homo Sapiens**) contains roughly **3,310,000,000** pairs of nucleotides.
 The house mouse (**Mus Musculus**) has a genome of about **3,300,000,000** base pairs. So, roughly the same size (actually, all mammals are of the same order).
 One popular species in bio-labs is the **Caenorhabditis Elegans** -- a small worm (a nematode) -- has a genome size of about **100,000,000**. So, considerably smaller.

The genome contains **genes** -- stretches of nucleotides that **code** for something. It has been known since Mendel that genes are the basic units of **inheritance** and that genes determine many things of the organism (e.g. colour of eyes).
 The number of genes in humans has been estimated to be around 30-40,000. The same number holds for the house mouse too.
 The number of genes in C. Elegans is estimated to be 19,000. So, we (humans) have not many more "basic units of inheritance" than a worm...

Protein synthesis

The **central dogma** states that the two strands of the DNA are separated at the site of the gene. An enzyme (RNA polymerase) **copies/transcribes the DNA into messenger RNA** (RNA is made of nucleotides as well but Uracil replaces Thymine). The messenger RNA is transported to the ribosome outside the nucleus of the cell. Using transfer RNA the ribosome **translates the mRNA into a sequence of amino acids which forms a protein**. Proteins make us.

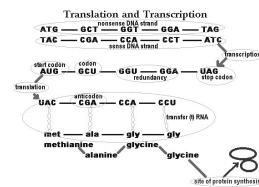


Biological Data?

Amino acids

G	Glycine	Gly	P	Proline	Pro
A	Alanine	Ala	V	Valine	Val
L	Leucine	Leu	I	Isoleucine	Ile
M	Methionine	Met	C	Cysteine	Cys
F	Phenylalanine	Phe	Y	Tyrosine	Tyr
W	Tryptophan	Trp	H	Histidine	His
K	Lysine	Lys	R	Arginine	Arg
Q	Glutamine	Gln	N	Asparagine	Asn
E	Glutamic Acid	Glu	D	Aspartic Acid	Asp
S	Serine	Ser	T	Threonine	Thr

The translation from RNA to amino acids is done by looking at three nucleotides at a time, and converting that triplet to the appropriate amino acid (see table below). There is some redundancy in the *genetic code/codon* (see the third base in particular).



Biologically (in the eukaryotic cell), the transcription to mRNA happens inside the nucleus, the translation happens on ribosomes in the cytoplasm. Also see http://www.accessexcellence.org/RC/VL/GG/protein_synthesis.html for an alternative description of transcription/translation.

TABLE 9.1 The genetic code (messenger RNA)

First base in the codon	Second base in the codon	A	G	Third base in the code
U	Phenylalanine	Serine	Tyrosine	Cysteine
	Leucine	Serine	Formylmethionine	Arginine
	Leucine	Serine	Formylmethionine	Tryptophan
C	Leucine	Proline	Histidine	Arginine
	Leucine	Proline	Histidine	Arginine
	Leucine	Proline	Glutamine	Arginine
A	Isoleucine	Threonine	Asparagine	Serine
	Isoleucine	Threonine	Asparagine	Serine
	Isoleucine	Threonine	Lysine	Arginine
G	Methionine	Threonine	Lysine	Arginine
	Valine	Alanine	Aspartic acid	Glycine
	Valine	Alanine	Aspartic acid	Glycine

