SGN-6156, Lecture 1
Biological sequence analysis

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

(part of the material by Juha Kesseli)

Department of Signal Processing,

Tampere University of Technology

01.04.2008

General motivation

- Computational biology/bioinformatics is almost always somehow connected to biological sequences.
- Three main types of biological sequences: **DNA**, RNA and **protein**.

Basic concepts: a simplified view

- Basic building blocks:
  - genome/DNA
  - genes, proteins
  - *cis*-regulatory elements
- Basic mechanisms:
  - transcription
  - splicing
  - translation (steps 1–3 = gene expression)
  - post-translational modifications/protein folding...
- Transcriptional regulatory mechanisms and other regulatory mechanisms (alternative splicing, microRNAs, protein modifications,...)
- See additional notes from (Ji and Wong, 2006).

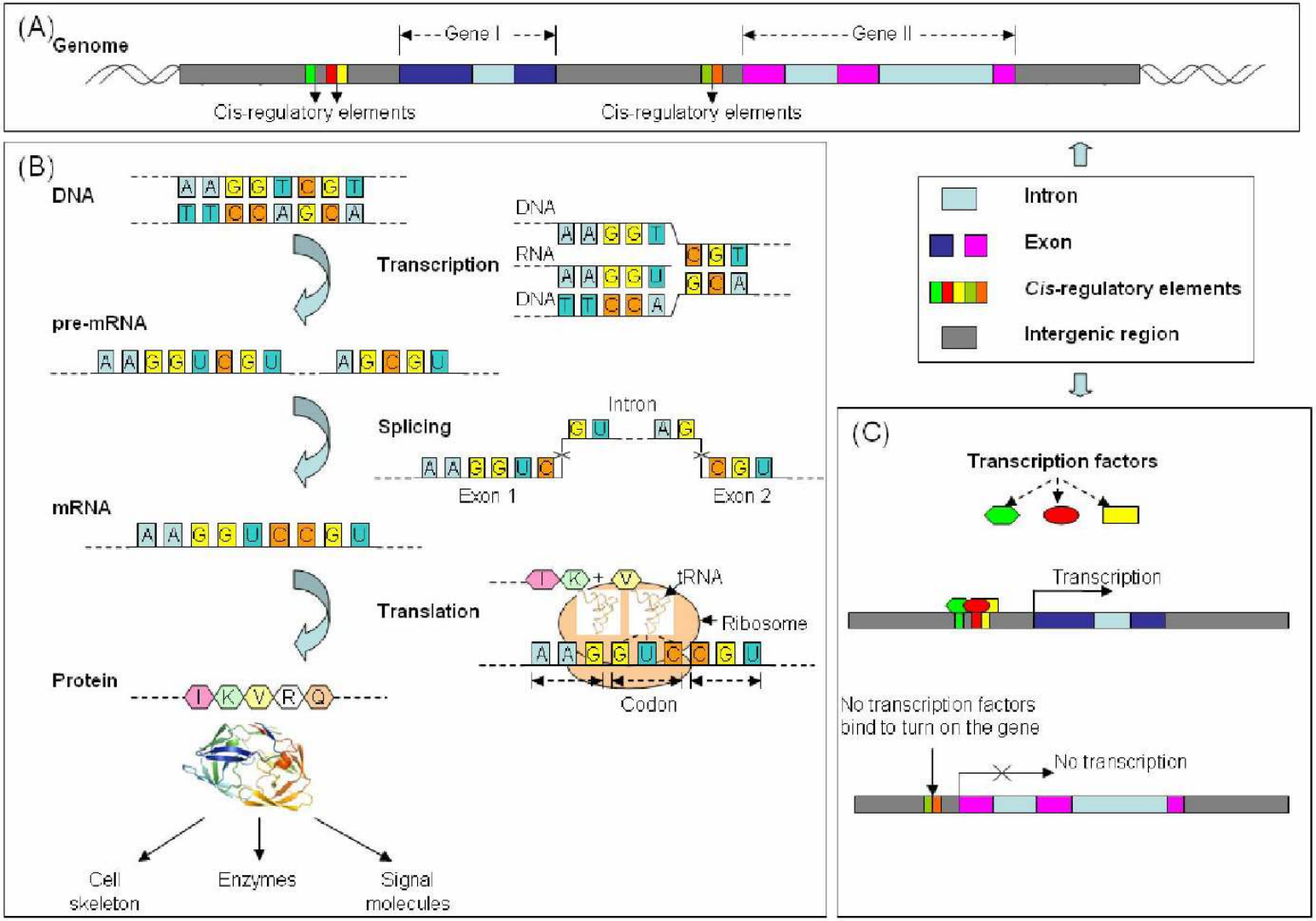

Figure from (Ji and Wong, 2006)

Computational analysis of biological sequences

- Here we emphasize the computational methods (and their underlying principles) that are used to analyze biological sequences.
- Little emphasis on practical sequence analysis or specific programs etc.
- Wet-lab experimentation is the most reliable way of determining a property or a feature of a biological molecule.
- Computational predictions (e.g. from sequence alone) are much easier and less expensive to perform and are thus of great importance.
- Sometimes direct experimentation might also be impossible and indirect computational analysis (statistical inference) is the only way to make biological conclusions.

- Before being able to start computational sequence analysis, one needs at least the sequence(s) to analyze.
  - Sequencing.
- Before been able to use the sequenced genome, one needs to know, at least approximately, the basic components: genes (protein-encoding regions), *cis*-regulatory regions, etc.
- Gene finding:
  - Extrinsic, utilizing sequence alignment.
  - *Ab initio* methods, utilizing statistical models of sequences.
- Several immediate questions are related to biological sequence similarity, homology and alignment.
- Most problems in computational biology are statistical in nature.

Some probabilistic models/concepts, recap

- An example of a biological sequence model: in the most simple setting, biological sequences are strings from an alphabet of size K (4 nucleotides or 20 amino acids).
- Consider a multinomial distribution $\theta = (\theta_1, \ldots, \theta_K)$
  - K outcomes, $\sum_{i=1}^{K} \theta_i = 1$.
- Assume that residues in sequences occur independently.
- The probability of a sample sequence $x = (x_1, \ldots, x_N)$ is

$$P(x|\theta) = \prod_{i=1}^{N} P(x_i|\theta) = \prod_{i=1}^{N} \theta_{x_i}$$

- Maximum-likelihood (ML) estimate: given a model with parameters θ and a set of data D, the maximum-likelihood estimate of θ is the value that maximizes $P(D|\theta)$, i.e.,

$$\hat{\theta} = \arg\max_{\theta} P(D|\theta).$$

- Consider again the above simple model and a sequence x.
- Observations can expressed as counts $n = (n_1, \ldots, n_K)$, and $N = \sum_i n_i$.
- ML parameter estimates are $\hat{\theta}_i = n_i/N$

- Likelihood of the data can be written as

$$P(x|\theta) = \prod_{i=1}^{N} P(x_i|\theta) = \prod_{i=1}^{N} \theta_{x_i} = \prod_{i=1}^{K} \theta_i^{n_i} = P(n|\theta).$$

- ML parameters $\hat{\theta}$ must satisfy $P(x|\hat{\theta}) > P(x|\theta)$ or $\log \frac{P(x|\hat{\theta})}{P(x|\theta)} > 0$ for any $\theta \neq \hat{\theta}$

$$\begin{aligned}
\log \frac{P(x|\hat{\theta})}{P(x|\theta)} &= \log \frac{P(n|\hat{\theta})}{P(n|\theta)} = \log \frac{\prod_{i=1}^{K} \hat{\theta}_i^{n_i}}{\prod_{i=1}^{K} \theta_i^{n_i}} = \log \prod_{i=1}^{K} \left(\frac{\hat{\theta}_i}{\theta_i} \right)^{n_i} \\
&= \sum_{i=1}^{K} n_i \log \frac{\hat{\theta}_i}{\theta_i} = N \sum_{i=1}^{K} \hat{\theta}_i \log \frac{\hat{\theta}_i}{\theta_i} \\
&= N \cdot H(\hat{\theta}||\theta) > 0,
\end{aligned}$$

where $H(\cdot||\cdot)$ is the relative entropy (Kullback-Leibler distance).

- The conditional probability of an event X given Y is (assuming $P(Y) \neq 0$)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}.$$

- The marginal probability of X

$$P(X) = \sum_Y P(X,Y) = \sum_Y P(X|Y)P(Y).$$

- The Bayes' theorem: the posterior probability of X given Y

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}.$$

- Bayesian model comparison among a set of models $\mathcal{M} = \{M_1, M_2\}$, given data D and priors $P(M_i)$

$$P(M_1|D) = \frac{P(D|M_1)P(M_1)}{P(D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_1)P(M_1) + P(D|M_2)P(M_2)}.$$

- Bayesian parameter estimation, given data D and prior $P(\theta)$

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

  where $P(D) = \int_{\theta'} P(D|\theta')P(\theta')d\theta'$.

- The prior $P(\theta)$ can be either informative or uninformative.
- $P(\theta|D)$ defines the full posterior distribution that can be used for/to compute:
  - full Bayesian analysis
  - maximum a posteriori (MAP) estimate
  - posterior mean.
- Both frequentist and Bayesian approaches will be used in the following, although Bayesian methods are preferred (e.g. in small sample settings and in model selection).

Motivation for sequence alignment

- Evolution and natural selection adapts new sequences from the existing ones.
- Sequences evolve by accumulating substitutions, insertions and deletions.
- A basic sequence analysis task is to ask if sequences are related/conserved.
- To answer that, first align the sequences and then determine if that alignment is statistically significant.
- Some potential issues:
  - What kind of alignments are considered as good?
  - How to score and rank different alignments?
  - How to find (computationally) good alignments?
  - How to evaluate significance?

- Known sequences in databases can be used to find close matches in arbitrary DNA or protein sequences.
- Match similar sequences in order to find, e.g.
  - homologs (sequences with shared ancestry and, thereby, possibly a shared function)
  - binding sites of similar molecules (can result from convergent evolution, typically transcription factors)
  - ...
- Finding homologous genes is the most common way of generating new annotations for genes (although homologous genes need not have the same or similar function).
- Aligning multiple sequences can also be used to study the phylogenetic tree.

Protein vs. DNA alignment

- Typically, it is recommended that proteins are aligned instead of DNA if possible.
  - With DNA, we need to consider the different reading frames.
  - It is simpler to incorporate probabilities of mutation for different amino acids into the alignment scores.
  - In particular with more distant sequences the comparison of nucleotides discards usable information.

Pairwise alignment

- From now on, the presentation mainly follows (Durbin et al, 1998; Section 2).
- In pairwise alignment we have two sequences that we want to compare.
- The alignment can be global or local.
  - In global alignment the two sequences are aligned from beginning to the end.
  - In local alignment subsequences with high similarity are found. This is often more interesting and convenient in practice since shorter similar subsequences often correspond with functionally similar domains.

- Pairwise alignment is also used by selecting a query sequence that is then pairwise compared with all the sequences in a database (e.g. BLASTing).
- An alignment example, Figure 2.1 in (Durbin et al. 1998)

An alignment scoring model

- In order to define how closely two sequences match, i.e. how well they can be aligned, we need to have a metric to determine their distance.
- To measure distances between two sequences with a common ancestor we need to know e.g. the probabilities of different point mutations occurring in one or both of the homologous sequences.
- We try to find evidence that sequences have developed (evolutionarily) from a common ancestor by a process of mutation and selection
  - Substitutions
  - Deletions/Insertions
- Evolutionary selection might have favored some type of mutations.
- The overall score is a combination of individual/point scores: identities, substitutions and gaps (deletions and insertions).

- A pair of sequences, $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$, assume $m = n$ first
- x_i and y_j take values from an alphabet $\mathcal{A}$ as above: $\mathcal{A} = \{A, C, G, T\}$ or the twenty amino acids.
- A random model R: symbols in x and y occur independently with probabilities q_a, q_c, q_g and q_t

$$P(x, y|R) = \prod_{i=1}^{n} q_{x_i} \prod_{j=1}^{n} q_{y_j}.$$

- A match model M: aligned pairs occur with a joint probability p_{aa}, p_{ac}, etc.

$$P(x, y|M) = \prod_{i=1}^{n} p_{x_i y_i}.$$

- $p_{x_i y_i}$ can be interpreted as the probability that both residues x_i and y_i have been independently derived from a common ancestor residue.
- Relative alignment score from the likelihood ratio (odds ratio)

$$\frac{P(x, y|M)}{P(x, y|R)} = \prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}.$$

- Logarithm of the likelihood ratio gives an additive score

$$S = \sum_{i=1}^{n} s(x_i, y_i) = \sum_{i=1}^{n} \log \left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right).$$

- Elements $s(a, b)$ form a substitution matrix.

Substitution matrices

- Substitution matrix contains estimates of the rates of DNA mutation for different amino acids or nucleotides.
- In a common 20-by-20 matrix the (i, j)th entry contains the probability that the ith amino acid mutates into the jth amino acid over a selected unit of time.
- Substitution matrices for nucleotides contain only little information.
- Common substitution matrices for protein sequences
  - BLOSUM
  - PAM
- Let us assume for now that a substitution matrix s is given (these can be estimated from data, as we'll see later).

BLOSUM62 substitution matrix

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	2	0
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-3
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-2	1
L	-1	-4	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-3
M	-1	-3	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-2
N	-2	1	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

Gap penalties

- The above scoring model does not yet take into account gaps (insertions/deletions).
- Gaps need to be penalized.
- Common gap penalty scores for a gap of length g are the linear score

  $$\gamma(g) = -dg$$

  or an affine score

  $$\gamma(g) = -d - e(g-1),$$

  where d is the gap open and e is the gap extension penalty.
- Typically $d > e$.

- The probability of a gap at a given location is the product of $f(g)$ (a function/density of the gap width) and the probability of inserted residues

$$P(\text{gap}) = f(g) \prod_{\text{residues in gap}} q_{x_i}$$

- Residues in the gap do not correlate with the length of the gap.
- Probabilities q_{x_i} above come from the random model.
- Log-likelihood ratio of the gap model to the probability of the random model gives $\gamma(g) = \log(f(g))$.

References

- R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.
- H. Ji and W. H. Wong (2006). Computational biology: toward deciphering gene regulatory information in mammalian genomes, *Biometrics*, vol. 62, pp. 645–663.