Computational searches of biological sequences

Special thanks to all the scientists that made public available their presentations throughout the web from where many slides were taken to elaborate this presentation.

The number of Fully Sequenced Genomes has been increasing very rapidly since the first report of the genome of a living organism only fourteen years ago.

- Mycoplasma genitalium (only ~5x10^5 pb)
- Apis mellifera (~2x10^6 pb)

Completely Sequenced Genomes © January 2009

Biología Computacional
ORFs from Complete genomes

Completely sequenced and published microbial genomes

<table>
<thead>
<tr>
<th>Year</th>
<th>No of ORFs in all genomes (incl. ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>350k</td>
</tr>
<tr>
<td>2004</td>
<td>500k</td>
</tr>
<tr>
<td>2005</td>
<td>750k</td>
</tr>
<tr>
<td>2006</td>
<td>1.5Mio</td>
</tr>
</tbody>
</table>

ORFs from Complete genomes vs Metagenomics ORFs

<table>
<thead>
<tr>
<th>Year</th>
<th>No of ORFs in all genomes (incl. ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>350k</td>
</tr>
<tr>
<td>2004</td>
<td>500k</td>
</tr>
<tr>
<td>2005</td>
<td>750k</td>
</tr>
<tr>
<td>2006</td>
<td>1.5Mio</td>
</tr>
<tr>
<td>2007</td>
<td>1.1Mio</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data analysis: the signs before the flood

Microbial genomes published per year

- Animal genomes (>100Mb, published, >95% cov)
- Metagenomics (>50Mb, not focussed, non-16S, published, deposited)

<table>
<thead>
<tr>
<th>Year</th>
<th>Oranges</th>
<th>Deep sea whale bones</th>
<th>Mammoth bones</th>
<th>Human gut Mouse gut Sludge</th>
<th>Global Ocean Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>30</td>
<td>5</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>20</td>
<td>10</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>30</td>
<td>5</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>40</td>
<td>10</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is the interest to sequence so many genomes?

What kind of data can we get from Sequence Comparisons?

Evolution

Some Important dates in history

- Origin of the universe: $10^9 \pm 2$ billion years
- Formation of the solar system: $-4.6 \pm 0.1$ billion years
- First self-replicating system: $-3.5 \pm 0.5$ billion years
- Prokaryotic-eukaryotic divergence: $-1.8 \pm 0.3$ billion years
- Plant-animal divergence: $-1.0$ billion years
- Invertebrate-vertebrate divergence: $-0.5$ billion years
- Mammalian radiation beginning: $-0.1$ billion years

$^a$Billions of years. From Doolittle et al., 1986.
Instead of list of genes we want to understand the metabolic capabilities of the organism.

Sequence conservation

Homology relationship

Functional relationship

Protein homology inferences depends on:

a) Extent of sequence similarity
b) The evolutionary nature of the protein
The divergence causes that the similarity of homologous proteins asymptotically resembles those from random sequences.

The good news. Proteins that share sequence similarity can also share common protein folding, and often share a similar function.

The bad news. Proteins that share sequence similarity might not share common protein folding.

The bad news. High levels of sequence identity do not “guarantee” the same enzymatic function.
**The bad news. In some cases, proteins that share important primary sequence and 3D similarity might not present related functions!!!**

- ArsC: System of arsenic resistance
- Spx: Transcriptional regulator

**Sequence and structural similarity of ArsC y Spx**

---

**Holology and some related concepts**

**Orthologs**: Homolog genes in different genomes that differ by speciation events. Orthologs commonly share the same function.

**Paralogs**: Homolog genes in the same genome that differ by a duplication event. Paralogs might have related but not identical function.

**Xenology**: The relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters.

---

**Walter M. Fitch. TIG 2000:16 (227-231)**

**Homology** is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. Characters can be any genic, structural or behavioral feature of an organism.

**Analogy** is distinguished from homology in that its characters, although similar, have descended convergently from unrelated ancestral characters.

**The cenancestor** is the most recent common ancestor of the taxa being considered.
Ernst Haeckel’s Monophyletic tree of organisms, 1866.

Holology and some related concepts

What does one call the relationship of the remaining two genes? Paralogy, of course.

The definition of the forms of homology does not change by virtue of the known, suspected, or unknown presence of a copy of a gene.

speciation

duplication

Lost of genes

Holology and some related concepts

The gene loss problem

Holology and some related concepts

Are these homologous proteins?
Sequence comparison is just the first step, but it could be very informative.

Different ways to compare biological sequences

1. **Strings** (Pattern matching, Pattern discovery, Pattern representation)
2. **Pairwise comparison**
3. **Structural search** (Threading)
4. **Position Specific Score Matrix PSSM searches**
5. **Hidden Markov Model HMM searches**

Similarity evaluation

The easiest way to determine the similarity of two sequences can be obtained by the evaluation of the number of their coincidences normalized in terms of the shortest sequence.

```
AAGTGAGATGCTTAG Length 17
 AGAGACTGA Length 10
8 identical residues \(\Rightarrow\) Similarity of 80%
```
In case of protein sequences, the similarity can be evaluated using different Scoring matrices.

A matrix with an evolutionary distance of 1 PAM would have numbers close to one on the main diagonal and small numbers off the main diagonal.

One PAM would correspond to roughly 1% divergence in a protein (one amino acid replacement per hundred).

The model of evolution that Dayhoff used assumed that proteins diverged as a result of accumulated, uncorrelated mutations.

PAM (Point Accepted Mutation)

It assumes that the rate of evolution in conserved proteins can be extrapolated to a less conserved families of proteins.

The elements of this matrix give the probability that the amino acid in one column will be replaced by the amino acid in some row after a given evolutionary interval. For example, a matrix with an evolutionary distance of 0 (zero) PAMs would have ones on the main diagonal and zeros elsewhere.

A matrix with an evolutionary distance of 1 PAM would have numbers close to one on the main diagonal and small numbers off the main diagonal.

They treat the PAM-1 matrix as a first order Markov chain transition model. To derive a mutational probability matrix for a protein sequence that has undergone N percent accepted mutations, a PAM-N matrix, the PAM-1 matrix is multiplied by itself N times. This results in a family of scoring matrices.

Log-odds = log of the conditional probability of one amino acid substitution,
= log(P-observed(Y|X)/P-expected(X|Y)) = PAM 1
The evolutionary rate of proteins may importantly differ.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Rate</th>
<th>Protein</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filamin</td>
<td>90</td>
<td>Thymosin beta 1</td>
<td>7.4</td>
</tr>
<tr>
<td>Growth hormone</td>
<td>37</td>
<td>Parathyroid</td>
<td>7.1</td>
</tr>
<tr>
<td>Ig kappa chain C region</td>
<td>37</td>
<td>Parvalbumin</td>
<td>7.0</td>
</tr>
<tr>
<td>Kappa casein</td>
<td>33</td>
<td>BPTI Protease inhibitors</td>
<td>6.2</td>
</tr>
<tr>
<td>Ig gamma chain C region</td>
<td>31</td>
<td>Igpepsin</td>
<td>5.9</td>
</tr>
<tr>
<td>Latrope beta chain</td>
<td>30</td>
<td>Metalloprotein beta</td>
<td>5.6</td>
</tr>
<tr>
<td>Ig lambda chain C region</td>
<td>27</td>
<td>Alpha-crystallin A chain</td>
<td>5.0</td>
</tr>
<tr>
<td>Complement C3a</td>
<td>27</td>
<td>Trypsin</td>
<td>4.8</td>
</tr>
<tr>
<td>LacZ</td>
<td>27</td>
<td>Cysteine carboxypeptidase b</td>
<td>4.2</td>
</tr>
<tr>
<td>Epidermal growth factor</td>
<td>36</td>
<td>Insulin</td>
<td>4.4</td>
</tr>
<tr>
<td>Somatostatin</td>
<td>25</td>
<td>Calcitonin</td>
<td>4.3</td>
</tr>
<tr>
<td>Pancreatic ribonuclease</td>
<td>21</td>
<td>Neumyosin 2</td>
<td>3.6</td>
</tr>
<tr>
<td>Lipoprotein beta</td>
<td>21</td>
<td>Pencysain</td>
<td>3.5</td>
</tr>
<tr>
<td>Haptenoglobin alpha chain</td>
<td>20</td>
<td>Lac reductase</td>
<td>3.4</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>19</td>
<td>Adenylate cyclase</td>
<td>3.2</td>
</tr>
<tr>
<td>Phospholipase A2</td>
<td>19</td>
<td>Thymidylate synthase</td>
<td>2.8</td>
</tr>
<tr>
<td>Protease inhibitor PSTI type</td>
<td>18</td>
<td>Vasoactive intestinal peptide</td>
<td>2.6</td>
</tr>
<tr>
<td>Proteinase</td>
<td>17</td>
<td>Cysteintripeptide</td>
<td>2.4</td>
</tr>
<tr>
<td>Paenaeus hemocyanin</td>
<td>17</td>
<td>Glyceraldehyde-3-PhD</td>
<td>2.2</td>
</tr>
<tr>
<td>Carbonic anhydrase C2</td>
<td>15</td>
<td>Cysteine carboxypeptidase</td>
<td>2.2</td>
</tr>
<tr>
<td>Latrope alpha chain</td>
<td>16</td>
<td>Pepsin endopeptidase</td>
<td>1.9</td>
</tr>
<tr>
<td>Hemoglobin alpha chain</td>
<td>12</td>
<td>Collagen</td>
<td>1.7</td>
</tr>
<tr>
<td>Hemoglobin beta chain</td>
<td>12</td>
<td>Topoisomer</td>
<td>1.5</td>
</tr>
<tr>
<td>Glyco-fibronectin A-I</td>
<td>10</td>
<td>Alpha-crystallin B-chain</td>
<td>1.5</td>
</tr>
<tr>
<td>Gastro</td>
<td>9.8</td>
<td>Glutamin</td>
<td>1.2</td>
</tr>
<tr>
<td>Animal lysozyme</td>
<td>9.8</td>
<td>Glutamyl-DH</td>
<td>0.9</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>8.9</td>
<td>Histone H2A</td>
<td>0.9</td>
</tr>
<tr>
<td>Amylase A</td>
<td>8.7</td>
<td>Histone H2A</td>
<td>0.8</td>
</tr>
<tr>
<td>Nerve growth factor</td>
<td>8.5</td>
<td>Histone H4</td>
<td>0.8</td>
</tr>
<tr>
<td>Acid phosphatase</td>
<td>8.4</td>
<td>Ubiquitin</td>
<td>0.1</td>
</tr>
<tr>
<td>Myelin basic protein</td>
<td>7.4</td>
<td>Histone H4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*percent 100 Ms

From (Daw, 1997; Henikoff et al., 1999)

**BLOSUM Matrices**

Steven Henikoff and Jorja G. Henikoff, 1992

**The blocks have no gaps.**

Do not have an evolutionary model.

BLOSUM n, Uses the blocks of sequences that share an identity of n% or more.

It also uses Log-odds.

**Examples:** BLOSUM 62, 50, 30.

**In a pairwise sequence comparison, the value of any pair of amino acids is the same regardless their position in the sequence.**

**Relationships with PAM and BLOSUM Matrices**

**PAM 250**

In a pairwise sequence comparison, the value of any pair of amino acids is the same regardless their position in the sequence.
**Basic sequence comparisons**

**Algorithm:** Move one of the two sequences and count the number of aligned similar residues.

```
TCAGACGATTG   (0)
ATCGGAGCTG
TCAGACGATTG    (1)
ATCGGAGCTG
TCAGACGATTG    (2)
ATCGGAGCTG
TCAGACGATTG    (3)
ATCGGAGCTG
TCAGACGATTG    (4)
ATCGGAGCTG
TCAGACGATTG    (5)
ATCGGAGCTG
TCAGACGATTG    (6)
ATCGGAGCTG
TCAGACGATTG    (7)
ATCGGAGCTG
TCAGACGATTG    (8)
ATCGGAGCTG
TCAGACGATTG    (9)
ATCGGAGCTG
TCAGACGATTG    (10)
ATCGGAGCTG
TCAGACGATTG    (11)
ATCGGAGCTG
```

A similar procedure can be done by counting the matches of diagonals.

```
pos:  1 2 3 4 5 6 7 8 9 10
X:  TCAGACGATTG (n=11)
Y:  ATCGGAGCTG (m=10)
pos:  1234567890
```

The diagonal with the greatest value represents the optimum alignment.

**A dot-plot analysis is an easy and graphical way to analyze two sequence alignment**

**Different kind of sequence alignments**

- Global
- Local
- Heuristic
- Exhaustive
Global or local alignments

Global. Includes all the protein.

Local. Only the zone of major similitude is considered.

Different kind of sequence alignments

Heuristic:
It is a method used to rapidly come to a solution that is hoped (but does not guaranty) to be close to an 'optimal solution'.
It is based on educated guesses, intuitive judgments or simply common sense.

Exhaustive:
Explores all possible solutions to find the one that guaranties to be the optimal.
It is considerably slower than heuristic methods

Similarity evaluation

In some cases, the observed similarity can be improved by the introduction of gaps.

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Length</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAGTGATG</td>
<td>GACTGA</td>
<td>11</td>
<td>36%</td>
</tr>
<tr>
<td>GAAGGACTGA</td>
<td>GAAGGACTGA</td>
<td>17</td>
<td>72%</td>
</tr>
</tbody>
</table>

8 identical residues -> Similarity of 72%

The evaluation of gaps has not been well defined.

Normally, a punishment is used (in the qualification) by the existence of gap, either by opening the sequence, by the length of the gap, or both.

The value of the punishment is still arbitrary.
Sequence alignment consists in the adequate introduction of spaces into the sequence to denote their similarities.

Global alignment by dynamic programming
Needelman-Wunsch

- The value, one, if Aj is the same kind of amino acid as Bi; if they are different amino acids, MATij is assigned the value, zero.
- The sophistication of the comparison is increased if, instead of zero or one, each cell value is made a function of the composition of the proteins.
- A penalty factor, a number subtracted for every gap made.
- The penalty factor could be a function of the size and/or direction of the gap.
- The maximum-match pathway, then, is that pathway for which the sum of the assigned cell values is largest.

Dynamic programming
Is a method of solving complex problems by breaking them down into simpler steps.

Commonly, programming involves formulating a complex calculation as a recursive series of simpler calculations.

Needelman-Wunsch
The cells calculate like in the global alignments, but when a value becomes negative it puts zero. The highest score of the table is the one of the final end of the best local alignment. Walk backward movement until finding a zero to reconstruct the alignment. Sometimes local alignments can be combined if they share neither columns nor lines.

Optimums score = 3

Alignment ABD

Similarity residues +1

Different residues -1

Indels -2

---

**FASTA**


**Step 1:** Identify the regions with the greatest identification number (ktup=1) or pairs of identities (ktup=2)

**Step 2:** Do new search considering similar amino acids (BLOSUM). Shorten the terms to only include the regions that contribute to score maximum. Each of these regions is a partial alignment without gaps.

**Step 3:** Consider if the extension of the ends can join two regions (consideration of mismatches)

**Step 4:** Join several partial alignments within a band (introduction of gaps or indels)

**Step 5:** Statistical evaluation of the best alignments

**Step 6:** Use of exhaustive algorithm Smith-Waterman to find alignment optimal

---

**BLAST**


**Query word W=3**

**Generation of exact BLAST words with a word size of W=3**

---

**Needelman-Wunsch**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>I</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>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Local alignment by dynamic programation**

Smith & Waterman, 1981

**Optimums score**

Alignment ABD

Similarity residues +1

Different residues -1

Indels -2

---
BLAST Basic Local Alignment Search Tool


• Speed is achieved by Pre-indexing the database before the search
• Uses a hash table that contains neighborhood words.

query word (W = 3)

<table>
<thead>
<tr>
<th>neighborhood words</th>
<th>neighborhood score threshold (T = 13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query: GUESTTOSQIALLKMOFPQGKLPIQHSLFIRAVERSAILQTMEL</td>
<td></td>
</tr>
<tr>
<td>PQS 10</td>
<td></td>
</tr>
<tr>
<td>PQS 12</td>
<td></td>
</tr>
<tr>
<td>PQS 14</td>
<td></td>
</tr>
<tr>
<td>PQS 16</td>
<td></td>
</tr>
<tr>
<td>PQS 18</td>
<td></td>
</tr>
<tr>
<td>PQS 20</td>
<td></td>
</tr>
<tr>
<td>PQS 22</td>
<td></td>
</tr>
<tr>
<td>PQS 24</td>
<td></td>
</tr>
<tr>
<td>PQS 26</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
</tbody>
</table>

Review of some basic concepts
Score distribution in a sequence database

A distribution is a plot showing the frequency of a given variable or observation.

The Score distribution in a sequence database is not a Gaussian distribution. It is a extreme value distribution.

The expectancy can be evaluated using the Karlin-Altschul statistics and can combine several compatible “matches”

It is associated to NCBI-GenBank

Null hypothesis

• The null hypothesis describes in a formal way some aspect of the statistical behavior of a set of data and this description is treated as valid unless the actual behavior of the data contradicts this assumption

• The Null hypothesis answer the question what is the probability of observing a value test statistic that is for the at least as extreme as the value that was actually observed
Examples of null hypotheses:

- Sequence comparison using shuffled sequences.
- A normal distribution of log ratios from a microarray experiment.
- LOD scores from genetic linkage analysis when the relevant loci are randomly sprinkled throughout the genome.

The picture shows a distribution of scores from a real database search using BLAST.

This distribution contains scores from non-homologous and homologous pairs.

What should be the null-hypothesis of this analysis?

The p-value is the probability of observing an effect as strong or stronger than you observed, given the null hypothesis.

I.e., “How likely is this effect to occur by chance?”

Pr(x > Simul)

The most common thresholds are 0.01 and 0.05

A threshold of 0.05 means you are 95% sure that the result is significant.

Is 95% enough? It depends upon the cost associated with making a mistake.

Examples of costs:
- Doing expensive wet lab validation.
- Making clinical treatment decisions.
- Misleading the scientific community.

Most sequence analysis uses more stringent thresholds because the p-values are not very accurate.
Database searching

- Say that you search the non-redundant protein database at NCBI, containing roughly one million sequences. What p-value threshold should you use?
- Say that you want to use a conservative p-value of 0.001
- Recall that you would observe such a p-value by chance approximately every 1000 times in a random database.
- A Bonferroni correction would suggest using a p-value threshold of $0.001 / 1,000,000 = 0.000000001 = 10^{-9}$.

E-values

- A p-value is the probability of making a mistake.
- The E-value is the expected number of times that the given score would appear in a random database of the given size.
- One simple way to compute the E-value is to multiply the p-value times the size of the database.
- Thus, for a p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is $0.001 \times 1,000,000 = 1,000$.

- $E$-value < $10^{-100}$
  Identical sequences. You will get long alignments across the entire query and hit sequence.
- $10^{-50} < E$-value $< 10^{-100}$
  Almost identical sequences. A long stretch of the query protein is matched to the database.
- $10^{-10} < E$-value $< 10^{-50}$
  Closely related sequences, could be a domain match or similar.
- $1 < E$-value $< 10^{-6}$
  Could be a true homologue but it is a gray area.
- $E$-value $> 1$
  Proteins are most likely not related
- $E$-value $> 10$
  Hits are most likely junk unless the query sequence is very short.

Summary

- Selecting a significance threshold requires evaluating the cost of making a mistake.
- Bonferroni correction: Divide the desired p-value threshold by the number of statistical tests performed.
- The E-value is the expected number of times that the given score would appear in a random database of the given size.
Different kind of BLAST programs

- 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 a nucleotide query sequence translated in all reading frames against a protein sequence database
- tblastn
  - compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- tblastx
  - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
**Practical consideration in data banks searches**

**Advise #1. Consider all the available biochemical / biological knowledge.**

**TRAP and Anti-TRAP as our initial working model**

Note: The gene that codes for TRAP and anti-TRAP in *B. subtilis* are called *mtrB* and *yczA*, respectively.

**Practical consideration in data banks searches**

**Advise #1. Consider all the available biochemical / biological knowledge.**

**Trp operon regulation in Bacillus subtilis**

- **Advise #1.** Consider all the available biochemical / biological knowledge.
- **Advise #2.** Filtering Low Complexity Regions

- A great percentage of a genomic DNA could be repetitive due to:
  - retrotransposons
  - ALU region
  - microsatellites
  - centromeric sequences, telomeric sequences
  - 5' Untranslated Region of ESTs

**Example of ESTs with simple low complexity regions:**

```
T27311
GGGTGCAGGAATTCGGCACGAGTCTCTCTCTCTCTCTCTCTCTCTCTC
TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
T27311
GGGTGCAGGAATTCGGCACGAGTCTCTCTCTCTCTCTCTCTCTCTCTC
TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
```
Programs like BLAST have the option of filtering out low complex regions, called Masking.

Repetitive sequences increase the chance of a match during a database search.

Example of a protein with non-informative segments: rab2 gtpase

>rab2 gtpase, putative [Eimeria tenella]

mtiilvgnkc dlerrevtfq egfrvwaltg qdfarqhnli fletsaktaq ymdrlrv

Example of a protein with informative segments: rab2 gtpase

>rab2 gtpase, putative [Eimeria tenella]

mtiilvgnkc dlerrevtfq egfrvwaltg qdfarqhnli fletsaktaq ymdrlrv

Example of a protein with informative segments: rab2 gtpase

>rab2 gtpase, putative [Eimeria tenella]

mtiilvgnkc dlerrevtfq egfrvwaltg qdfarqhnli fletsaktaq ymdrlrv

Advise #2. Filtering Low Complexity Regions

Avoid using protein sequences over DNA sequences.

Sequence search based on antiTRAP DNA sequence.
Practical consideration in data banks searches

Advise #3. Preferentially use protein sequences over DNA sequences

Sequence search based on antiTRAP Protein sequence

Practical consideration in data banks searches

Advise #4. Consider the size of your protein

Consider the size of your protein over DNA sequences

Practical consideration in data banks searches

Advise #5. Re-search with the likely distant homologs

Re-search with the likely distant homologs over DNA sequences
Using Sfum_2476 as a query, we identify the antiTRAP (YczA) Blast hit.

Using DnaJ as a query, we identify the antiTRAP (YczA) Blast hit.

Since we only got DnaJ proteins in the default (100) first sequences. You should increase this number to see long distant homologs.

Advise #5. Re-search with the likely distant homologs.

Advise #6. Consider the effect of the database size in the E-value.
Advise #6. Consider the effect of the database size in the E-value

Sometimes you might get a wrong conclusion if you do not consider the size of small databases.

Compare these two E-values

Advise #7. Consider the multidomain nature of some proteins

Does the query really have a relationship with the results? One way to check is to run the search in the opposite direction… …but often not reversible even when true homology.
Web sites used in our practice
Figures are linked to their corresponding web sites

- EMBL-EBI
- PSI-PRED
- RSA Tools
- RegulonDB
- Pfam
- MEME
- MAST
- ClustalW
- UniProtKB/Swiss-Prot
- WEBLOGO
- TRANSFAC
- NCBI
- PubMed
- BLAST
- FASTA
- MUSCLE