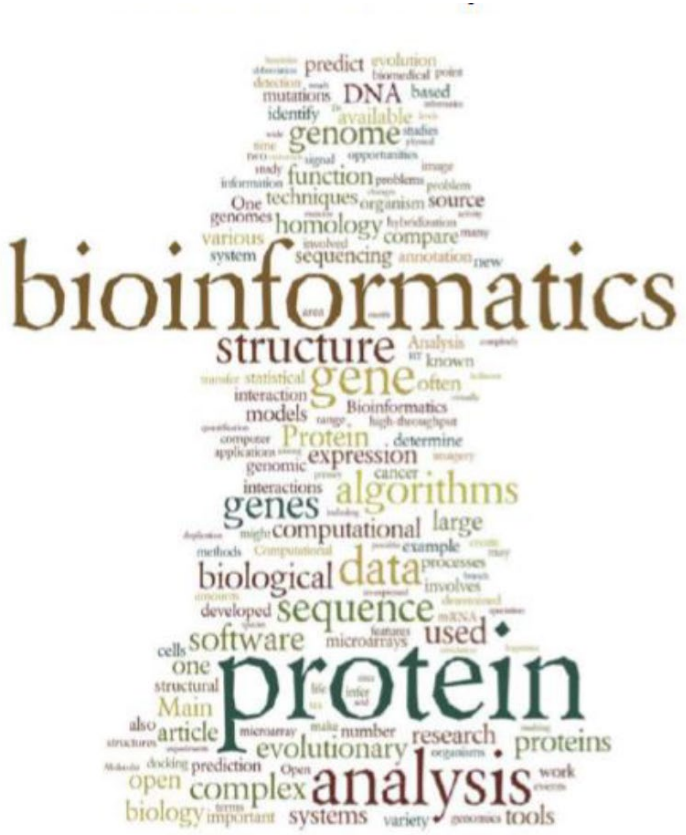
Models in Bioinformatics

Mikael Pontarp
Department of Biology, Faculty of Science, Lund University
mikael.pontarp@biol.lu.se

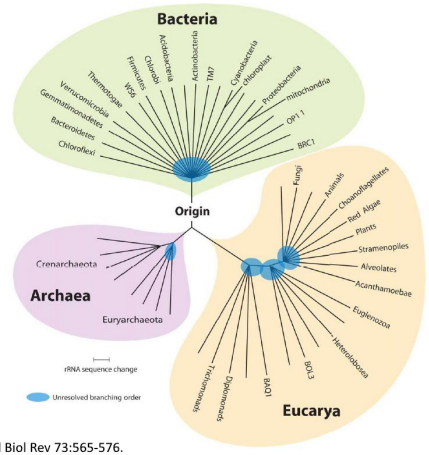


Mannion et al. 2014

LUND UNIVERSITY

On Models in Bioinformatics

Tree of life – three domains

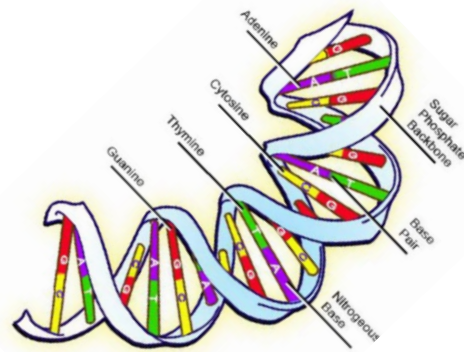


Source: Pace, N. R. (2009) Microbiol Mol Biol Rev 73:565-576.

Reminder:

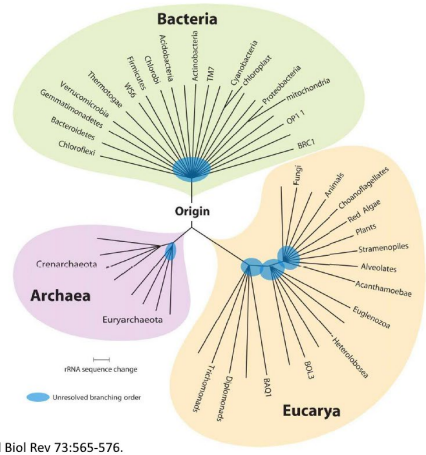
- A model is a:
  - Simplification of reality
  - Abstraction of reality
  - Generalization reality
- "A model is always wrong, but sometimes useful" George E. P. Box
  - Scale and degree of simplification depends on the question (e.g. maps of different resolution)
- Different models for different biological organization
  - Models in bioinformatics is commonly focused on nucleotides, genes and genomes

Biological organization



On Models in Bioinformatics

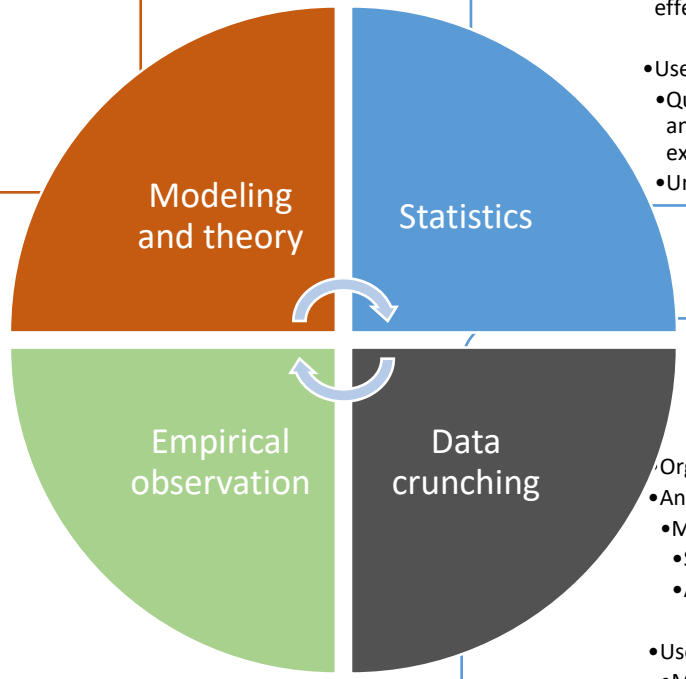
Tree of life – three domains



Source: Pace, N. R. (2009) Microbiol Mol Biol Rev 73:565-576.

- Phenomenological models
- Mechanistic models
- Process based model

- Used for:
  - Conceptualization
  - Understanding
  - Prediction
  - Develop theory



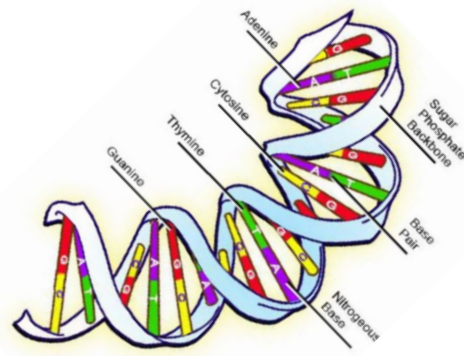
- Linear models
- Models of variation
- Models of fixed and random effects

- Used for:
  - Quantifying observed patterns and relationships (response vs explanatory)
  - Understanding and predict

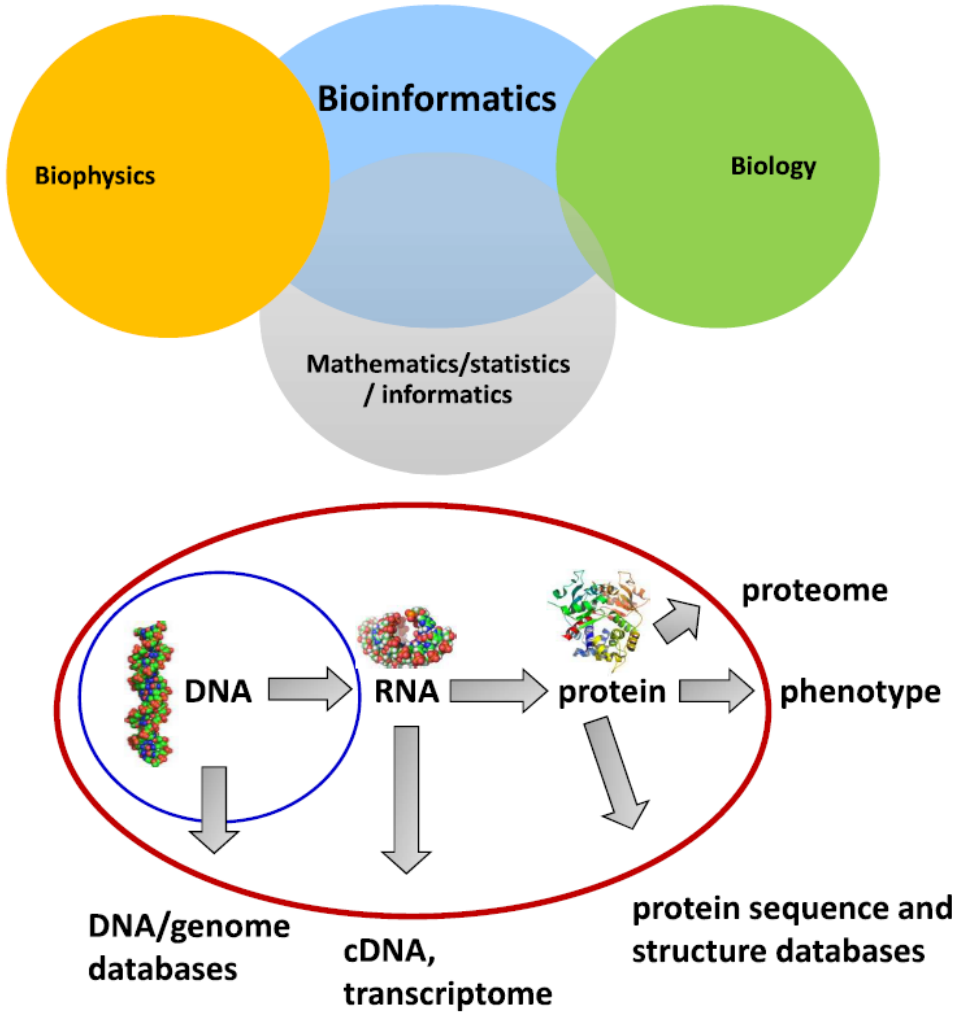
- Organize, clean and save data
- Analyze data
  - Models
    - Statistics
    - Algorithms

- Used for:
  - Making info in data accessible
  - Quantifying and understanding observed patterns

Biological organization



Key components of Bioinformatics



The National Center for Biotechnology Information (NCBI) definition.

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

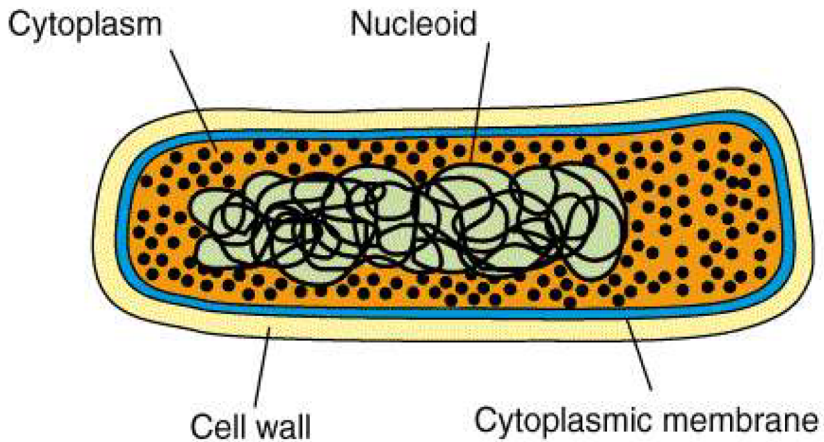
There are three important sub-disciplines within bioinformatics:

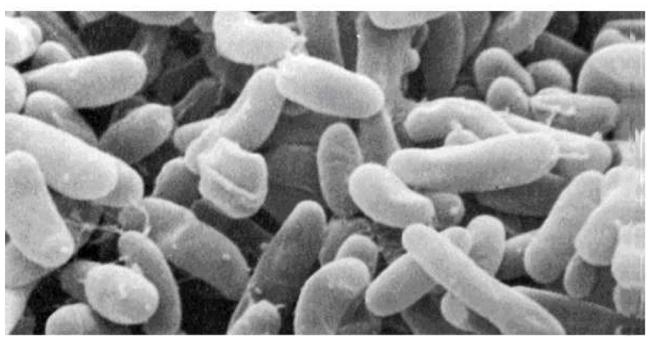
1. The development of new algorithms and statistics with which to assess relationships among members of large data sets.
2. The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures.
3. The development and implementation of tools that enable efficient access and management of different types of information.

Outline

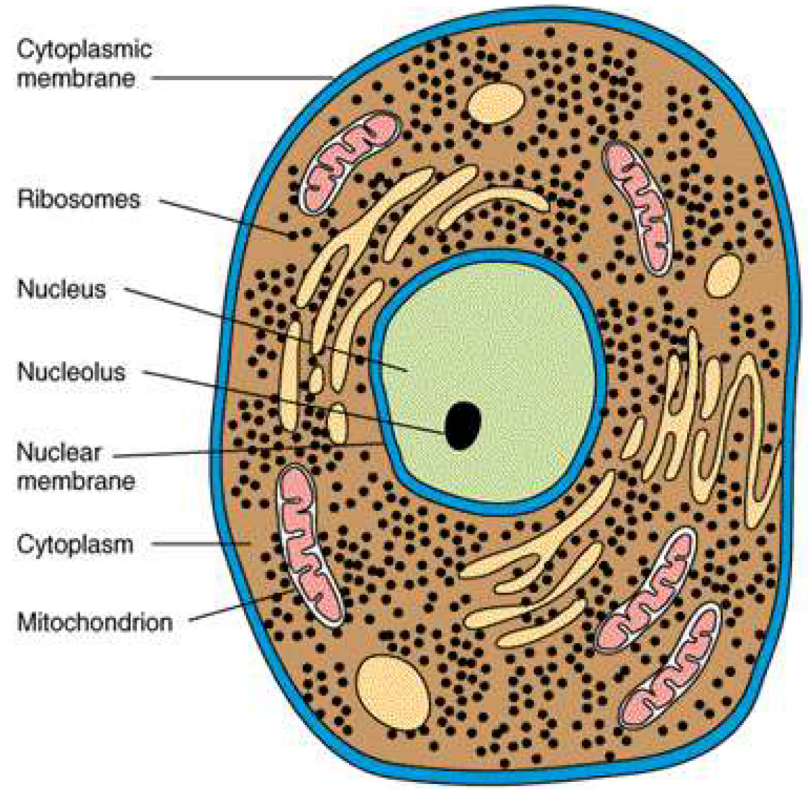
- Some basic biological concepts
- Some models and methods for sequence analyses
- A roadmap of the path from databases, sequence analysis, protein structure, organismal phenotype and phylogeny

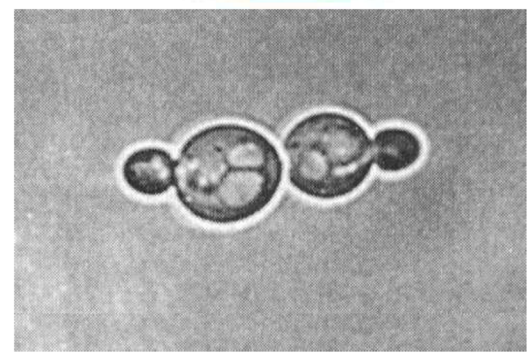
Bacterial/Archaeal cell



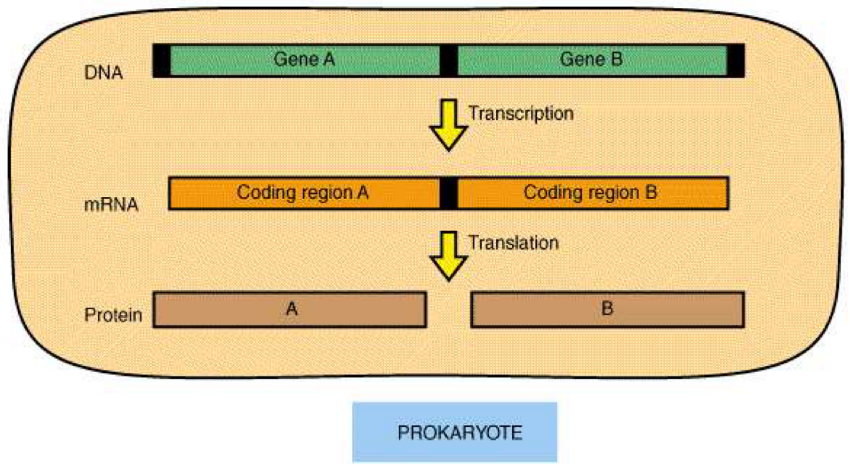


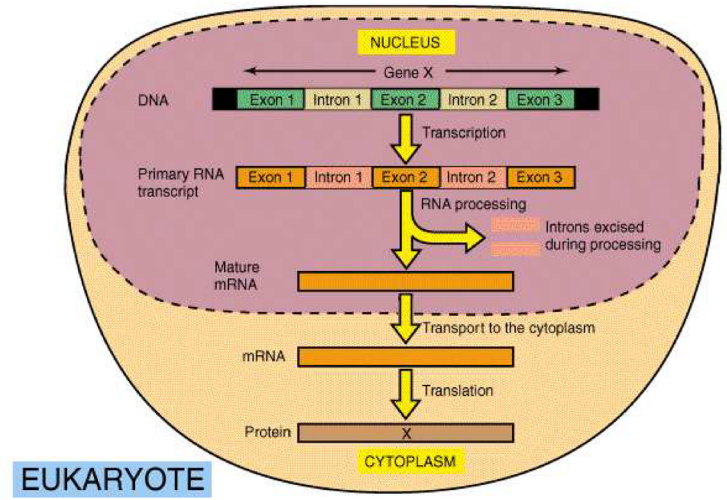
Eukaryotic cell





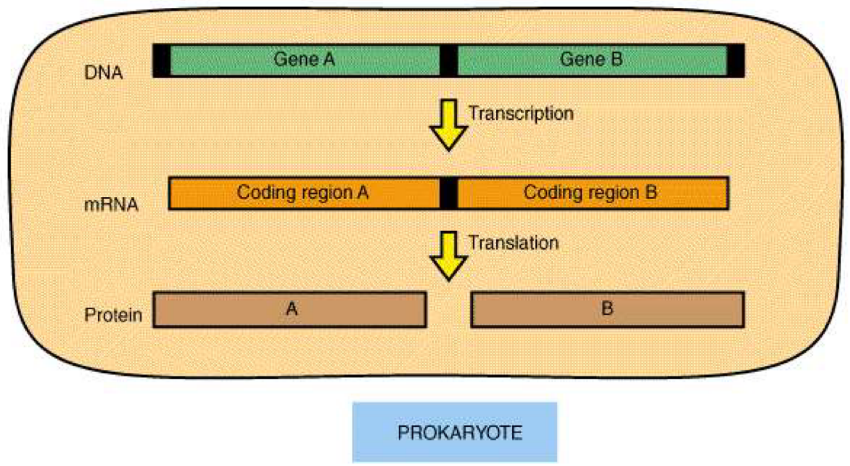
The Central Dogma of Molecular Biology

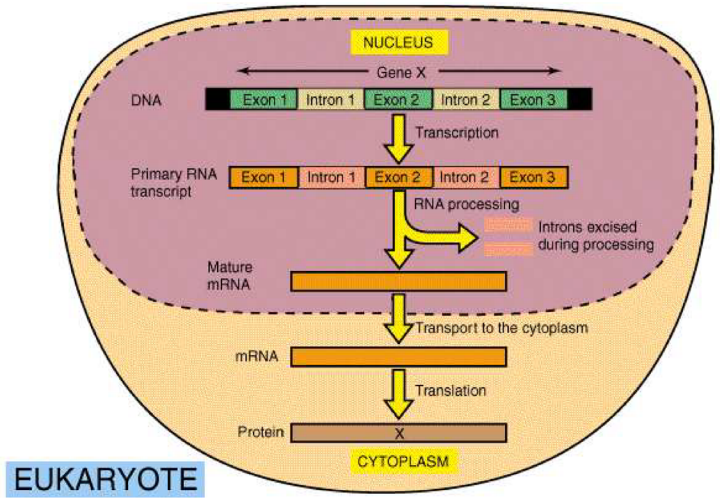


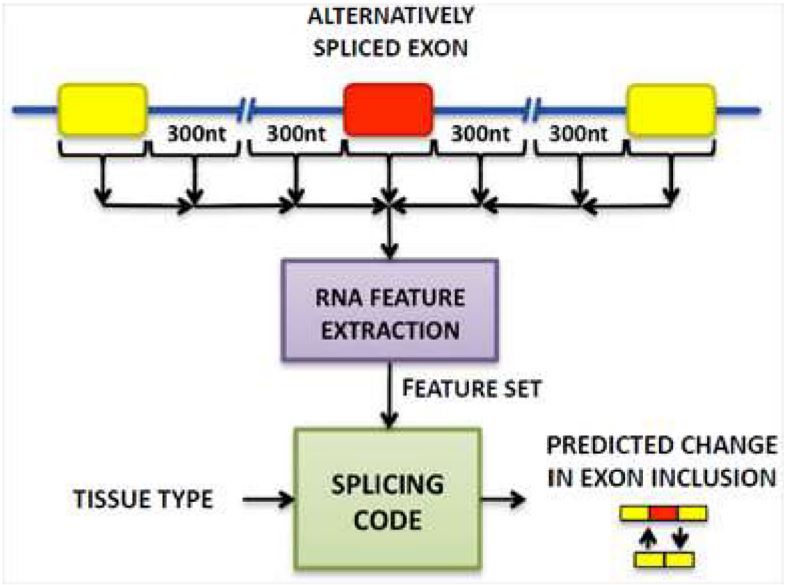


Source: Y. Barash, et al. Deciphering the Splicing Code. Nature, 465:7294, May 6, 2010.

The Central Dogma of Molecular Biology





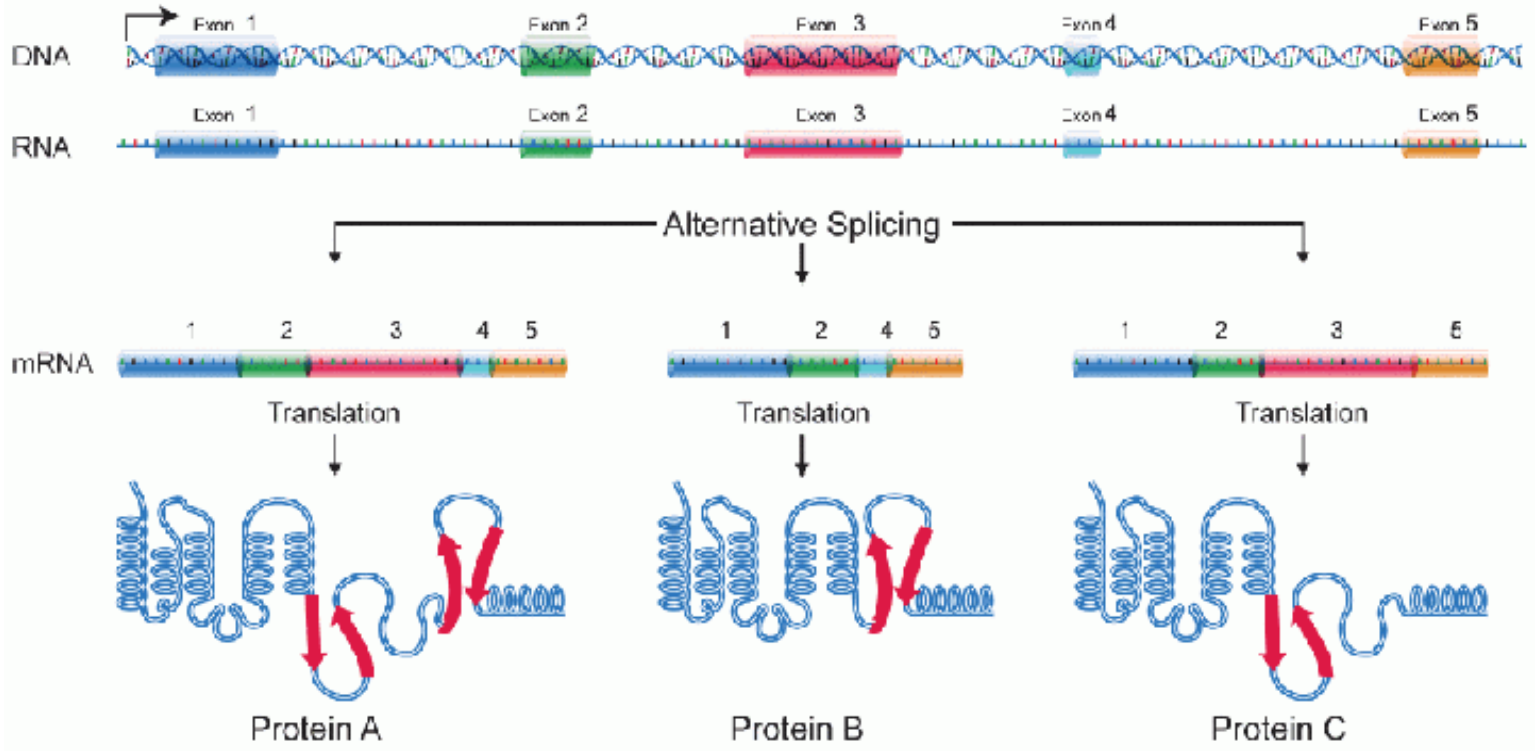


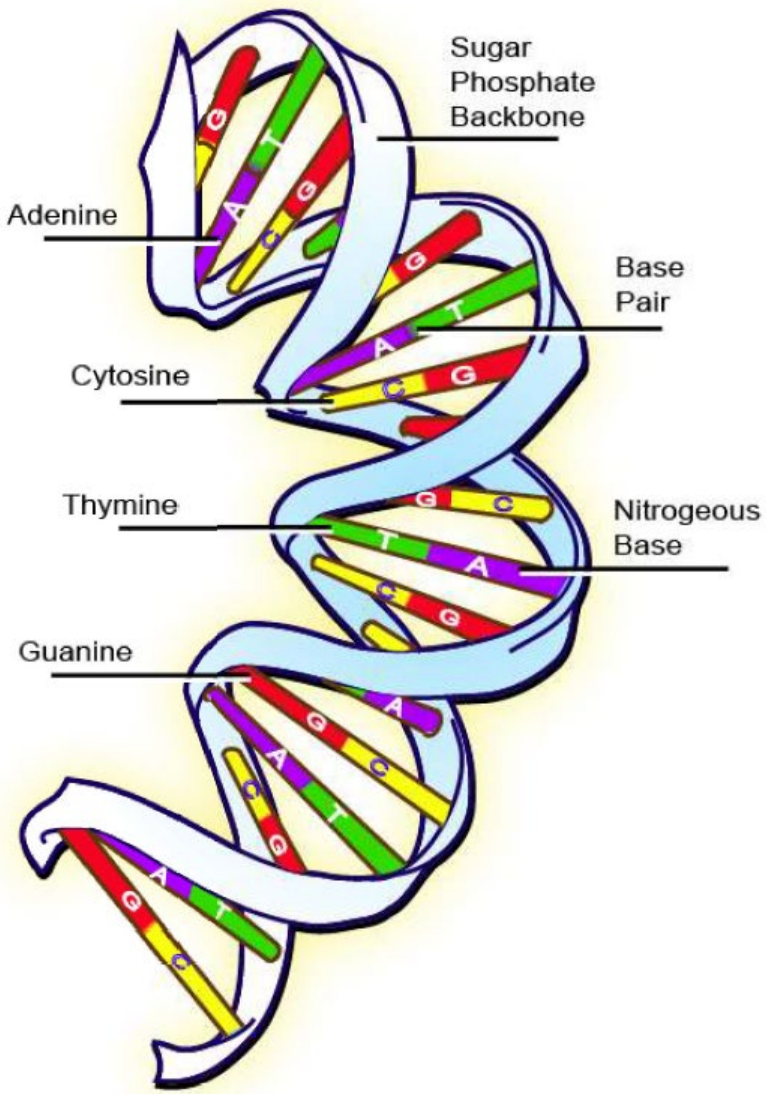
- Can we identify genes?
- Can we infer gene function?
- Can we understand the complexity of alternative splicing?

Source: Y. Barash, et al. Deciphering the Splicing Code. Nature, 465:7294, May 6, 2010.

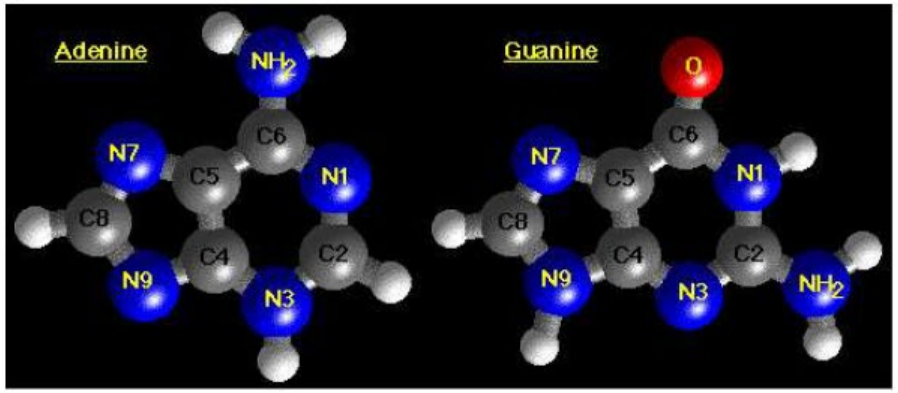
Genetic and genomic mechanisms, basis for function

Alternative splicing

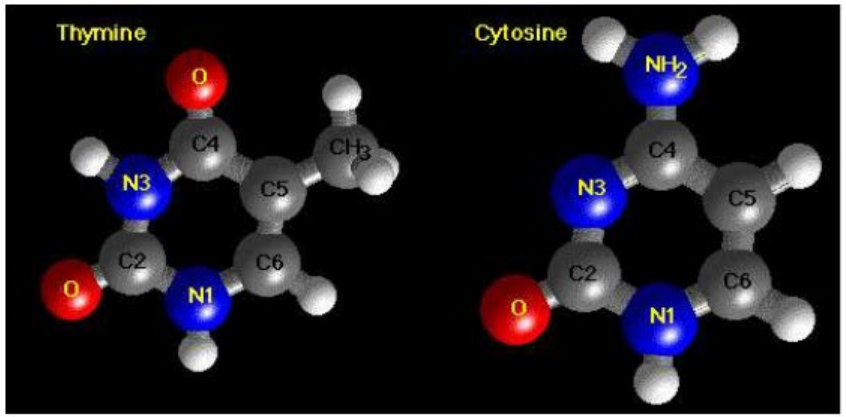




Purine Bases. **A**denine and **G**uanine. Purines are the larger of the two types of bases found in DNA.



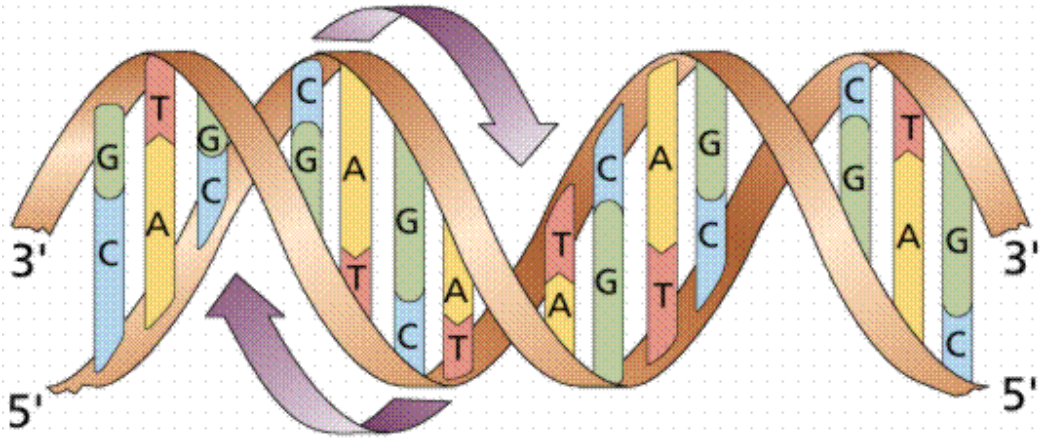
Pyrimidine Bases. **T**hymine **and** **C**ytosine.



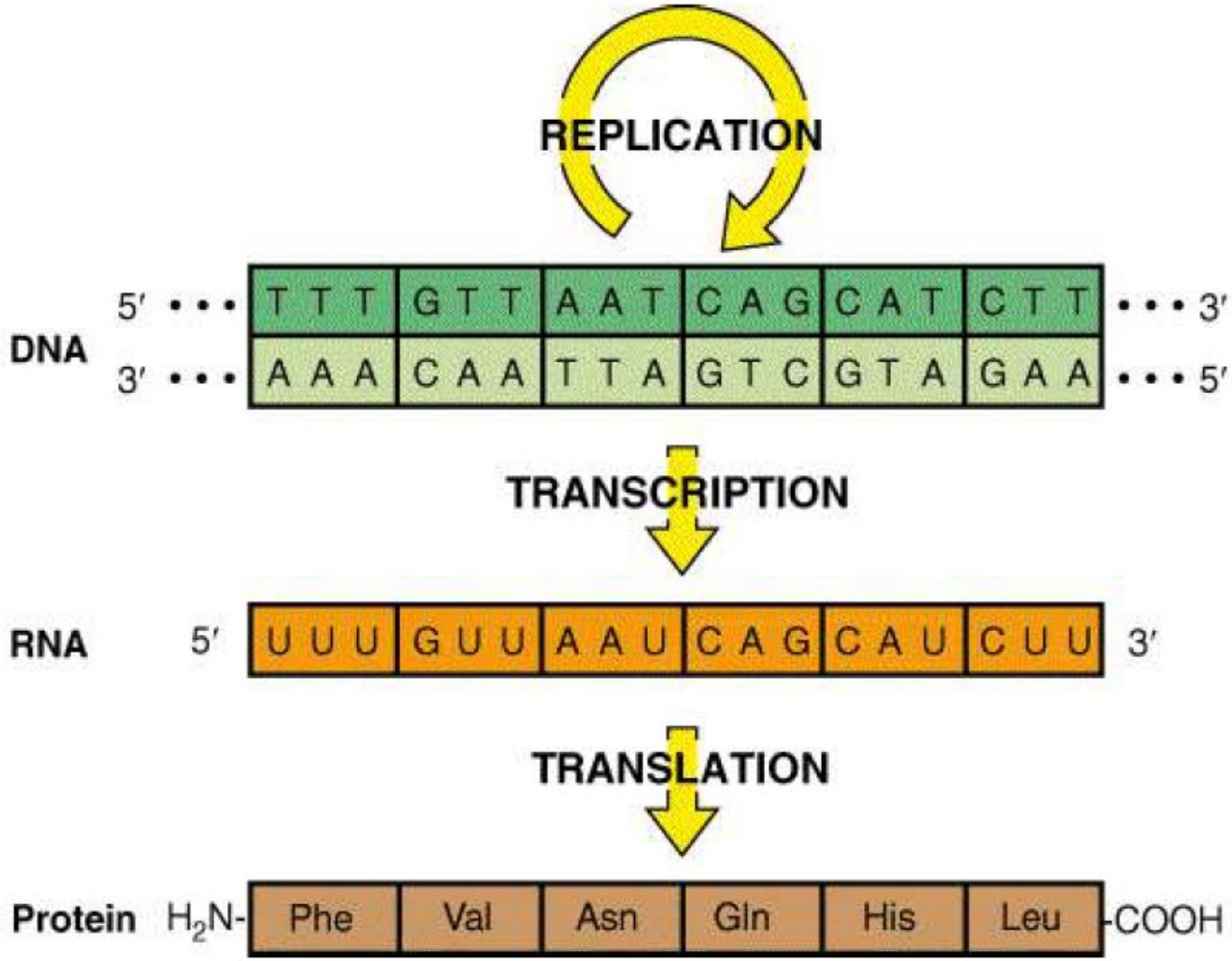
5′- ATCCGCGTCGATTCGATGCGTTAGATGCTGTGCCAAAACA-3′

3′- TAGGCGCAGCTAAGCTACGCAATCTACGACACGGTTTTGT-5′

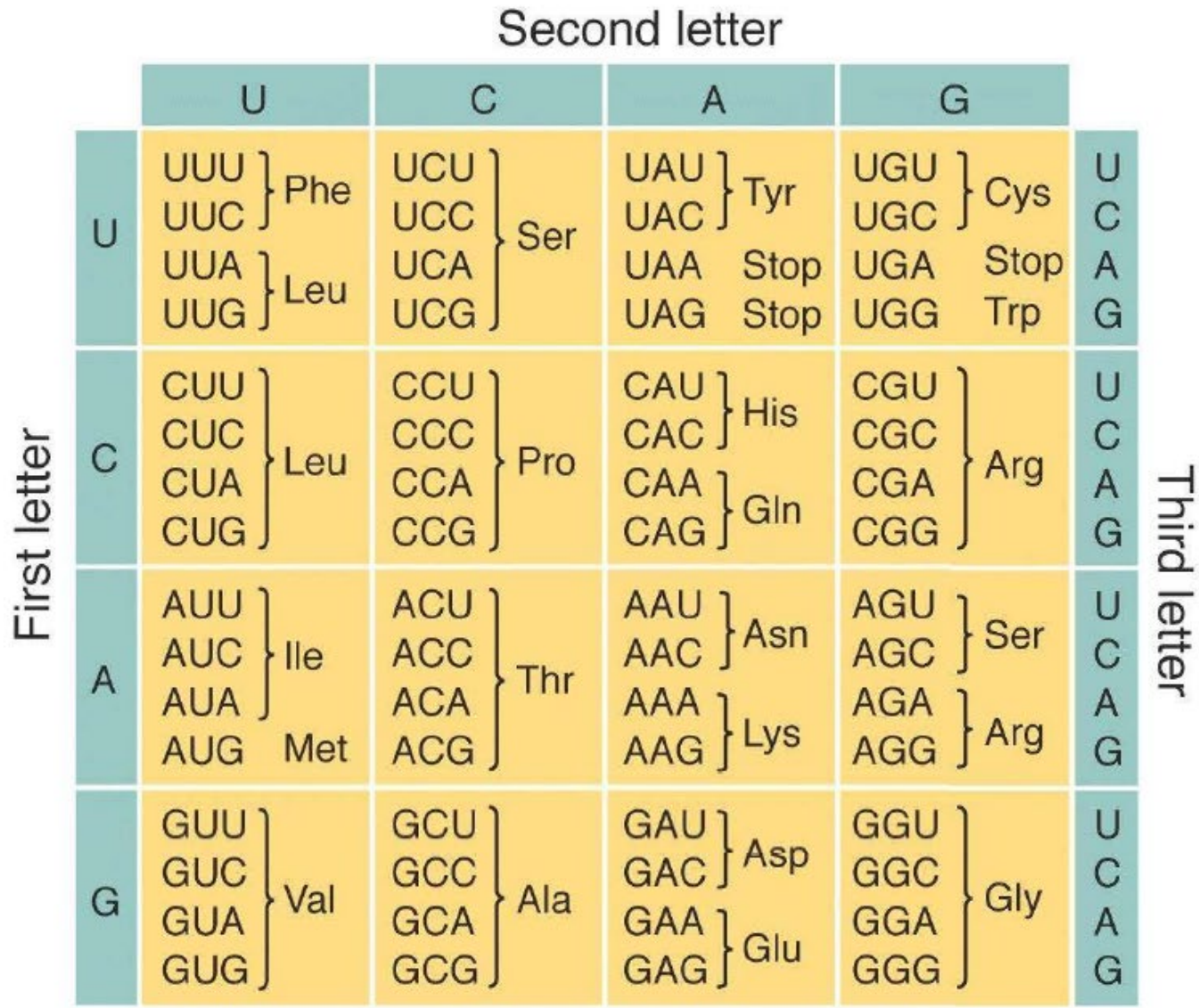
Complimentary strand:



Bioinformatics is about genetically coded information (DNA/RNA) and Proteins



The (standard) genetic code



The genetic code consists of 64 triplets of nucleotides ($4*4*4=4^3=64$)

Sequence analysis

Save and import your data

>NC_012920.1 Homo sapiens mitochondrion, complete genomeGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCT
CCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTG
CGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGA
TTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGC
GAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAA
TAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATA
ACAAAAAATTTCCACCAAACCCCCCCTCCCCCGCTTCTGGCCACAGC
ACTTAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCTAACAC
CAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCACTTTTA
ACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACT
ACTAATCTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACAC
ACCGCTGCTAACCCCATACCCCGAACCAACCAAACCCCAAAGACACC
CCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATACACTGAAAATG
TTTAGACGGGCTCACATCACCCCATAAACAAATAGGTTTGGTCCTAG
CCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCC
AGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCA
AGCACGCAGCAATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCC
ACGGGAAACAGCAGTGATTAACCTTTAGCAATAAACGAAAGTTTAAC
TAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGT
TTTAGATCACCCCCTCCCCAATAAAGCTAAAACTCACCTGAGTTGTAA
AAAACTCCAGTTGACACAAAATAGACTACGAAAGTGGCTT…..

Know your data

5′- ATCCGCGTCGATTCGATGCGTTAGATGCTGTGCCAAAACA-3′

Length L: 40 bp

Base composition: 25% A
25% T
25% C
25% G

G+C in some organisms

- *Homo sapiens* 46%
- *Halobacterium* 68%
- *Plasmodium falciparum* 19%

Local variation

Find genes

Open reading frame ORF

- An ORF is a continuous stretch of codons that begins with a start codon (usually AUG) and ends at a stop codon (usually UAA, UAG or UGA).

- Identifying the start and stop codons for translation determines the protein coding section, or open reading frame (ORF), in a sequence.

- Once you know the ORF for a gene or mRNA, you can translate a nucleotide sequence to its corresponding amino acid sequence.

Compare sequences

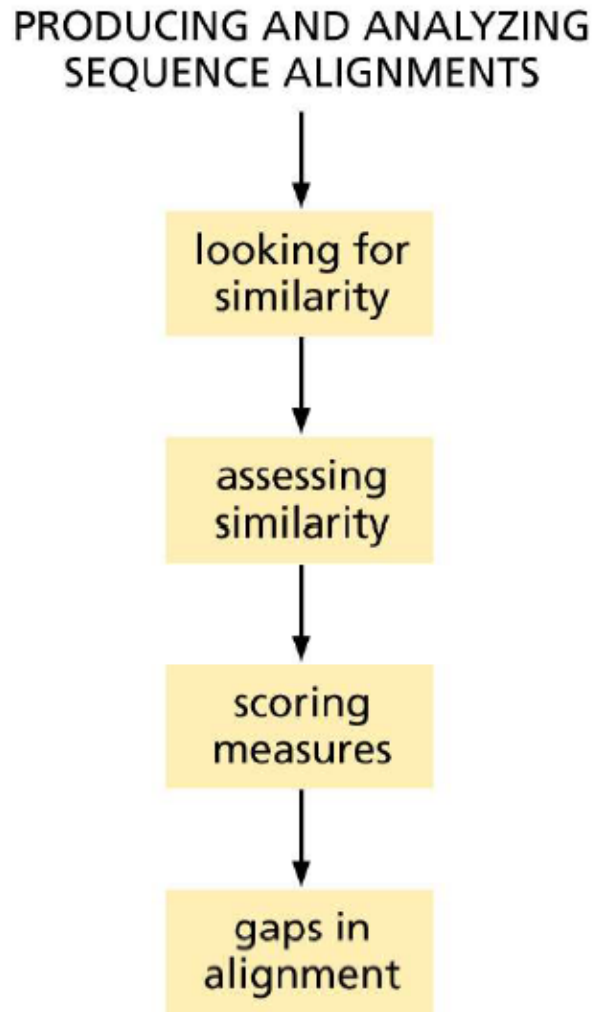
Comparing two sequences

1. **Choose two sequences**
2. **Select an algorithm that generates a score**
3. **Allow gaps (insertions, deletions)**
4. **Score reflects degree of similarity**
5. **Estimate probability that the alignment occurred by chance**

Two alternative strategies for alignments

Global alignment (Needleman & Wunsch 1970)

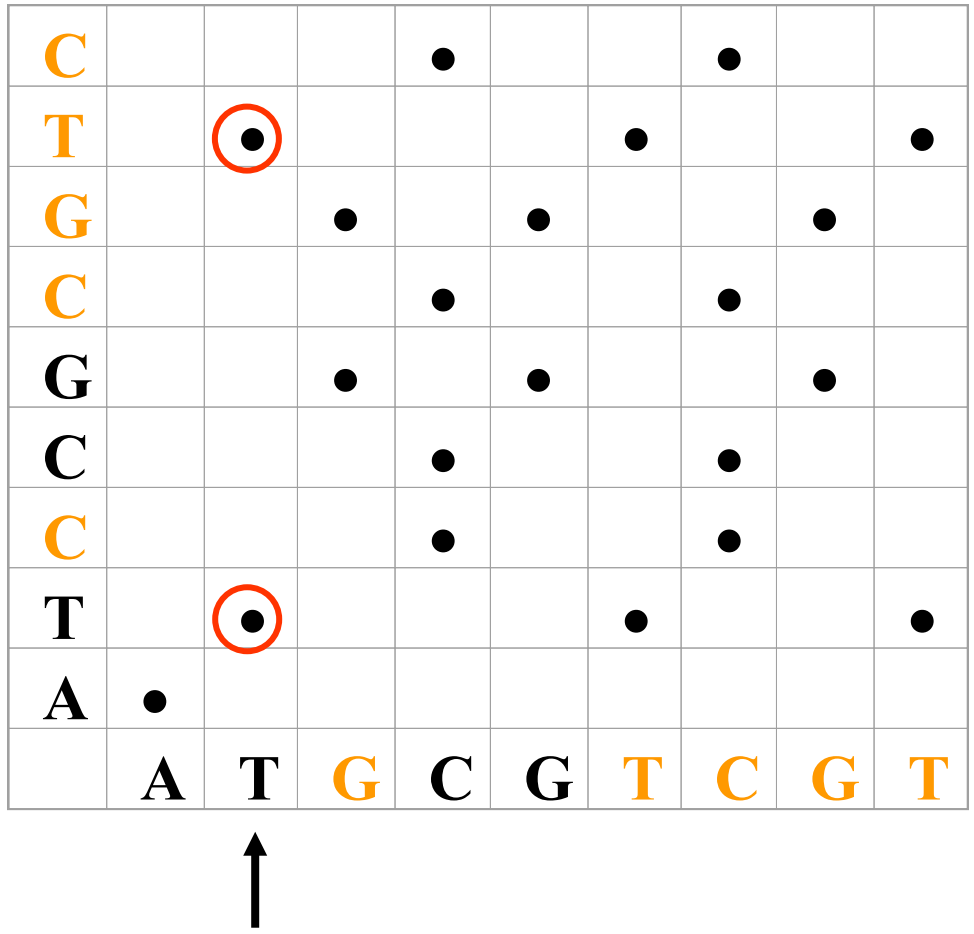
Local alignment (Smith & Waterman 1981)



Two sequences:

ATCCGCGTC

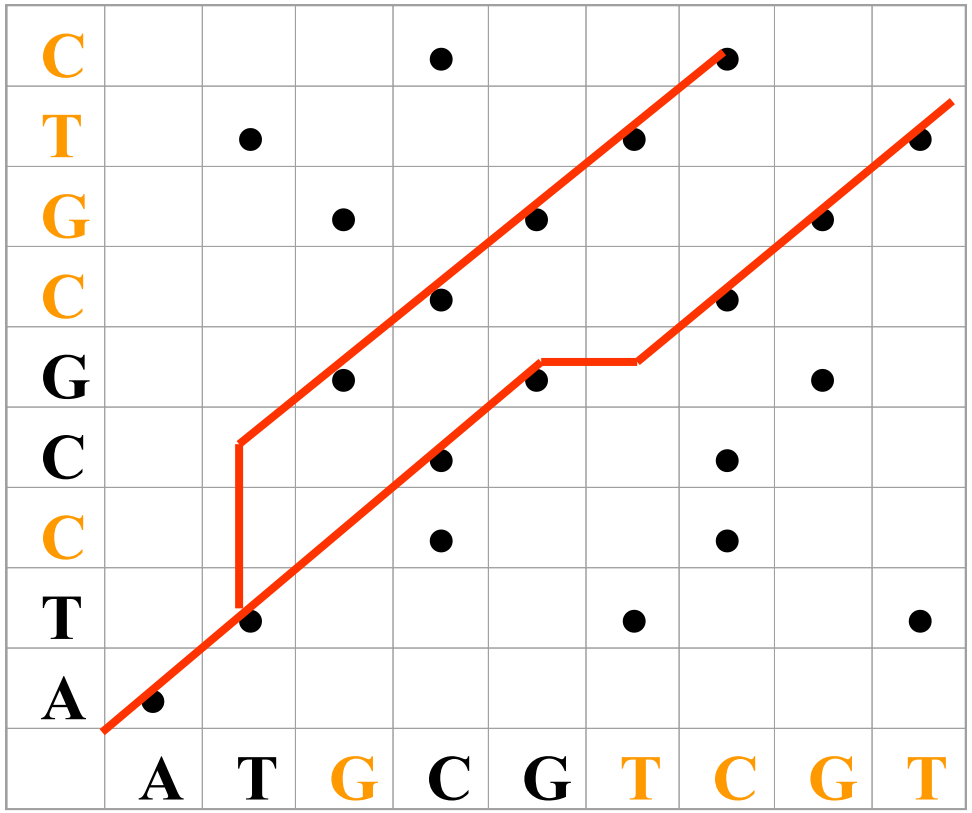
ATGCGTCGT



How do we determine their relative arrangement, sliding left/right, gaps…

2

```
ATCCGCGTC
||   |||||
AT--GCGTCGT
```



1

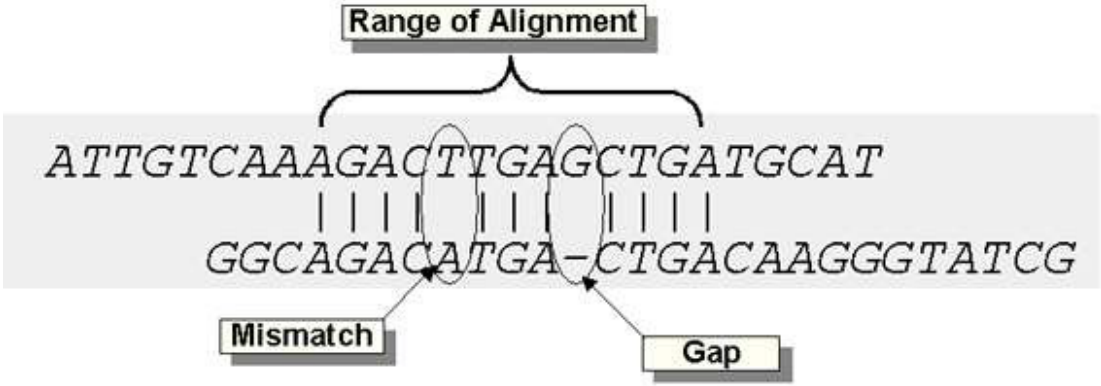
```
ATCCG-CGTC
||  ||  |||
ATGCGTCGT
```

Which alignment is correct?

Penalty scores

Calculation of an alignment score.

The score describes the overall quality of the alignment.



$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

How do we find the best alignment?

- Brute force approach:

  generate all possible alignments between two sequences.

* Problem – very time consuming two sequences of length 250 requires ~10^{149} alignments.

Dynamic programming

(Richard Bellman 1953)

Dynamic programming is a very general optimization technique that can be applied any time a problem can be recursively subdivided into two similar sub problems of smaller size, such that the solution to the larger problem can be obtained by piecing together the solutions to the two sub-problems.

Global: The Needleman–Wunsch algorithm

Local: The Smith-Waterman algorithm

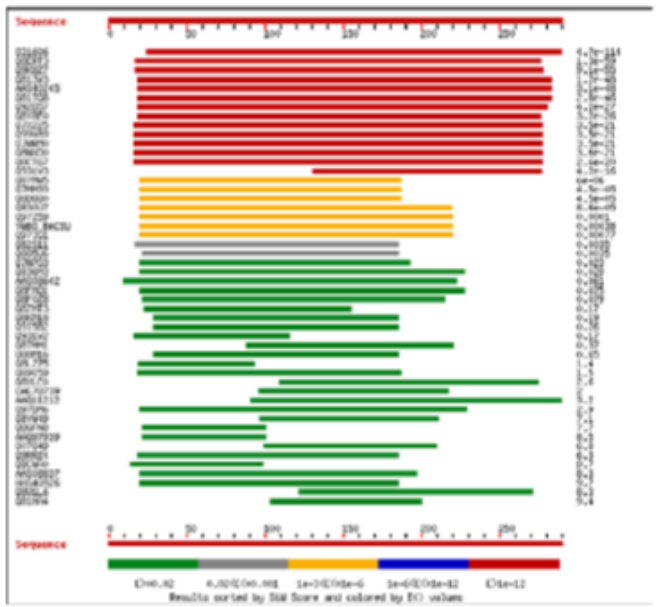
Video: Global alignement with Needleman-Wunsch

Sequence analysis

Database search

Database searching methods

In database searches a query sequence is aligned with each of the sequences (called targets) in a database. A score is computed and the query/target pairs with the best score are reported.



Applications include

- identifying orthologs and paralogs
- discovering new genes or proteins
- discovering variants of genes or proteins
- exploring protein structure and function

Sequence analysis

Protein structure and organismal phenotype

Animation

Genetic distance

```
OTU1   ATCCGCGTCGATTCGATGCGTTAGATGCTGTGCCAAAACA
OTU2   ATTCCCGTCGATACGAAGCGTTACCTGCTGAGCGAAATGA
OTU3   ATCCCCGTCGATTCGAAGCGTTACATCCTGTGCCAAAACA
```

OTU = operational taxonomic unit

1. Number of differences

	OTU1	OTU2	OTU3
OTU1		10	4
OTU2	25%		8
OTU3	10%	20%	

But what does the differences mean?

Structural basis for alignment

- When two protein sequences have more than 20 to 30% identical residues aligned the corresponding 3-D structures are almost always very similar.
- Overall folds are identical but structures may differ in detail.
- Form often follows function.
- Sequence alignment is an approximate predictor of the underlying 3-D structural alignment (utilized in comparative modelling).
- Homologous proteins (proteins that evolved from a common ancestor) always have similar structures, and sometimes have similar functions.
- Non-homologous proteins have different structures

Bioinformatics Methods to Predict Protein Structure and Function

A Practical Approach

Yvonne J. K. Edwards* and Amanda Cottage

Abstract

Protein structure prediction by using bioinformatics can involve sequence similarity searches, multiple sequence alignments, identification and characterization of domains, secondary structure prediction, solvent accessibility prediction, automatic protein fold recognition, constructing three-dimensional models to atomic detail, and model validation. Not all protein structure prediction projects involve the use of all these techniques. A central part of a typical protein structure prediction is the identification of a suitable structural target from which to extrapolate three-dimensional information for a query sequence. The way in which this is done defines three types of projects. The first involves the use of standard and well-understood techniques. If a structural template remains elusive, a second approach using nontrivial methods is required. If a target fold cannot be reliably identified because inconsistent results have been obtained from nontrivial data analyses, the project falls into the third type of project and will be virtually impossible to complete with any degree of reliability. In this article, a set of protocols to predict protein structure from sequence is presented and distinctions among the three types of project are given. These methods, if used appropriately, can provide valuable indicators of protein structure and function.

Index Entries: Molecular modeling; sequence similarity searches; multiple sequence alignment; identification and characterization of domains; secondary structure prediction; solvent accessibility prediction; automatic protein fold recognition.

Sequence analysis

Phylogeny

Genetic distance

```
OTU1   ATCCGCGTCGATTCGATGCGTTAGATGCTGTGCCAAAACA
OTU2   ATTCCCGTCGATACGAAGCGTTACCTGCTGAGCGAAATGA
OTU3   ATCCCCGTCGATTCGAAGCGTTACATCCTGTGCCAAAACA
```

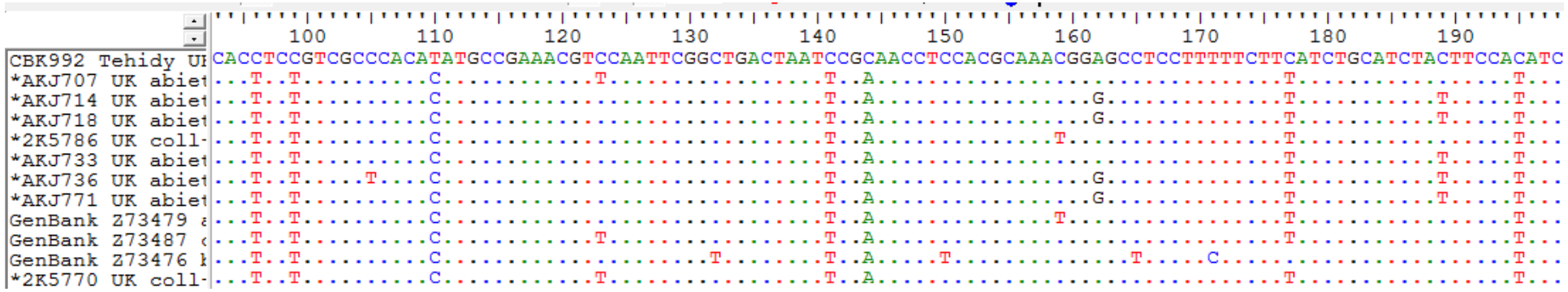
OTU = operational taxonomic unit

1. Number of differences

	OTU1	OTU2	OTU3
OTU1		10	4
OTU2	25%		8
OTU3	10%	20%	

But what does the differences mean?

Not all mutations are equally likely



C ⟷ T

[G ⟷ A]

Kinds of substitutions

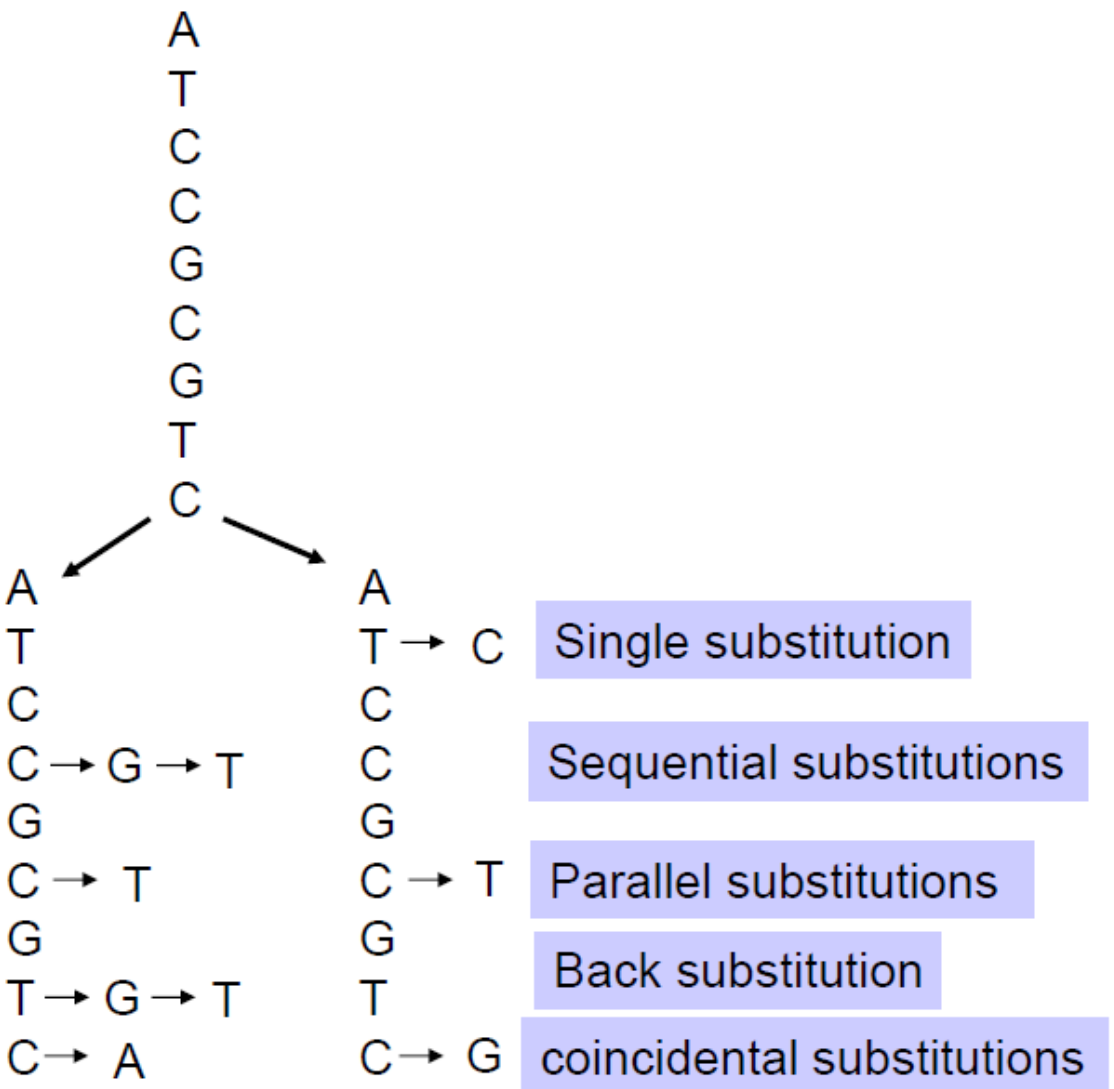
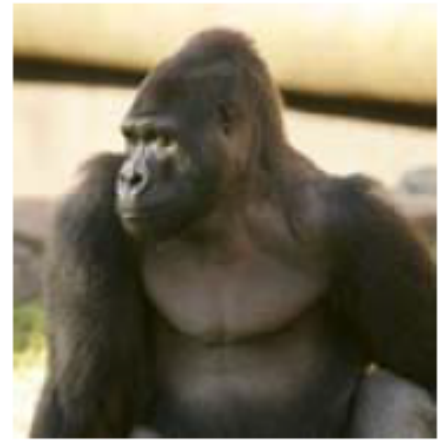


Fig 7.15

Phylogenetic reconstruction from DNA sequences






- Find/decide/assume a model representing how the DNA has evolved

Genetic distance

```
OTU1   ATCCGCGTCGATTCGATGCGTTAGATGCTGTGCCAAAACA
OTU2   ATTCCCGTCGATACGAAGCGTTACCTGCTGAGCGAAATGA
OTU3   ATCCCCGTCGATTCGAAGCGTTACATCCTGTGCCAAAACA
```

- By using the probabilities for substitutions in a evolutionary model we can estimate the likelihood of different changes occurring between the OTU's analysed.
- By analysing the number of changes we can thus also estimate "true" evolutionary distances

Genetic similarity – Evolutionary similarity – Relatedness – Phylogenetic tree