Aalto University
School of Science

CS-E5865 Computational genomics

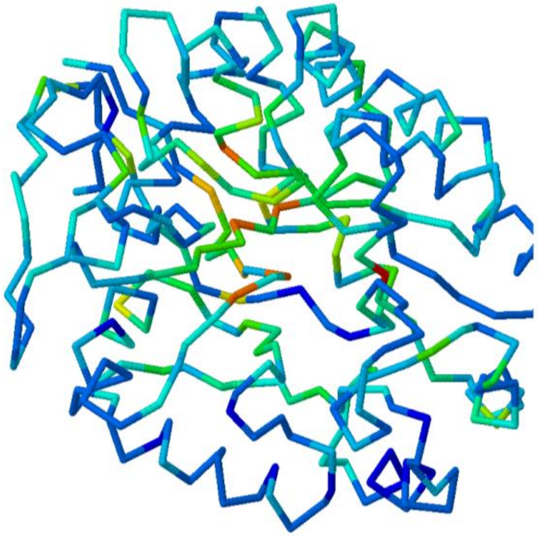
Autumn 2020, Lecture 2: Gene finding

Lecturer: Pekka Marttinen

Assistants: Alejandro Ponce de León, Zeinab Yousefi, Onur Poyraz

Introduction to genes and proteins

- What are proteins?
- Central workhorses of the cell, performing a wide variety of functions:
  - catalyzing metabolic reactions, replicating DNA, responding to stimuli, transporting molecules, etc.
- They consist of a chain of amino-acids that folds itself into a 3-dimentional shape which ultimately determines its function
  - Errors in the amino-acid sequence can lead to malfunctioning proteins
- There are 20 amino-acids that can form a huge number of proteins

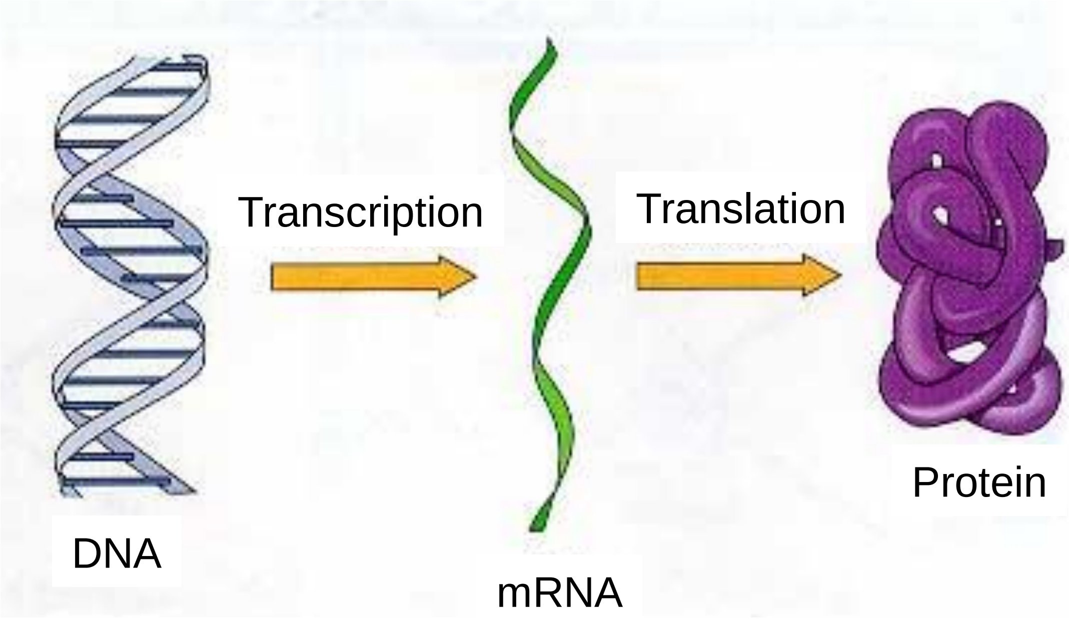


Marttinen et al., 2006, Bioinformatics

Aalto University
School of Science

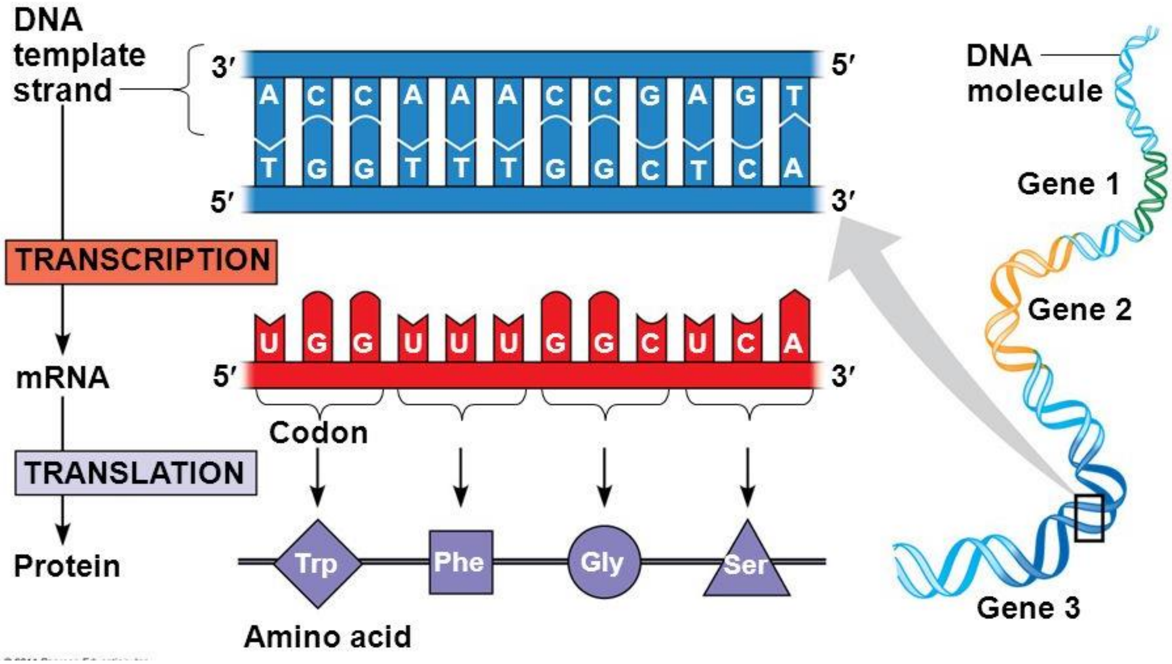
From genes to proteins

- Central dogma of molecular biology:



https://biochemist01.wordpress.com/tag/what-is-central-dogma/

From genes to proteins



http://slideplayer.com/slide/7225692/

Reading frames

- Not all regions of an mRNA molecule are translated.

- The translational machinery must know on which nucleotide to start the translation
  - Depending on the start position there are 3 different ways to decompose a sequence into codons.
  - Example: Consider the sequence ACTCGGGCTGGACACAC

ACT CGG GCT GGA CAC AC

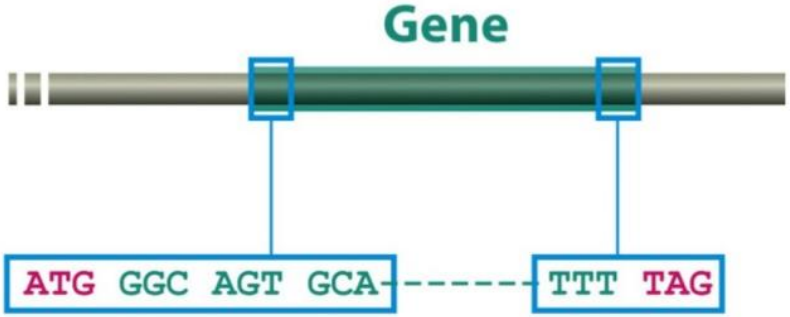
A CTC GGG CTG GAC ACA C

AC TCG GGC TGG ACA CAC

- Reading frame: each of the three ways to decompose the DNA sequence into codons

Open reading frame (ORF)

- Translation starts at codon ATG (methionine)
- 3 stop codons signal the end of the translation: TGA, TAA and TAG
- Open reading frame: a stretch of DNA whose length is a multiple of 3, that begins with the start codon and ends with one of the 3 stop codons
  - internal start codons are accepted



Exercise

- Below is a DNA fragment from the beginning of a gene. Determine which strand is transcribed, indicate the polarity of the two DNA strands, and the sequence of bases in the resultant mRNA.

ACATACGCCTTTCAGGTT
TGTATGCGGAAAGTCCAA

- Slightly modified from https://www.youtube.com/watch?v=gAm1ASjAMf8

Frame-shift mutations

- Mutations in DNA changing a nucleotide to another will typically only change one amino acid to another
  - May not affect the function of the protein
- Mutations that insert or delete a nucleotide are called *frameshift mutations*
- Frameshift mutations usually have drastic consequences
  - The rest of the amino acid sequence is changed
  - The resulting protein might not be functional

Gene finding (aka Gene predition)

- Task: given a genomic sequence, find the Open Reading Frames (ORF’s)
  - delineated by start (ATG) and stop codons (TAA,TAG,TGA).
- What’s the difficulty here?
  - Cannot we just mark down all start and stop codons that we can find in the genome and declare a stretch between a start and stop codon as an ORF?
- Two challenges:
  1) Triplets of nucleotides looking like start and stop codons may appear by chance
  2) In eukayrotic genes, one should also find the introns and exons inside the coding region.

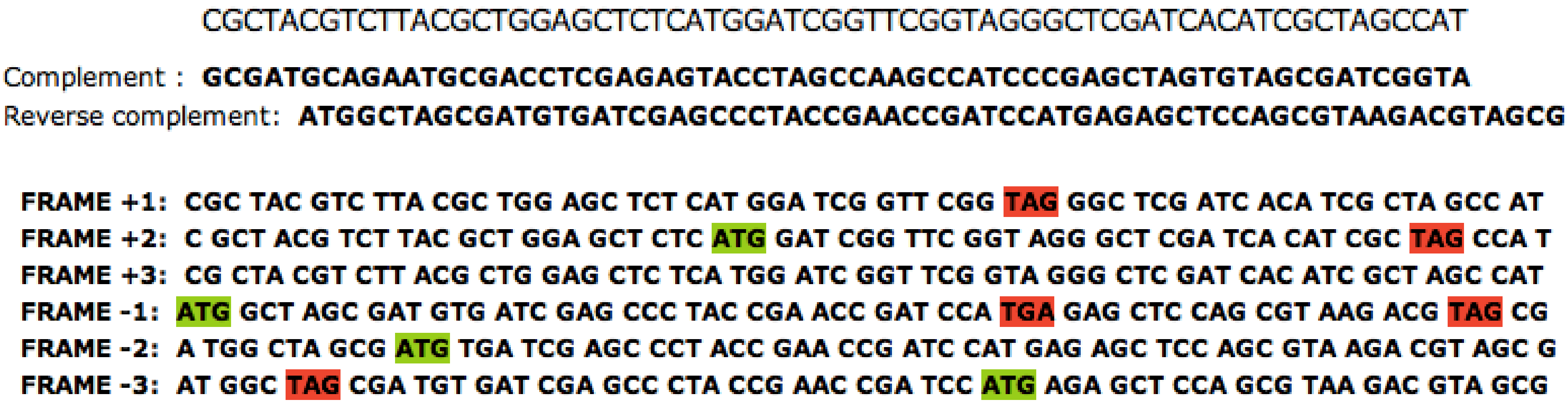
Spurious start and stop codons

- If the correct reading frame (=codon boundaries) is not known, there may be several candidates for start and stop codons within the sequence

```
TCTCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGAACA
TCGCAGCACATCTTTTACGCACCTCTCCATCTCTGCTCACACGCACCACCACAACCA
CAAGATTTCTGCCCTTCTCTACGTCTTCGTTCTTAAACCATGGCCATTTGAAAAAAC
CGAAACCAGGCGAAGAACTGAAGATAACTTTTATTCTGAAGGATGGCTCCCAGAAGA
CGTACGAAGTCTGTGAGGGCGAAACCATCCTGGACATCGCTCAAGGTCACAACCTGG
ACATGGAGGGCGCATGCGGCGGTTCTTGTGCCTGCTCCACCTGTCACGTCATCGTTG
ATCCAGACTACTACGATGCCCTGCCGGAACCTGAAGATGATGAAAACGATATGCTCG
ATCTTGCTTACGGGCTAACAGAGACAAGCAGGCTTGGGTGCCAGATTAAGATGTCAA
AAGATATCGATGGGATTAGAGTCGCTCTGCCCCAGATGACAAGAAACGTTAATAACA
ACGATTTTAGTTAATGCCCTGC
```

Finding genes on the complementary strand

- In DNA genes lie on both strands
- To find genes on a single strand of DNA, we also need to consider the reverse complement
- We have in total six reading frames to consider
  - Three in one direction
  - Three in the reverse direction, with reverse complement start and stop codons



```
                   CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGGTAGGGCTCGATCACATCGCTAGCCAT

Complement :       GCGATGCAGAATGCGACCTCGAGAGTACCTAGCCAAGCCATCCCGAGCTAGTGTAGCGATCGGTA
Reverse complement:  ATGGCTAGCGATGTGATCGAGCCCTACCGAACCGATCCATGAGAGCTCCAGCGTAAGACGTAGCG

FRAME +1:  CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT
FRAME +2:  C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T
FRAME +3:  CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT
FRAME -1:  ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2:  A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3:  AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG
```

Gene prediction: main approaches

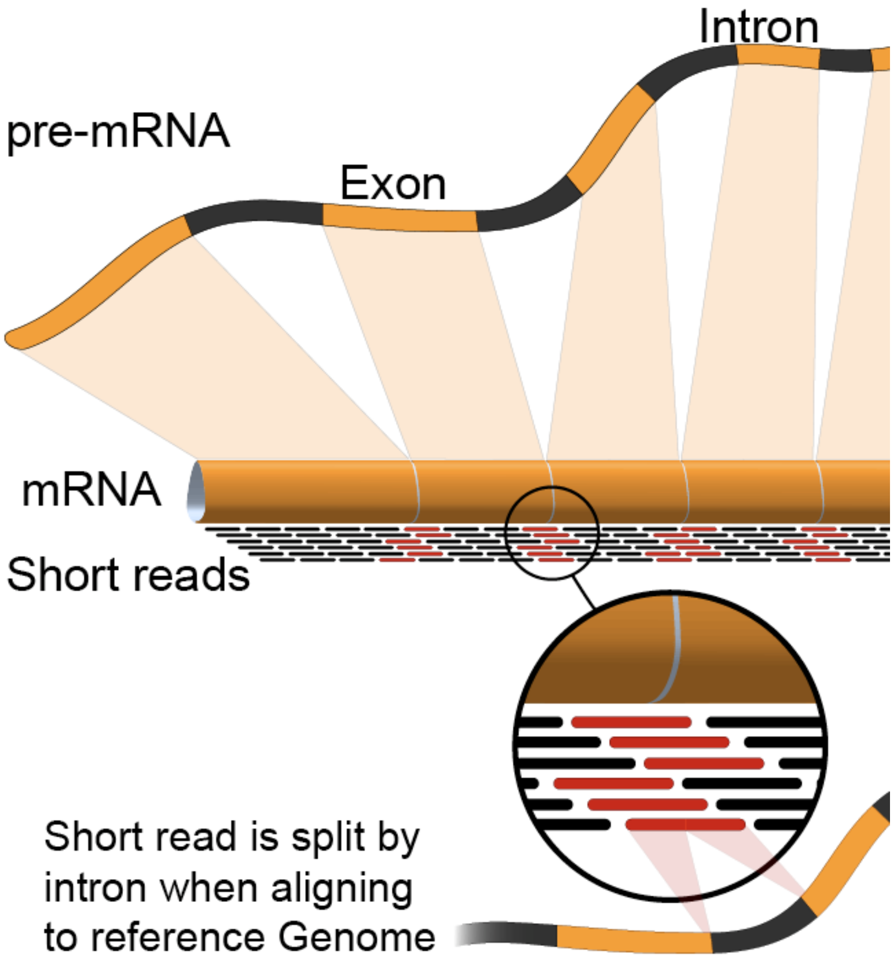
- Evidence-based gene finding: identify genes by inspecting the products of the genes, mRNA and protein sequences in the cell, and map them back to the genome
  - Note: not discussed in C&H book; the techniques became mainstream after 2007
- Ab initio gene prediction: detecting the 'signal' of functional elements via statistical approaches or matching against a database of known motifs
- Comparative genomics approaches: detect conserved DNA regions by comparing a large set of related genomes

Evidence-based gene finding

- In evidence-based gene finding, one assumes that there is access to mRNA or protein sequences expressed by the organism
  - RNA-seq is one suitable experimental technique for mRNA
  - Peptide sequencing via tandem mass spectrometry gives amino acid sequences
- Target genome is searched for sequences that match the expressed mRNA or protein sequences
  - Sequence alignment problem using, e.g. BLAST for prokaryotic genes, relatively straight-forward
  - Exon-intron structure of eukaryotic genes is a complication

RNA-seq for gene finding

- Two alternative approaches:
  1. Assemble mRNA from short reads and match the mRNA transcript to the genome, taking introns into account (right).
  2. Align the short reads of cDNA directly to the genome and vote for exons.



Eukaryotic Gene finding with known protein sequences

- Consider matching known protein sequence to the target genome
- As only exons are translated, when matching the protein sequence into the target genome, one needs to consider where the introns might be located
- By computational means one can find the best alignment between the protein sequence and the DNA sequence
  - Sequence of predicted exons interleaved by introns
  - Sequence alignment algorithms

Limitations of evidence-based gene finding

- Major limitation of evidence-based approach is coverage
  - mRNA approach:
    - Not all genes are expressed all the time or in all tissues, so mRNA will not in general cover the all genes
  - Known protein approach:
    - Not all proteins have been sequenced, corresponding genes would be missed
    - What if the target genome contains previously unknown genes?
- For larger coverage, we need *ab initio* tools that do not require observing the gene products

Ab initio gene prediction

- What can be deduced just by looking at the genome?
- In this lecture we discuss some basic *ab initio* methods used for prokaryotic gene finding
  - Hidden Markov Model (HMM) –techniques can be used for eukaryotic gene finding. (later in the course)

- In prokaryotic cells all genes are DNA sequences beginning with a start codon and ending with a stop codon
  - Already non-trivial for prokaryotes as not all start codon – stop codon pairs (*open reading frames*, ORFs) correspond to genes.

Detecting spurious signals: hypothesis testing

- When searching a genome for *patterns* (k-mers, ORFs, exons,...) we need to consider the probability of them being created by chance

- Need methods for separating “true findings” (or “signal”) from “false” (or “noise”)

- In statistics, *hypothesis testing* refers to calculating these probabilities and making inferences based on them

Statistical hypothesis testing

- Ingredients:
  - Null model or null hypothesis, denoted H_0 (e.g. ORF is generated by a random process)
  - Alternative hypothesis H_1, generally the logical complement of H_0 (e.g. ORF has been generated by a biologically relevant process)
  - Probability distribution for data under the null hypothesis (e.g. the i.i.d multinomial distribution)
  - Test statistic of interest (e.g. length of the ORF)
  - Significance level: a fixed probability α wrongly rejecting the null hypothesis H_0
  - p-value: the probability of the test statistic obtaining as extreme or more extreme value by chance, if null hypothesis H_0 is true

Statistical hypothesis testing

- We consider the probability of a given pattern (e.g ORF) being created by chance under the null hypothesis
- An occurrence of a pattern (k-mer, ORF, exon) is *significant* if it has a smaller p-value than the given significance level α, i.e. it is highly unlikely to appear under the null model
- Note that by the means of statistical hypothesis testing, we cannot **guarantee** that the pattern is **not** created by chance.

False positive and false negative findings

- Two types of errors in hypothesis testing
- False positive (FP)
  - We incorrectly reject the null hypothesis, i.e. call the pattern significant
  - Also denoted to Type I error in statistics
- False negative (FN)
  - We incorrectly accept the null hypothesis, i.e. call the pattern not significant

	Sequence is not a gene	Sequence is a gene
Test significant	FP	TP
Test non-significant	TN	FN

Significance levels

- The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis H_0.
- Commonly used levels for statistical significance:
  - 5% is generally considered as “almost significant”,
  - 1% significant and
  - 0.1% very significant.
- However, the levels are *conventions*, not arising from theory
- In bioinformatics, it is also possible to use the p-values to rank the discovered patterns, without using the arbitrary significance level cut-off
  - List of highly ranked patterns can then be presented to a human expert for further analysis

Multinomial sequence model

- The simplest model for DNA sequences
- Assumes that nucleotides appear independently from each other and with a fixed probability, according to a given distribution (i.i.d assumption)

$$p = (p_A, p_C, p_G, p_T)$$

- The probability of observing a nucleotide is independent of the position $p_x = p(\mathbf{s}(i) = x)$
- Probability of a sequence s obtained by multiplying the observed nucleotide probabilities

$$P(s) = \prod_{i=1}^{n} p(\mathbf{s}(i)) = \prod_{x \in \mathcal{N}} p_x^{n(x,s)}$$

Example: Computing the probability of an ORF

- For an already identified ORF in a sequence, what is the probability of finding an ORF of equal length (or longer) in a random sequence?
- What is the probability of an ORF of k or more codons arising by chance?
- First approximations:
  - assume an i.i.d multinomial model
  - assume all 64 codons are equally likely
  - need to consider a sequence of k codons that do not contain a stop codon

$$P(\text{'run of } k \text{ non-stop codons'}) = \left(\frac{61}{64}\right)^k$$

Significance of an ORF

- What is the sequence length k such that 95% of randomly created ORFs are shorter than k?

$$P(\text{'at least } k \text{ non-stop codons'}) = (61/64)^k$$

- Try different k to discover

$$P(\text{'at least 63 non-stop codons'}) = (61/64)^{63} = 0.049$$

- By accepting only ORFs of length 65 (63 non-stop codons + start & stop codons) or more, 95% of the spurious ORF’s are removed.
- For significance level of 99% (α=0.01), the threshold would be 98 codons
- More details in *note_on_orf_significance.pdf*.

Computing the probability of an ORF

- To get a more refined model, we can drop the assumption of equal codon frequencies
- Consider the probabilities of observing a stop codon

$$P(stop) = P(TAA) + P(TAG) + P(TGA)$$

- Now the probability of an ORF of k non-stop codons under an i.i.d model is given by

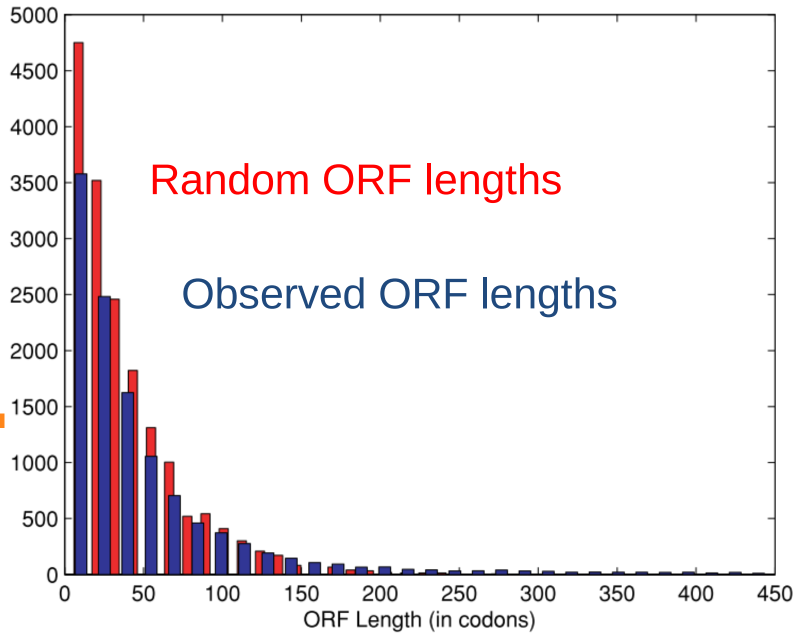
$$P(\text{'run of } k \text{ non-stop codons'}) = (1 - P(stop))^k$$

Randomization tests

- Sometimes it may be difficult to exactly compute p-values for observations.
- For example, it may not be clear what kind of null model to use, or the null model leads to very complicated equations
- In these cases we can use randomization tests
- In randomization testing, one creates a large set of data that is consistent with the chosen null model, but otherwise resembles the observed data

Randomization tests

1. Simulate random data that are consistent with the null model.
2. Check the distribution of the test statistic (e.g. ORF length) in the simulated data.
3. Check the rank of our observed pattern in this distribution (lengths of randomly created ORFs).
4. p-value is the fraction of simulated data that have test statistic values greater than or equal to the test statistic for the observed pattern.



Randomization tests

- Several ways to obtain randomized sequences
- In permutation testing, one shuffles the original sequence randomly. Several choices, capturing different aspects
  - Shuffle nucleotides independently
  - Shuffle the codons
    - If the ORFs have been already predicted, this is straightforward
    - Otherwise, needs a method to pick the reading frame (codon boundaries).
- In Bootstrapping, one samples with replacement from the original sequence
    - Again, can be done for individual nucleotides or longer stretches of DNA

Multiple testing

- In computational genomics, hypothesis testing is typically conducted for 100s or 1000s of patterns
- p-values determine the significance of a single test
- The number of tests should be taken into account
- This is called the multiple testing correction

Multiple testing

- False positive rate
  - The probability of getting a significant result for a random sequence.
  - Formally: FP / (FP + TN)
  - 5 “significant” results in 100 tests does not mean that the significance tests meant anything biologically (with $\alpha=0.05$).
- False discovery rate
  - The proportion of false findings among all significant tests.
  - Formally: FP / (FP + TP)

	Sequence is not a gene	Sequence is a gene
Test significant	FP	TP
Test non-significant	TN	FN

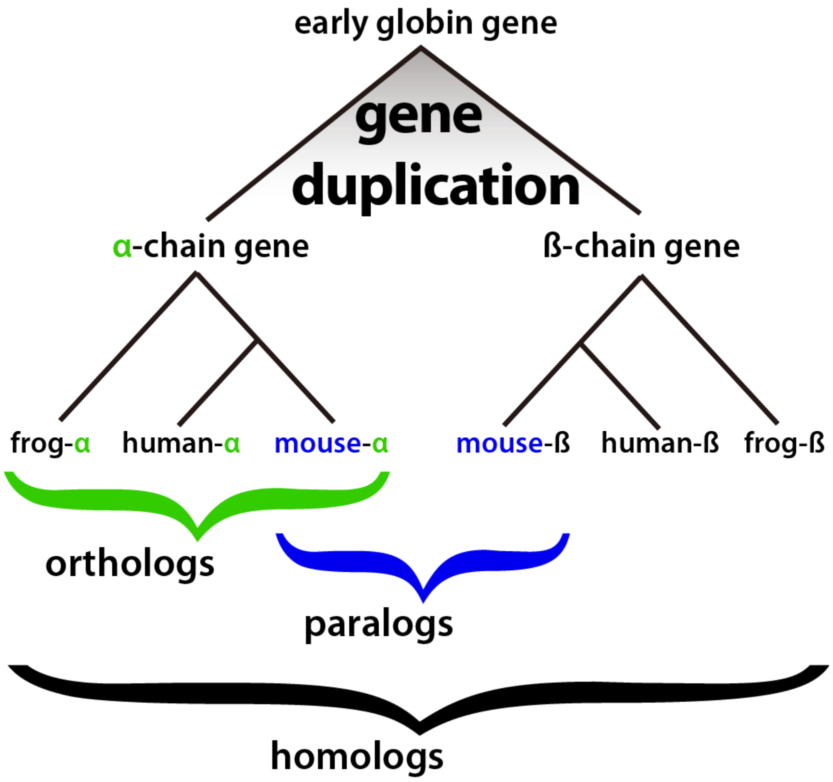
Comparative genomics approach: Sequence homology

- Genomic studies rely heavily on the notion of genes in different organisms having evolutionary relationships
- For example, humans, mice and fruit fly share a large number of genes that are assumed to have a common ancestor gene
  - Such genes are said to be *homologs**
- Groups of homologous genes form *gene families*

*from Greek *homologos*: *homo* = agreeing, equivalent, same + *logos* = relation

Orthology and paralogy

- Homologous genes come in two flavors
- Orthologous genes are copies of the descendants of the ancestral gene in different organisms
- Paralogous genes are copies of the ancestral genes within the same organism
  - arise via duplication of genes in genomes
  - enable function evolution via divergence of the copies



Sequence homology and similarity

- Homology is tricky to detect directly with computational means (phylogenetic analysis deals with this problem)
- Typically, sequence similarity is used as an alternative concept
  - Idea: if two genes share an ancestor, their nucleotide sequences will probably be similar
- Note: homology is a binary concept (common ancestor/no common ancestor), similarity is a multi-valued concept (e.g.80% similar is possible)

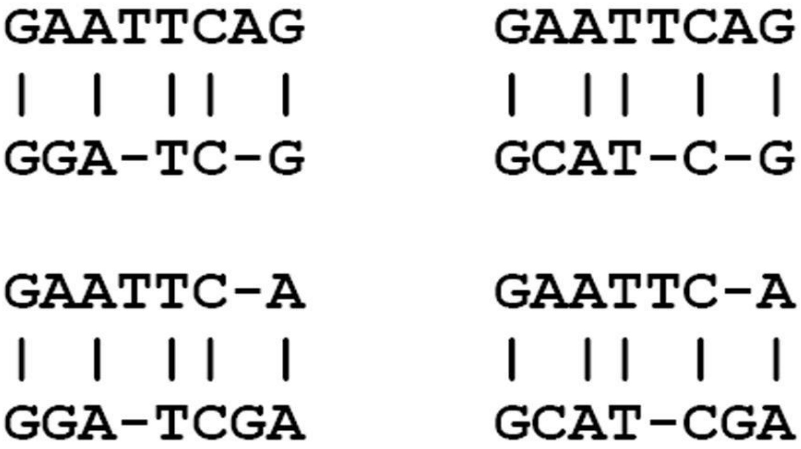
Sequence alignment

- The purposes of sequence alignment are
  - to measure the sequence similarity of two sequences
  - to reveal which parts of the sequences match and which do not
- Commonly used way to visualize pairwise alignments on the right:

  "|" denote matching pair of symbols

  "-" denotes a gap symbol inserted in the sequence to improve alignment

Example: align the 2 sequences
GAATTCAG
GGATCGA



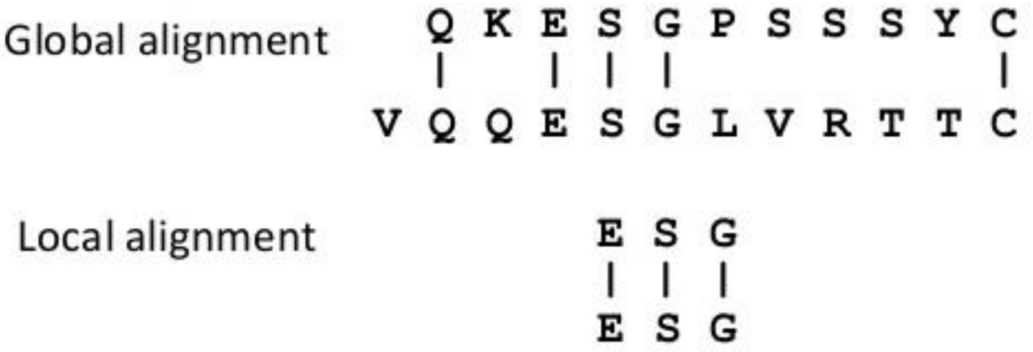
```
GAATTCAG     GAATTCAG
| | || |     | || | |
GGA-TC-G     GCAT-C-G

GAATTC-A     GAATTC-A
| | || |     | || | |
GGA-TCGA     GCAT-CGA
```

Uses of sequence alignment in biology

- Prediction of function: given a similar gene with a known function, one can predict the function for a new gene by transferring the annotation
- Database searching: searching for similar genes (with known or unknown function) in a large databases
- Gene finding:
  - comparison of whole genomes of sets of related organisms can reveal gene locations
  - Evidence-based approaches: aligning the expressed mRNA or protein sequences against the genome
- Sequence assembly: aligning short DNA sequences against a reference genome or each other

Global and local alignment

- Two types of alignment:
- Global alignment aims to maximize the alignment quality over the whole sequences
  - leaving gaps typically penalized
- Local alignment looks to match sub-regions of the sequences
  - gaps typically not penalized



```
Global alignment     Q K E S G P S S S Y C
                     |   | | |         |
                   V Q Q E S G L V R T T C

Local alignment          E S G
                         | | |
                         E S G
```

http://www.slideshare.net/avrilcoghlan/the-smith-waterman-algorithm

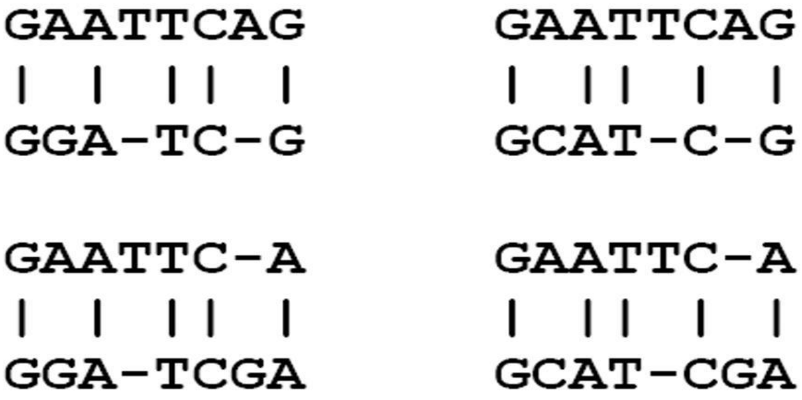
Global alignment scoring functions

- By inserting gaps in different places, we get different alignments
- We wish to find the best one
- We define a simple scoring function σ(x,y) for a pair of symbols in the alignment
- The alignment score is the sum

$$M = \sum_{i=1}^{c} \sigma(x_i, y_i)$$

where i indexes the positions in the alignment

Example: align the 2 sequences
GAATTCAG
GGATCGA



```
GAATTCAG        GAATTCAG
| | || |        | || | |
GGA-TC-G        GCAT-C-G

GAATTC-A        GAATTC-A
| | || |        | || | |
GGA-TCGA        GCAT-CGA
```

Global alignment scoring functions: Example

- Simple scoring function:

$$\sigma(-,a) = \sigma(a,-) = -1$$

$$\sigma(a,b) = \begin{cases} -1 & a \neq b \\ 1 & a = b \end{cases}$$

- Scores of the alignments on the left 1*5 -1*3 = 2

```
GAATTCAG    GAATTCAG
| | || |    | || | |
GGA-TC-G    GCAT-C-G

GAATTC-A    GAATTC-A
| | || |    | || | |
GGA-TCGA    GCAT-CGA
```

Substitution matrices

- We can collect the scores of the function σ into a matrix (right)
- In general, the scores can depend on the pair of symbols
- Matrix S containing the σ values is called the substitution matrix
- For DNA simple scoring schemes are typically used
- For amino acids more rich substitution matrices are used
    - PAM
    - BLOSUM

$$\begin{array}{c|ccc} & a & b & - \\ \hline a & +1 & -1 & -1 \\ b & -1 & +1 & -1 \end{array}$$

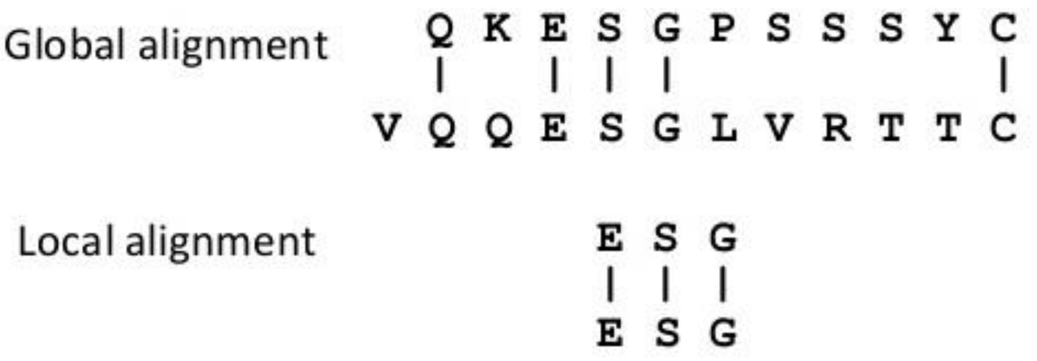
$$S = \begin{array}{c|ccccc} & a_1 & a_2 & \ldots & a_l & - \\ \hline a_1 & \sigma(a_1, a_1) & \sigma(a_1, a_2) & \ldots & \sigma(a_1, a_l) & \sigma(a_1, -) \\ a_2 & \sigma(a_2, a_1) & \sigma(a_2, a_2) & \ldots & \sigma(a_2, a_l) & \sigma(a_2, -) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_l & \sigma(a_l, a_1) & \sigma(a_l, a_2) & \ldots & \sigma(a_l, a_l) & \sigma(a_l, -) \\ - & \sigma(-, a_1) & \sigma(-, a_2) & \ldots & \sigma(-, a_l) & \sigma(-, -) \end{array}$$

Optimal global alignment

- The optimal alignment A* between two sequences s and t is the alignment A(s,t) that maximizes the alignment score M over all possible alignments.
- There are $\binom{2n}{n}$ possible alignments between two sequences of length n, so brute-force enumeration of all of them is not feasible
- Can be solved efficiently with so called Needleman-Wunsch algorithm, which is based on dynamic programming (we take a closer look in the next lecture)
  – Basic idea: solve the problem for prefixes of length 1,2,...,n incrementally making use of the optimal solutions for the prefixes

Local alignment

- Finding two subsequences of sequences s and t, that will have the best alignment score
- Biological motivation: perhaps part of the gene has been conserved, e.g.
  - a functional part (a domain) of a protein, or
  - a binding site of a regulatory protein in the promoter region
- Smith-Waterman algorithm (next lecture)



```
Global alignment      Q K E S G P S S S Y C
                      |   | | |         |
                    V Q Q E S G L V R T T C

Local alignment             E S G
                            | | |
                            E S G
```

http://www.slideshare.net/avrilcoghlan/the-smith-waterman-algorithm