MSc Bioinformatics
Course
Fundamentals of Bioinformatics

Lecture 3: Genome Principles, Evolution and Homology Searching

Centre for Integrative Bioinformatics VU (IBIVU)
Faculty of Exact Sciences / Faculty of Earth and Life Sciences
http://ibi.vu.nl, heringa@few.vu.nl, 87649 (Heringa), Room P1.28
OH BRAD, THEY SAY THERE'S DNA IN MY BODY!

WHO CARES, DARLING, ...WHO CARES...
C-G base pairing is more stable than A-T (A-U) base pairing.

3rd codon position has freedom to evolve (synonymous mutations).

Species can therefore optimise their G-C content (e.g. thermophiles are GC rich) *(consequences for codon use?)*
DNA compositional biases

- Base compositions of genomes: G+C (and therefore also A+T) content varies between different genomes

- The GC-content is sometimes used to classify organism in taxonomy

- **High** G+C content bacteria: Actinobacteria e.g. in *Streptomyces coelicolor* it is 72%

  Low G+C content: *Plasmodium falciparum* (~20%)

- Other examples:
  - *Saccharomyces cerevisiae* (yeast) 38%
  - *Arabidopsis thaliana* (plant) 36%
  - *Escherichia coli* (bacteria) 50%
<table>
<thead>
<tr>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are about 20,000 – 23,000 genes in the human genome (~ 3% of the genome)</td>
</tr>
<tr>
<td>Average gene length is ~ 8,000 bp</td>
</tr>
<tr>
<td>Average of 5-6 exons per gene</td>
</tr>
<tr>
<td>Average exon length is ~ 200 bp</td>
</tr>
<tr>
<td>Average intron length is ~ 2000 bp</td>
</tr>
<tr>
<td>8% of the genes have a single exon</td>
</tr>
<tr>
<td>Some exons can be as small as 1 or 3 bp</td>
</tr>
</tbody>
</table>
The largest known human gene is DMD, which stands for “Dystrophin (muscular dystrophy, Duchenne and Becker types)”

The gene encodes the protein dystrophin: the gene’s size is $\sim 2.4$ million bp over 79 exons.

X-linked recessive disease (affects boys)

Two variants: Duchenne-type (DMD) and becker-type (BMD)

Duchenne-type: more severe, frameshift-mutations

Becker-type: milder phenotype, “in frame”- mutations
Nucleic acid basics

- Nucleic acids are polymers
- Each monomer consists of 3 moieties
Nucleic acid basics (2)

- A base can be of 5 rings

**Purines**
- purine
- adenine
- guanine

**Pyrimidines**
- pyrimidine
- cytosine
- uracil
- thymine

Purines and Pyrimidines can base-pair (Watson-Crick pairs)

![Watson and Crick, 1953](image)

Most stable bond
Nucleic acid as hetero-polymers

- DNA and RNA strands

 REMEMBER:

- **DNA** = deoxyribonucleic acid;
- **RNA** = ribonucleic acid (OH-groups at the 2’ position)
- Note the **directionality** of DNA (5’-3’ & 3’-5’) or RNA (5’-3’)
So … 

DNA

RNA

Diagram showing the structures of DNA and RNA with their respective nucleotides and base pairing.
Human DNA

• There are at least 3bn ($3 \times 10^9$) nucleotides in the nucleus of almost all of the trillions ($\sim 5-10 \times 10^{12}$) of cells within a human body (an exception is, for example, red blood cells which have no nucleus and therefore no DNA) – a total of $\sim 10^{22}$ nucleotides!

• Many DNA regions code for proteins, and are called genes (1 gene codes for 1 protein as a base rule, but the reality is a lot more complicated)

• Human DNA contains $\sim 23,000$ expressed genes

• Human DNA is not dense when judged using protein coding genes: only about 3% is used.
Human DNA (Cont.)

• All people are different, but the DNA of different people only varies for 0.1% or less. Evidence in current genomics studies (Single Nucleotide Polymorphisms or SNPs) imply that on average only 1 nucleotide out of about 1000 is different between individuals. Over the whole genome, this means that some 3 million letters would differ between individuals.

• Deoxyribonucleic acid (DNA) comprises 4 different types of nucleotides: adenine (A), thiamine (T), cytosine (C) and guanine (G). These nucleotides are sometimes also called bases.

• The structure of DNA is the so-called double helix, discovered by Watson and Crick in 1953, where the two helices are cross-linked by A-T and C-G base-pairs (nucleotide pairs – so-called Watson-Crick base pairing).
Genetic diseases: **cystic fibrosis**

- Autosomal, recessive, hereditary disease (Chr. 7)
- CF is most common among Caucasians and Ashkenazi Jews; one in 25 people of European descent carry one gene for CF. Approximately 30,000 Americans have CF, making it one of the most common life-shortening inherited diseases

- Symptoms:
  - Exocrine glands (which produce sweat and mucus)
  - Abnormal secretions
  - Respiratory problems
  - Reduced fertility and (male) anatomical anomalies
Gene product: CFTR (cystic fibrosis transmembrane conductance regulator)

CFTR is an ABC (ATP-binding cassette) transporter or traffic ATPase.

These proteins transport molecules such as sugars, peptides, inorganic phosphate, chloride, and metal cations across the cellular membrane.

CFTR transports chloride ions (Cl\(^-\)) ions across the membranes of cells in the lungs, liver, pancreas, digestive tract, reproductive tract, and skin.
CF gene CFTR has 3-bp deletion leading to Del508 (Phe) in 1480 aa protein (epithelial Cl⁻ channel)

Protein degraded in Endoplasmatic Reticulum (ER) instead of inserted into cell membrane

The deltaF508 deletion is the most common cause of cystic fibrosis. The isoleucine (Ile) at amino acid position 507 remains unchanged because both ATC and ATT code for isoleucine. There are a number of other mutations known, all leading to a disfunctional protein.
Protein structure hierarchical levels

PRIMARY STRUCTURE (amino acid sequence)

VHLTPEEKSAVTALWGKVNVDE
VGGEALGRLLVYPWTQRFEE
SFGLSTPDVMGNPKVKAHG
KKVLGAFAEGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGN
VLVCVLAHFGEFTPPVQAAY
QKVVAGVANALAHKYH

SECONDARY STRUCTURE (helices, strands)

QUATERNARY STRUCTURE (oligomers)

TERTIARY STRUCTURE (fold)
Main classes of proteins

Proteins were first named by the Dutch chemist Gerhardus Johannes Mulder (1802-1880) and described by the Swedish chemist Jöns Jakob Berzelius in a publication in 1839.

A protein is a polymer (polypeptide) consisting of amino acids as building blocks, each divided in a main- and side-chain part.

Proteins can occur at all localisations (and organelles) within a cell and outside of it. There are two main types:

• **Soluble proteins**: emerged in a watery environment (e.g. cell lumen)
• **Membrane proteins**: attached to a cellular membrane (lipid bi-layer)

Three different computer-rendered representations of the three-dimensional structure of the protein triose phosphate isomerase
Back to genome information

• Craig Venter (2001)
  – “genome size dictates complexity of the organism” -> not true
<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$X-174 virus</td>
<td>5,386</td>
</tr>
<tr>
<td>Epstein Bar Virus</td>
<td>172,282</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>580,000</td>
</tr>
<tr>
<td><em>Hemophilus Influenza</em></td>
<td>$1.8 \times 10^6$</td>
</tr>
<tr>
<td>Yeast (<em>S. Cerevisiae</em>)</td>
<td>$12.1 \times 10^6$</td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td><strong>$3.2 \times 10^9$</strong></td>
</tr>
<tr>
<td>Wheat</td>
<td>$16 \times 10^9$</td>
</tr>
<tr>
<td><em>Lilium longiflorum</em></td>
<td>$90 \times 10^9$</td>
</tr>
<tr>
<td>Salamander</td>
<td>$100 \times 10^9$</td>
</tr>
<tr>
<td><em>Amoeba dubia</em></td>
<td>$670 \times 10^9$</td>
</tr>
</tbody>
</table>
Genome size and number of genes: not related!

<table>
<thead>
<tr>
<th>Organism</th>
<th>DNA length</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mycoplasma genitalium</em></td>
<td>0.5 Mb</td>
<td>470</td>
</tr>
<tr>
<td><em>Deinococcus radiodurans</em></td>
<td>3 Mb in 4-10 copies!</td>
<td>3 200</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4.5 Mb</td>
<td>4 400</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>12 Mb</td>
<td>6 200</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>97 Mb</td>
<td>19 000</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>120 Mb</td>
<td>18 000</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>3200 Mb</td>
<td>23.000</td>
</tr>
<tr>
<td><em>Arabidopsis</em></td>
<td>120 Mb</td>
<td>26.000</td>
</tr>
</tbody>
</table>
Genome size and number of genes: not related?

What is life?

• NASA astrobiology program:
  “Life is a self-sustained chemical system capable of undergoing Darwinian evolution”
“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

“Nothing in bioinformatics makes sense except in the light of Biology”
Divergent Evolution

Four requirements:

• **Template** structure providing stability (DNA)
• **Copying** mechanism (meiosis)
• Mechanism providing **variation** (mutations; insertions and deletions; crossing-over; etc.)
• **Selection**: some traits lead to greater fitness of one individual relative to another. Darwin coined “*survival of the fittest*”

Evolution is a conservative process: the vast majority of mutations will not be selected (i.e. will not make it as they lead to worse performance or are even lethal) – this is called **negative** (or **purifying**) selection
Properties of Evolution

• By repeated selection, evolution works as an optimisation process

• **However:**
  – The objective function that is optimised changes all the time ("survival of the fittest")
  – Evolution is often a (spatially) local process
  – So, evolution is a spatio-temporal locally (de)coupled process

• In the 1970s and early 80s, evolution became strongly viewed as an optimisation process (‘optimal foraging’, ‘optimal mating’, etc.) – sometimes with amusing consequences
  – Optimality, non-optimality, pluralism, selection, drift
  – The idea of overriding importance of selection became changed by the work of Motoo Kimura. His neutral selection theory was published in 1980.
Orthologous genes are homologous (corresponding) genes in different species -- red arrow

Paralogous genes are homologous genes resulting from a duplication event within the same species (genome) – blue arrow
After a gene duplication event, paralogous sequences typically accumulate many mutations as a result of lowered selection pressure.
Changing molecular sequences

- **Mutations**: changing nucleotides (‘letters’) within DNA, also called ‘*point mutations*’

- A & G: purines, C & T/U: pyrimidines:
  - **Transition**: purine -> purine or pyrimidine -> pyrimidine
  - **Transversion**: purine -> pyrimidine or pyrimidine -> purine
Types of point mutation

- **Synonymous mutation**: mutation that does not lead to an amino acid change (where in the codon are these expected?)

- **Non-synonymous mutation**: does lead to an amino acid change
  - **Missense mutation**: one a.a. replaced by other a.a.
  - **Nonsense mutation**: a.a. replaced by stop codon (what happens with protein?)
Ka/Ks Ratios

• **Ks** is defined as the number of synonymous nucleotide substitutions per synonymous site

• **Ka** is defined as the number of nonsynonymous nucleotide substitutions per nonsynonymous site

• The **Ka/Ks** ratio is used to estimate the type of selection exerted on a given gene or DNA fragment

• Need aligned orthologous sequences to do calculate **Ka/Ks** ratios (we will talk about alignment later).
The frequency of different values of $Ka/Ks$ for 835 mouse–rat orthologous genes. Figures on the $x$ axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1.
Ka/Ks ratios

Three types of selection:
1. Negative (purifying) selection -> Ka/Ks < 1
2. Neutral selection (Kimura) -> Ka/Ks ~= 1
3. Positive selection -> Ka/Ks > 1
Divergent evolution

Ancestral sequence: ABCD

ACCD (B→C) mutation

ACCD or ACCD

ABD (C→∅) deletion

AB→D or A→BD

Pairwise Alignment
Divergent evolution

Ancestral sequence: ABCD

- ACCD (B→C) mutation
- ABD (C→∅) deletion

Pairwise Alignment

ACCĐ or ACCĐ
AB—D or A—BD

true alignment
A protein sequence alignment
MSTGAVLIY--TSILIKECHAMPAGNE------
---GGILLFHRTHELIKESHAMANDEGGSNNS
   *   *    *  ****  ***

A DNA sequence alignment
atccgtttggcaaatcgcctcctatccgggcttttaa
att---tggcggtgatcg--cctctctacgggcc----
  ***  *****  *****  **  **********
What can be observed about divergent evolution

Ancestral sequence

Sequence 1  Sequence 2

1: ACCTGTAATC
2: ACGTGCAGATC

D = 3/10 (fraction different sites (nucleotides))

(a) G C
One substitution - one visible
(b) G A
Two substitutions - one visible
(c) A A
Two substitutions - none visible
(d) G A
Back mutation - not visible
Evolution and three-dimensional protein structure information

What do we see if we colour code the space-filling (CPK) protein model?

• E.g., red for conserved alignment positions to blue for variable (unconserved) positions.
Evolution and three-dimensional protein structure information

Isocitrate dehydrogenase:

The distance from the active site (in yellow) determines the rate of evolution (red = fast evolution, blue = slow evolution)
Divergent evolution

• Common ancestor
• Sequences change over time
• Protein structures typically remain the same
• Therefore, function normally is preserved within orthologous families

“Structure more conserved than sequence”
Three main principles

- DNA makes RNA makes Protein

- Sequence → Structure → Function

- Structure more conserved than sequence
  - Function in turn is also more conserved than sequence
Convergent evolution

• Much less common than divergent evolution
• Often with shorter motifs (e.g. active sites)
• Motif (function) has evolved more than once independently, e.g. starting with two very different sequences adopting different folds
• Sequences and associated structures remain very different, but (functional) motif can become identical
• Classical example: serine proteinase and chymotrypsin
• If a given function in an organism is taken over by a gene (protein) that has undergone convergent evolution, this is called **non-orthologous displacement**
• **Important corollary:** sequences related through convergent evolution should **not** be aligned, since there is no common ancestry
Convergent evolution example
Serine proteinase (subtilisin), chymotrypsin and carboxypeptidase C

• Different evolutionary origins
• As proteinases these proteins chop up other proteins
• Similarities in the reaction mechanisms. Chymotrypsin, subtilisin and carboxypeptidase C have a **catalytic triad** of serine, aspartate and histidine in common: serine acts as a nucleophile, aspartate as an electrophile, and histidine as a base.
• The geometric orientations of the catalytic residues are similar between families, despite different protein folds.
• The linear arrangements of the catalytic residues reflect different family relationships. For example the catalytic triad in the chymotrypsin clan is ordered \( \text{HDS} \), but is ordered \( \text{DHS} \) in the subtilisin clan and \( \text{SDH} \) in the carboxypeptidase clan.
Serine proteinase (subtilisin) and chymotrypsin

Catalytic triads

chymotrypsin
serine proteinase
carboxypeptidase C

Read http://www.ebi.ac.uk/interpro/potm/2003_5/Page1.htm
Chymotrypsin

alpha-chymotrypsin (bovine)

elastase (porcine)

trypsin (Streptomyces griseus)

N-terminal residue (Ile 16) of active α-chymotrypsin

Tertiary structure of bovine alpha-chymotrypsin, with detail of residues of the catalytic triad

Ile 16

bovine chymotrypsinogen
Serine proteinase (subtilisin)

Structure is very different from chymotrypsin (preceding slide)
Convergent evolution

- No common ancestry!
- Protein sequence and structure are very different
- Functional motif can arise leading to similar function
- If analogous protein resulting from divergent evolution takes over function in cell, this is called *non-orthologous displacement*
Evolutionary and functional relationships

Reconstruct evolutionary relationship by establishing a putative **homologous** relationship between sequences (so it is likely they have a common ancestor)

- Based on sequence:
  - Identity (simplest method)
  - Similarity
- Based on other information (e.g., 3D structure)

Functional (causal) relationship:  
**Sequence** → **Structure** → **Function**

**homology ≡ common ancestry**
Searching for similarities

**Homology (common ancestry)** makes it more likely that genes share the same structure and function

**Homology**: sharing a common ancestor
– a binary property (yes/no)
– it is a nice tool:
When (an unknown) gene X is *homologous* to (a known) gene G it means that we gain a lot of information on X: what we know about G can be transferred to X as a good suggestion.
Searching for similarities

- The main question: what is the function of the new gene?
- The “lazy” investigation without doing experiments:
  - Find a set of similar proteins
  - Identify similarities and differences
  - For long proteins it is often good to identify domains first and then compare the corresponding (sub)sequences separately
Inferring homology from similarity

• Homology: sharing a common ancestor
  – a binary property (yes/no)

• Common ancestry makes it more likely that genes share the same function
  – It’s a nice tool:
    When (a known gene) $G$ is *homologous* to (an unknown gene) $X$, we gain a lot of information on $X$ by transferring what we know about $G$
Can we just transfer information about structure and/or function?

• Structure (and function) more conserved than sequence

• Sequence -> structure -> function

• So, if the sequences already tell us it’s the same thing (homolog), then certainly the structures and functions are supposed to be the same.

• This works most of the time, but there are cases where likely homology does not bear out.
What function does your gene have

- We are going to use the homology principle
- We are going to seriously search through sequence databases
  - Non-redundant (NR) database > 4.5 million sequences
  - Each and every sequence should be considered
Sequence searching - challenges

- Exponential growth of databases
Sequence searching - challenges

- Exponential growth of databases
PRALINE web-interface

**PRALINE multiple sequence alignment**

Paste your sequences in FASTA format (MAX 500 sequences, length 2000):

Or Upload a FASTA file (MAX 500 sequences, length 2000):

Enter a name for your job:

**Options**

Exchange weights matrix: BLOSUM62

Associated gap penalties:

- [ ] Help
- [ ] 12 Open
- [ ] 1 Extension

Global progressive alignment strategy:
- [ ] Standard progressive strategy
- [ ] Pre-profile local processing

Probe:
- [ ] No

Score Cutoff:
- [ ] 0

PS-BLAST pre-profile processing (homology-extended alignment) (new option)

PS-BLAST iterations:

Start e-value cutoff:

Secondary structure prediction:

- [ ] No

DSSP defined secondary structure search:

Tree representation of the final alignment:

Customize alignment representation colours:

Final alignment file format:

E-mail:

If you would like to be notified when your job has completed, please tick the box below and enter the e-mail address the notification should be sent to.

Submit

Interface written by Victor A. Semevski and Jaap Herings
Frequently used format to describe protein sequences:

**Fasta Format**

- **Sequence start indicator**
- **Sequence name**
- **Sequence**

Fasta files can contain many sequences starting with a ‘>’ symbol
Alignments are useful ...

Conserved patterns
Evolutionary analysis
Structure prediction
Motifs
Function prediction
A protein sequence alignment
MSTGAVLIY--TSILIKECHAMPAGNE------
---GGILLFHRTHELIKESHAMANDEGGSNNS
  *  *    *  **** ***

A DNA sequence alignment
attcgtttggcaaatcgccccctatccggccttaa
att---tggcggatcg-cctctacgggccc----
 ***   ****  **** **    ******

Alignment should only be applied to (putative) homologous sequences!! All sequences are supposed to derive from a common ancestor. Ideally, an orthologous set of sequences gets aligned.
1970 **Needleman-Wunschch global pair-wise alignment**


1981 **Smith-Waterman local pair-wise alignment**

How many pair-wise alignments

\[
\begin{array}{cccccccc}
T & D & W & V & T & A & L & K \\
T & D & W & L & - & - & I & K
\end{array}
\]

Combinatorial explosion

- 1 gap in 1 sequence: \( n+1 \) possibilities
- 2 gaps in 1 sequence: \((n+1)n\)
- 3 gaps in 1 sequence: \((n+1)n(n-1)\), etc.

\[
\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \sim \frac{2^{2n}}{\sqrt{\pi n}}
\]

2 sequences of 300 a.a.: \(~10^{88}\) alignments
2 sequences of 1000 a.a.: \(~10^{600}\) alignments!
Technique to overcome the combinatorial explosion: Dynamic Programming (DP)

- Break alignment problem up in smaller subproblems and solve these iteratively
- Alignment is simulated as a Markov process, all sequence positions are seen as independent (i.i.d)
- Chances of sequence events are independent
  - Therefore, probabilities per aligned position are multiplied
  - Amino acid matrices contain so-called log-odds values ($\log_{10}$ of the probabilities), so probabilities can be summed [$\log(ab)=\log(a)+\log(b)$]
Pairwise sequence alignment
Global dynamic programming (DP)

**Search matrix**

MDAGSTVILCFVG

MDAAST

ILC

GS

**Residue Exchange Matrix**

**Gap penalties** (open, extension)

Evolution

MDAGSTVILCFVG
Substitution Matrices: DNA

define a score for match/mismatch of letters

Simple:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Used in genome alignments:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>91</td>
<td>-114</td>
<td>-31</td>
<td>-123</td>
</tr>
<tr>
<td>C</td>
<td>-114</td>
<td>100</td>
<td>-125</td>
<td>-31</td>
</tr>
<tr>
<td>G</td>
<td>-31</td>
<td>-125</td>
<td>100</td>
<td>-114</td>
</tr>
<tr>
<td>T</td>
<td>-123</td>
<td>-31</td>
<td>-114</td>
<td>91</td>
</tr>
</tbody>
</table>
Amino acids are not equal:
1. Some are similar and easily substituted:
   • biochemical properties
   • structure
2. Some mutations occur more often due to similar codons

The two above give us substitution matrices

http://www.cimr.cam.ac.uk/links/codon.htm
http://www.people.virginia.edu/~rjh9u/aminacid.html
Amino acid exchange matrices

How do we get one?

And how do we get associated gap penalties?

• First systematic method to derive a.a. exchange matrices by Margaret Dayhoff et al. (1968) – *Atlas of Protein Structure*
### PAM250 matrix (Dayhoff)

#### amino acid exchange matrix (log odds)

Positive exchange values denote mutations that are more likely than randomly expected, while negative numbers correspond to avoided mutations compared to the randomly expected situation.

<table>
<thead>
<tr>
<th>A</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>-2</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>-2</td>
</tr>
<tr>
<td>Q</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
</tr>
<tr>
<td>L</td>
<td>-2</td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
</tr>
<tr>
<td>F</td>
<td>-4</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>-6</td>
</tr>
<tr>
<td>Y</td>
<td>-3</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
</tr>
</tbody>
</table>

| A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
### PAM250 matrix (Dayhoff)

#### Amino acid exchange matrix (log odds)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
<th>K</th>
<th>M</th>
<th>F</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>Y</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-2</td>
<td>-4</td>
<td>-4</td>
<td>-5</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>-5</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-3</td>
<td>3</td>
<td>1</td>
<td>-2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-5</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-5</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-4</td>
<td>-4</td>
<td>-6</td>
<td>-4</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-5</td>
<td>0</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>-6</td>
<td>2</td>
<td>-4</td>
<td>-7</td>
<td>-8</td>
<td>-5</td>
<td>-7</td>
<td>-7</td>
<td>-3</td>
<td>-5</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>0</td>
<td>-6</td>
<td>-2</td>
<td>-5</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>-3</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>4</td>
<td>2</td>
<td>-2</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-6</td>
<td>-2</td>
<td>4</td>
</tr>
</tbody>
</table>

Positive exchange values denote mutations that are more likely than randomly expected, while negative numbers correspond to avoided mutations compared to the randomly expected situation.
# BLOSUM 62 substitution matrix

(Henikoff & Henikoff, PNAS 89:10915; 1993)

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

http://www.carverlab.org/testing/epp.html
A number of different schemes have been developed to compile residue exchange matrices.

However, there are no formal concepts to calculate corresponding gap penalties.

Empirically determined values are therefore recommended; e.g. PAM250, BLOSUM62, etc.
M = BLOSUM62, P_o = 0, P_e = 0
$M = \text{BLOSUM62}, P_o = 12, P_e = 1$
M = BLOSUM62, $P_o = 60$, $P_e = 5$
Dynamic programming

Scoring alignments

– Substitution (or match/mismatch)

• DNA

• proteins

– Gap penalty

• Linear: \( gp(k) = \alpha k \)

• Affine: \( gp(k) = \beta + \alpha k \)

• Concave, e.g.: \( gp(k) = \log(k) \)

The score of an alignment is the sum of the scores over all alignment columns

\[
S_{a,b} = \sum_i s(a_i, b_i) - \sum_k N_k \cdot gp(k)
\]
Dynamic programming

Scoring alignments

Score: \( s(T,T) + s(D,D) + s(W,W) + s(V,L) - P_o - P_x + s(L,I) + s(K,K) \)
Pairwise sequence alignment

Global dynamic programming

MDAGSTVILCFVG

M
D
A
A
S
T
I
L
C
G
S

Amino Acid Exchange Matrix

Search matrix

MDAGSTVILCFVG-
MDAAST-ILC--GS

Evolution

Gap penalties (open, extension)
DP is a two-step process

- **Forward step**: calculate scores
- **Trace back**: for global alignment, start at bottom-right cell and reconstruct the path leading to that cell
  - These two steps lead to the highest scoring alignment (the optimal alignment)
  - **This is guaranteed when you use DP!**
Dynamic Programming

\[ H(i,j) = \max \begin{cases} 
H(i-1,j-1) + S(i,j) \\
H(i-1,j) - g \\
H(i,j-1) - g 
\end{cases} \]

This is a recursive formula.
There are three kinds of alignments

- Global alignment (preceding slides)
- Semi-global alignment
- Local alignment
Variation on global alignment

- **Global** alignment: uses all letters from both sequences, all gaps are penalised
  
  CAGC
  
  CAGCACTTTGGATTCTCGG
  
  CAGC-----G−T-----GG

- **Semi-global** alignment: uses all letters but does not penalise for end gaps
  
  CAGC
  
  CAGCA−CTTGGATTCTCGG
  
  CAGCGTGG----------
Semi-global alignment

Internal-gaps

MSTGAVLIY--TS-----

---GGILLFHRGTSN

End-gaps

End-gaps
Semi-global alignment

Applications of *semi-global*:
- Finding a gene in genome
- Placing marker onto a chromosome
- One sequence much longer than the other

*Risk*: if gap penalties high -- really bad alignments for divergent sequences

Protein sequences have N- and C-terminal amino acids that are often small and hydrophilic
Local alignment

Result: alignment of the two highest scoring subsequences from either sequence

Applications of *local alignment*:

– aligning multi-domain proteins
– aligning a sequence against a sequence database
– in general when you suspect that homology only holds for parts of one (or both) sequences
Local dynamic programming
(Smith & Waterman, 1981)

LCFVMLAGSTVIVGTR

E D A S T I L C G S

Search matrix

AGSTVIVG
A - STILCG

Negative numbers

Amino Acid Exchange Matrix

Gap penalties
(open, extension)
Measuring Similarity

- **Sequence identity** (number of identical exchanges per unit length)
- **Raw alignment score**
- **Sequence similarity** (alignment score normalised to a maximum possible)
- **Alignment score normalised** to a randomly expected situation (database/homology searching)
What can sequence alignment tell us about structure

HSSP Sander & Schneider, 1991

Sequence identity implies structural similarity!

Don't know region

≥30% sequence identity
Global or Local Pairwise alignment
A multiple sequence alignment is an alignment of three or more sequences.
Multiple sequence alignment

Why?

- It is the most important means to assess relatedness of a set of sequences
- Gain information about the structure/function of a query sequence (conservation patterns)
- Construct a phylogenetic tree
- Putting together a set of sequenced fragments (Fragment assembly)
- Many bioinformatics methods depend on it (e.g. secondary/tertiary structure prediction)
Progressive multiple alignment

Guide tree

Multiple alignment

Scores
Similarity matrix

Scores to distances
Iteration possibilities

5×5

Scores 1-2
Score 1-3
Score 4-5
Progressive alignment strategy

All individual pairwise alignment and construction of distance matrix

Calculating a guide tree; C & D the closest pair; A & B the next closest pair

Figure adapted from Xiong, J. “Essential Bioinformatics”
Progressive alignment strategy

Methods:

- Biopat (Hogeweg and Hesper 1984 -- first integrated method ever)
- MULTAL (Taylor 1987)
- DIALIGN (1&2, Morgenstern 1996)
- PRRP (Gotoh 1996)
- ClustalW (Thompson et al 1994)
- PRALINE (Heringa 1999)
- T-Coffee (Notredame et al. 2000)
- POA (Lee 2002)
- MAFFT (Katoh, Misawa, Kuma, Miyata 2002)
- MUSCLE (Edgar 2004)
- PROBSCONS (Do, 2005)
PRALINE alignment (http://www.ibi.vu.nl/programs/pralinewww/)

The current color scheme of the alignment is by similarity. You can also view the alignment by:
- Conservation
- Hydrophobility
- Reliability
- Sec. Structure

The colour assignments have been adopted from the defaults in CLUSTAL (Thompson et al., 1993) default.

CLUSTAL X (1.64b) multiple sequence alignment Flavodoxin-cheY

1fx1     -PKALIVYGSTGTGNTYTAEATLQRANAG-Y-EDVSQDAASVEAGLFGFELVLLGCSTWGQDSIE--LQDFIPPLFD-SLEETQAGQRK
FLAV_DESVH MPKALIVYGSTGTGNTYTEATLADERAG-Y-EDVSQDAASVEAGLFGFELVLLGCSTWGQDSIE--LQDFIPPLFD-SLEETQAGQRK
FLAV_DESGI MPKALIVYGSTGTGNTVEALNSEG-M-ETTVVNVADTVPALGELGTELGCSTWGQDSIE--LQEDFVPLYE-LLRAGLKDVC
FLAV_DESSA MSKVLIVFGSSTGNTETAEAYVAEPAFNE-E-I-DVEKNTVDVSDDNLGNTLMLEELGCGSTWGQGIEE--LQDFIPPLFD-SLEDINKGK
FLAV_DESE MSKVLIVFGSSTGNTESIATQKLELIAG-H-ETVLNAADASARAYDYGAVLFGSNAWGMDL--MQDDFLSE-ENFRFLAGQRK
FLAV_CLOAB -MKISILYSSKTKTGRVAKLIEEGYKVSGKRI-ETVKMTNLAVDKFLQE-SEGIFGTPYVAN--I-SWEMKWIID-ESSEFWLEGLK
FLAV_MEGEL -MVEIVYWGNTGEAAMEVAAK-A-DVESFRFDTNVDVASS-KDIVILGCPMAXGSE-E--LEDVVPEAFF-TDLAPKLGK
4fxn     --MKIVYWSGTGTENKAEIAKIIEGK--K-DVNTINSVDNSIDELLD-EDILLSAMGDE-V--LEESIEFPI-EIISTKISGK
FLAV_ANASP SKKIGLFYGTQGKESVEAIIRDEFGDNVT-----LHDVQAEVTDLND-YQYLIQCPWNIGELQ--SD------WEGLYS-ELDDVDFNKL
FLAV_AZOVI -AKIGLFFGSNTGKTRKAVSIKKRFDEMTSD--ALNVRSAEFDAQ-YQFLILGTPTLGEGELPGLSSCENESWFEFLP-KIEGLDFGKT
2fcr     --KIGIFSTSTGTNTEVADFIGTKLGAADAP--IDVDDVTDPQALKD-YDLLFLGAPTVNTGADTSGT-SWDFELDKLPEVMDKLP
FLAV_ECOLI --AIGIFGFFSNTGKTGRKAVKLSKIERFDEMTSD--VHDIASKEKLEDAY-DILLIGIPSTYHGEA-CD------WDDFFP-TLEEIDFNKL
3chy     --ADKELKFLVYDDSTMRIRYRNLLKELG------FNNVEEAEDGVDA--I------SDWNMPNMDG-LELLKTR--

1fx1        VACFPCGDSHYEYF--CGAVDAIEELKNLGAIEVQD-----------LRIDGDPRAADDIVGAHVDVRAI-----------------
FLAV_DESVH VACFPCGDSHYEYF--CGAVDAIEELKNLGAIEVQD-----------LRIDGDPRAADDIVGAHVDVRAI-----------------
FLAV_DESGI VGVFCGDSSTTFYF--CGAVDVKREAMGRGLTASS--LKDGEPOSAE-VDLWAREVLRV-------------------------
FLAV_DESSA VSVFPCGDSSTTFYF--CGAVDAIRELEKSGAVVIGDS--LKDGDPERDE-IVWGGIGADKI-----------------------
FLAV_DESE VVAFASPQGEYEFH--CGVAPAIREKGDGIATIAEAG---------LMKMEGADSNDPEEAVASFADVLQK-----------------
FLAV_CLOAB GAAFSANTASIAGGS--DIALLTLNLMDVMSGVLVSGA-----FGPKTHLGYHINEIQENEDARNFRGERIANVQXI------
FLAV_MEGEL VGLFPGSYGWSCEG--WMWAKQRTEDTATIGVTA----------VIN-EMPNAPEKE-LGAAAAKA-----------------
4fxn        VALFGSVEGWGDGK--WNRDFPEEMGNYCVGVTFE--LQMPNEPEEAEDCIFEGKKN-------------------------
FLAV_ANASP VAYFTGDQIGYADNFQDAIGLEKISQRGKGTKYVEWHDIGPNDSKALRG-NGKFVGLALSEDQNQSDLTDDRIKSVAQKLSEFGL--
FLAV_AZOVI VALFPGDLQVPENYALDAEGLYSFPPKDGRAKIVGVSSTDGEYFESSEAVV-DGKFVGLALLDNQGKTDRVAAAMLAQIAEPFLGSL--
2fcr        VAIYFLGDVAEQYFPNFCDAEIDHNRCFQAKAKPVGFSPNPDYDYEESKSVR-DGKFVGLALPLDVMNDQIPMEKRWAVGEAIVVSETG-
FLAV_ECOLI VALFPGDLQNLGSNFSAMRLYDLVIRAGACGVCWNPQEREYKFSFSALENNEFVGLLQDENYQDLTEERIDSWKELKPAVL--
FLAV_CLOAB VALFPGCQEDYAELYFCDALGRTIDIEFPGATIVGVWMQYHFEASKDADDVFGLALTIDQPELTAMERKVWQKQIEELHLDEILNA
3chy        AD--GAMSALPVY-----MVTAEAKKENIIAQAQGAS--GYV-VKPFDDAATLEELKLNKEFLG--------

The secondary structures of 4 sequences are known and can be used to assess the alignment (red is β-strand, blue is α-helix)
Flavodoxin-che Y: Praline (prepro≥1500)

1fx1

TGNT-EYTAETARQLANAG-YEVDSRDAASVEAGGLFGDFDVLLLGCSTWGDSDI------ELQDDFIPLF-DSLEETGAQGRKVACF
FLAV_DESDE

MKVLHVGFGSTGNT-ES1aQLLELIAAGG-HEVTLNAADASEANLADYDVALFgCSAWMEDL------EMQDDFLSLF-EEFNRFGLAGRKVAAf
FLAV_DESVH

MPKALIVYGFGSTGNT-EYTaETIRELADAG-YEVDSRDAASVEAGGLFGDFDVLLlGCSTWGDSDI------ELQDDFIPLF-DSLEETGAQGRKVACF
FLAV_DESSA

MSKALIVYGFGSTGNT-ETAaEYVAEFAENKE-IDVLEKNDTVDSVADLNGYDILFgCSTWGDGEI------ELQDDFIPLF-DSLENADLKGGVSVf
FLAV_DESIGI

MPKALIVYGFGSTGNT-EGVaEAIAKTLNSEG-METTVNVAADVTAGLAEYGDVLLlGCSTWGDDEI------ELQEDFVPL-ELDLRAKLKDKKVf
2fcr

--GKIFFGSTNT-TEVDIFGKTLGA------KADAPAIVDVTDDQALKDYDDLFLGAPTWNTG------ADTERGTSWSDEFLYDKLPEVMDKDLPLVAIF
FLAV_AZOVIO

-ALKILEFGSNTGTK-RKvaSSIKKFRFDET-MSAD-LNVNRV-ÆEDFAQYQFLfIgTPTLELGELPGLSADCENESWEEP-PIEIGLDSGKTVALf
FLAV_ENTAG

MATIGIFFGSDTQTG-RKvaKLHQQKLDG------IADAPLDRRAT-REQLSYPVLLlGTPTLGDGELPGVEAGSßYDSWQEXT-NTLSEADLTGKVALf
FLAV_ANASP

SKKIGLYFTGQTKG-ESaVaEIEIDFGN------DDVTLHDSQAE-VTDMLNQYLIgiCPWNIGEL------QSDWEGLY-SELDDVDFNGKLVAYf
FLAV_ECOLI

-ATIGIFFGSDNTGNT-ENiaKMIQKQLG------DVADVHDIAKSS--KEDLEADILLlGIPTWYEG------AQCDWDFFP-TELIEIFNGKVLAlf
4fxn

-MK--IVYWStGDTGN-EKMEIAELAEIGS-VDNTINVSDVNIDEL-NEDILICGSAEMDEVL------ESEEFPEI-EEIS-TKIGSKKVALF
FLAV_MECEL

MVE--IVYWStGDTGN-EMaNEIAEAAVKAAG-ADVESRFEDTNVDVSA-SKDVLlGCpmGeEEL------EDSVEPF-TDA-PLKLGGKVelf
FLAV_CLOAB

-MKISILYSSKSTG-ERVaKLIEEGVrKNVNEKTMNDVAD-KKLQFSEGIIIFgTPTYfAN------ISWEMKKWI-DESSEFNLEKGLf
3chy

ADKEKLFLVVDFFSTMIRIVNLKLGFN--NVVEAEYGVDAINLQAGGYGFVf--SDWMPNF----------DGELEL-KTIRADVAMSALVf

1fx1

GCGDS-SY-EYFCGA-VDAIEEKLBKNLNGAEIVQD------------------------GLRIDG-----PRAARDIVGAHVDVRAI------
FLAV_DESDE

ASQGD-SY-EHFCGA-VPAIÈREKAgelTIAlAE------------------------GLKMgdG-ASNDPEVASCADVLKQL-----
FLAV_DESVH

GCGDS-SY-EYFCGA-VDAIEEKLKNLgAEIVQD------------------------GLRIDG-----PRAARDIVGAHVDVRAIG-----
FLAV_DESSA

GCGDS-SY-TYFCGA-VDAIEEKLKMgAVVfGD------------------------SLKIDGD---PE--REILISWSGSIADK-----
FLAV_DESIGI

GCGDS-SY-TYFCGA-VDVIÈKEADgATlVAS------------------------SLKIDGD--PE--SAEVLDwARVLRVA-----
2fcr

GLGDAEGYPDNFCDA-IEEIHDCFAKQGAKPVGFSNPDYDYEESKS-VRDKFlGFLGPlDMVNDQPfMEKRVAGWVEAVVSfGV------
FLAV_AZOVIO

GLGqVqVYPENYLDa-LEGlYSFfKDRgAKIVGWSDTDGyEFSESSA--VVDGKFLGVAldLDDQSSGKTDERVAAwLaAAIPEFGLS-L--
FLAV_ENTAG

GLGqLqYNKSNFvSA-MRsILYDLVLaRqACVGVNfREGYFKSFSALENNNEVFGFLQDfEYDfLEfTIERSwLEKLPKAV-L----------
FLAV_ANASP

GTQDGqYgYADNFQDA-IGILÈÈKSIQgRGTVSTGITDSGFDNSQ--LNGKFVLGALDgEQDSTDfDRIgSWaQIgLKSfEGL-----
FLAV_ECOLE

GCGqEDAYAENFCDA-GLTIRfDIEPrGATIVGHWPTTAGYFEASKGLADDHDFVGLAILADERDQPfELTAERVKeWgKqiSeELfHDLfLaNA
4fxn

G------SY-GWDgGKrMRDPFEERMNGYGVVVVf-----PLIVQNE--PDEAEQDCEIFGKfRIA------
FLAV_MEGEL

G------SY-GWGSgGWNDAfKQRTEDfTGATVfG------AIVNEM--PDNA-PCKEGfGAAK-----
FLAV_CLoAB

STANISAGSDIA----LTLTLNMLVMgMVLVSg---gVAFgPKTHLGyVHINEqIqnedARIRFrgerIAnKQfIr---
3chy

VTAEfAK---ENIIAaA--AQAGAS---GYVV------KFTPAALEfLEKfIKfEGKLM---
Flavodoxin-cheY: Praline (locprepro≥300)

1fx1

--PKALIVYGGTSTGNTETAIATIRAGLNAGYEVDSRDAASEAVAGLELGFDVLVGLGCSTWGDDASI------ELQDDFIPL--FDSLEETGAGGRKVACF
FLAV_DESVH
-MPKALIVYGGTSTGNTETaETIARELADAYEVSDSAASEAVAGLELGFDVLVGLGCSTWGDDASI------ELQDDFIPL--FDSLEETGAGGRKVACF
FLAV_DESSA
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF
FLAV_DESGI
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF
FLAV_DESCD
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF
FLAV_MEGEL
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF
FLAV_DESSA
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF
FLAV_DESVH
-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF

4fxn

-MSKSLIVYGGTSTGNTETaATVAYAEKENEDVELKVNTDVSADLNGYDVLFGCSTWGEESI------ELQDDFIPL--YDSLKENALGKGGVVSF

2fcr

---KIGIFFFSTGNTTEVADFDIGKTLGAGADAPI--DDVDDVTDFQALKDYGLLLLGAPVTNTGAD----TERSGTSWDEFL-YDLKFEVDMKDLPLVAIF
FLAV_ANASP
-SSKIGILEFYGTGTTKESVaTTRIRFEDGNDVTVLHL--DVSQAEV-TLDNDYQYLLIGCPTWNIGEL----QSDWEGL-YSELDDVFNGKLVAYF
FLAV_AZOVl
-AKIGLFFGSNTGKTRKvaKSIKRRFDETSMDA-LNNRNSA-EDFAQQYQFLIlGTPGTGEELPGLSSCENESWEF----LPKIEGLDFSGKTVALF
FLAV_ENTAG
-MATIGIGFFSDETDQTRKvaKHLHQKLDG-IAADAPLDRRARAT-EQFLSFLYVLLGTPTLGDELPGVEAGSQDWSQEF--TNLSEADLTGKTVALF
FLAV_ECOLI
-AITGIGFFSDETDGNTENVaAIIQKQLKGDVADVH--DIAGSSK-EDLAEYDILLQIPTWVGEA----QCDWDFD--FPTLEEIDFNKLVALF
FLAV_CLOAB
-MKISILYSSKTGKTERVaKLIIEGVRKSISSKTMNLDAVDKKFLQSESEGIIfTPTYYA----NISWEMKKIDESFENLLEGKLAA

3chy

ADKELKFLVVDDDFSTMRRIRVNLLNELGPFTNVVEEAEDGVDOALNLQ-AGGGYFVI-SDWNMPNM--DGLEL--LKTIRADGAMSALMVLM

1fx1

GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_DESVH
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_DESSA
GGCGDS--SY-TYFGCA-VD--AIEEKlKMGAVVIGD---------------------SLKDID--GDPE--RDEIVSWSGIGAKIK1--
FLAV_DESGI
GGCGDS--SY-TYFGCA-VD--VIEEKlAEtTLAVS----------------------SLKDID--GDPE--SAEVLDwAREVLARYV1--
FLAV_DESCD
GGCGDS--SY-TYFGCA-VD--VIEEKlAEtTLAVS----------------------SLKDID--GDPE--SAEVLDwAREVLARYV1--
FLAV_MEGEL
GGCGDS--SY-TYFGCA-VD--VIEEKlAEtTLAVS----------------------SLKDID--GDPE--SAEVLDwAREVLARYV1--
FLAV_ANASP
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_AZOVl
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_ENTAG
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_ECOLI
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--
FLAV_CLOAB
GGCGDS--SY-EYFGCA-VD--AIEEKlKNLGAEIVQD---------------------GLRID--GDPREARDDIVGHAVDVRGA1--

4fxn

--AAKQRURDVTGGATVIG-----AI--VNE--EMPDNA-PACKETEAAKA--

2fcr

GLCDAE-GBPNDNCDA-IE--EIHDCFAKQGAKPVFGSNPDYDYEESKSVRD-GKFGLPLLDMVNDQIPMEKRVAQGWEAVEVSETGV--
FLAV_ANASP
GTGDQ1-GYADNFQDA-IG--ILEEISQRgKTVGYWSTDGYFNDSDKALRN-KGFVGLALDDNQSLDTDRIKSWAYAQKLEFGLG
FLAV_AZOVl
GLGDQV-GYPENYLD3A-LG--ELYSFFPKDrAgAKIVGSWSTDGYFSESEAVVD-KGFVGLALDDNQSLKTDERVAAwLAQIAPEFGLS--
FLAV_ENTAG
GLGDQV-GYPENYLD3A-LG--ELYSFFPKDrAgAKIVGSWSTDGYFSESEAVVD-KGFVGLALDDNQSLKTDERVAAwLAQIAPEFGLS--
FLAV_ECOLI
GLGDQV-GYPENYLD3A-LG--ELYSFFPKDrAgAKIVGSWSTDGYFSESEAVVD-KGFVGLALDDNQSLKTDERVAAwLAQIAPEFGLS--
FLAV_CLOAB
GLGDQV-GYPENYLD3A-LG--ELYSFFPKDrAgAKIVGSWSTDGYFSESEAVVD-KGFVGLALDDNQSLKTDERVAAwLAQIAPEFGLS--

3chy

VTAEAKKENIIAA-------AQAGAAS---------------------GYVVK------PFTAATLEELNKIFKELGM--
The BLAST suite

- Computer program for homology searching

- Given a protein query sequence (for which the function is unknown), the program searches through a non-redundant sequence database (NR) of >4 million sequences

- BLAST aligns a given query sequence with each database sequence and calculates similarity

- If sequence similarity is high enough (low BLAST e-value), the query and database sequence are deemed homologous, so that the database sequence’s function (if known) can be transferred to the query.

- BLAST is a fast heuristic local alignment tool
BLAST entry page

1. Paste your query sequence
2. Choose the BLAST program you want
1 - This portion of each description links to the sequence record for a particular hit.

2 - Score or bit score is a value calculated from the number of gaps and substitutions associated with each aligned sequence. The higher the score, the more significant the alignment. Each score links to the corresponding pairwise alignment between query sequence and hit sequence (also referred to as subject or target sequence).

3 - E Value (Expect Value) describes the likelihood that a sequence with a similar score will occur in the database by chance. The smaller the E Value, the more significant the alignment. For example, the first alignment has a very low E value of $e^{-117}$ meaning that a sequence with a similar score is very unlikely to occur simply by chance.

4 - These links provide the user with direct access from BLAST results to related entries in other databases. ‘L’ links to LocusLink records and ‘S’ links to structure records in NCBI's Molecular Modeling DataBase.
X’ residues denote low-complexity sequence fragments that are ignored
BLAST ‘flavours’

- **blastp** compares an amino acid query sequence against a protein sequence database.
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx** compares the six-frame conceptual protein translation products of a nucleotide query sequence against a protein sequence database.
- **tblastn** compares a protein query sequence against a nucleotide sequence database translated in six reading frames.
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
PSI (Position Specific Iterated)
BLAST

• basic idea
  – use results from BLAST query to construct
    a profile matrix
    • a profile matrix is a table that corresponds to a multiple
      alignment (or a BLAST master-slave (many-to-1)
      alignment) and represents the likelihood of each amino
      acid type at a given alignment position
  – search database with profile instead of
    query sequence
• iterate
A profile is a $L$ (number of aligned positions) by 20 matrix holding the propensities for each of the 20 possible amino acids at each position.

- Propensities can be simple frequencies encountered in the alignment but can also be calculated in a more complex way.
- The PSI-BLAST profiles are referred to as Position-Specific Scoring Matrix (PSSM)
PSI-BLAST iteration

iterate

Query sequence

Gapped BLAST search

Database hits

PSSM

iterate

Query sequence

Gapped BLAST search

Database hits

PSSM
Another PSI-BLAST iteration graphic...

Run query sequence against database

Run PSSM against database

Q

______________

hits

T

Discarded sequences

DB
PSI-BLAST steps in words

• Query sequences are first scanned for the presence of so-called *low-complexity regions* (Wooton and Federhen, 1996), *i.e.* regions with a biased composition (“ACACACACACA...”) that are likely to lead to spurious hits; are excluded from alignment.

• The program then initially operates on a single query sequence by performing a gapped BLAST search.

• Then, the program takes significant local alignments (hits) found, constructs a multiple alignment (master-slave alignment) and calculates a position-specific scoring matrix (PSSM) from this alignment.

• PSI/BLAST then rescans the database in a subsequent round, using the PSSM, to find more homologous sequences. Iteration continues until user decides to stop or search has converged.
During iteration, new hits can come in and hits can drop out of the hit-list.

At each iteration a new profile is made of the master-slave alignment.
Sequence searching

QUERY

DATABASE

POSITIVES

NEGATIVES

True Positive

False Positive

True Negative

False Negative
The PFAM Database

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures
- Search with Hidden Markov Model (HMM) for each alignment
The PFAM Database

• Pfam is a database of two parts, the first is the curated part of Pfam containing about 9000 protein families (Pfam-A). Pfam-A comprises *manually crafted* multiple alignments and profile-HMMs.

• To give Pfam a more comprehensive coverage of known proteins *automatically* a supplement called Pfam-B is generated. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A.

• Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.
The PFAM Database

Sequence coverage Pfam-A : 74% (Yellow)
Sequence coverage Pfam-B : 13% (Blue)
Other (Grey)

74% of proteins have at least one match with Pfam.

Version 21.0 - November 2006: Pfam-A contains 8957 families
Clan pages in Pfam. (A) A screen shot of a clan summary page, containing the description, annotation and membership of the clan. From this page, the user can view the family relationship diagram (B). Each family in the clan is represented by a blue box and its relationship to other families is represented by solid lines (significant profile–profile comparison score) or dashed lines (non-significant profile–profile comparison score). Beside each line, the profile–profile comparison E-value score is presented. This score is also linked to a visualization of the profile–profile comparison alignment (C). The clan summary page also provides a link to the clan alignment (D). The clan alignment is a multiple sequence alignment of all of the clan members seed alignments (each set of seed sequences are separated by the alternate background shading). The alignments are coloured using Jalview.
Wrapping up

• Facts about DNA
• DNA deleterious effects leading to disease
• Genome size and number of genes
• DNA, evolution, speciation
• homology, orthology, paralogy
• (non-)synonymous mutation, ka/ks ratio
• Pairwise alignment: global, semi global, local
• Dynamic programming (DP): an algorithm to carry out and score alignments
  – Note: understanding how the algorithm works exactly will be part of the course Algorithms in Sequence Analysis (ASA)
• Multiple sequence alignment (MSA)
  – Examples of alternative alignments for a given sequence set
• Blast, PSI-Blast
• PFAM database
• Note: MSA and PFAM have not been explained during the lecture but are easy to revise using the slides