Bioinformatics Exercises

AP Biology Teachers Workshop
Susan Cates, Ph.D.

Evolution of Species
Phylogenetic Trees show the relatedness of organisms

Phylogenetic Tree of Life

Common Ancestor (Root of the tree)
Rooted vs. Unrooted Trees

Rooted

Unrooted

Molecular Evolution

FlavoHbs

Plant Hb

Bacterial 2-on-2 Hb

Molecular Evolution

FlavoHbs

Plant Hb

Bacterial 2-on-2 Hb
Sequence comparison

Healthy vs. diseased
Identify genes involved in diseases

One organism vs. another
How closely related are two organisms
Unknown function vs. known
Lots of genes are not understood

Sequence comparison

One protein within a family vs. another
Identify mechanisms of disease, identify favorable characteristics (stability, specificity of substrate, affinity for substrate, etc.)
Vocabulary

If the same letter occurs in two aligned sequences then this position has been **conserved** in evolution. If the letters differ it is assumed that the two derive from an ancestral letter (which could be one of the two or neither).

Evolutionary processes in biology can introduce **insertions** or **deletions** in sequences.

In a sequence alignment, a letter or a stretch of letters may be paired up with dashes in the other sequence, called **gaps**, to signify an insertion or deletion.

If a biologist makes the statement that two sequences are related, he means that they are believed to have a **common evolutionary origin**.
### Identity

Indicates exact match in two (or more) sequences

### Similarity

Indicates chemical or structural similarity between unidentical aligned residues in two (or more) sequences

### Homology

The source of the similarity between unidentical aligned residues in two (or more) sequences is biological, such as evolutionarily related sequences in different species (same origin and function) or relationship between members of a chromosome pair in diploid organisms (homologous sequences are similar, but similar sequences are not always homologous)

### Specificity

The ability to reject false relationships, measured by the ratio of the number of true negatives to the sum of false positives and true negatives.

\[
\text{true negatives} \\
\quad \text{(true negatives + false positives)}
\]

### Sensitivity

The ability to detect all true relationships, measured by the ratio of the number of true positives to the sum of true positives and false negatives.

\[
\text{true positives} \\
\quad \text{(true positives + false negatives)}
\]
Studying distantly related sequences:
1. Use protein sequence.

Studying closely related sequences (identity, homology, paralogy):
1. Nucleotide sequence might be preferred (can see subtle changes that might be invisible in protein sequences)

use protein sequences rather than DNA when possible (why?)

Higher signal to noise ratio in protein sequences - what are the causes?

I. Mathematical Probability:
From a strictly mathematical point of view, assuming that there is an equal likelihood of any nucleotide appearing at any point in a sequence (which is generally NOT true biologically), what are the chances that a G in a nucleotide sequence will be randomly matched by a G in the same position in a different sequence? 1/4

From the same point of view, what are the chances that a G in a protein sequence will be randomly matched by a G in the same position in a different sequence? 1/20
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

   UUC to UUU mutation:
   - UUC encodes PHE (F)
   - UUU encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)

   AAG to AGG mutation:
   - AAG encodes LYS (K)
   - AGG encodes ARG (R)

Similarity “signals” contribute more information in protein sequences than in nucleotide sequences

a. Many categories, some can be weighted more heavily than others (nonpolar, polar, positively charged, negatively charged, aromatic, structural similarity)

b. Nucleotides -
   - transitions  purine to purine, pyrimidine to pyrimidine
   - transversions  purine to pyrimidine, pyrimidine to purine

---

Transitions:

- A to G
- C to T
- G to A
- T to C

Transversions:

- A to T
- A to C
- G to T
- G to C

---

Purine

Pyrimidine
URLs or google searches for bioinformatics students:

- The Human Genome Project:  
  http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

- The Human Genome Sequencing Center at Baylor College of Medicine  
  http://www.hgsc.bcm.tmc.edu/

- Cells Alive:  
  www.cellsalive.com

- The Biology Workbench:  
  http://workbench.sdsc.edu/

- National Center for Biotechnology Information:  

- Expasy (Swiss Institute of Bioinformatics)  
  http://us.expasy.org/tools/

- European Bioinformatics Institute  
  http://www.ebi.ac.uk/Tools/

---

Emphasize the factors that contribute to the dependence of biological studies on computers

How many bases in a genome?

- Human ~ 3 billion
- Rat ~ 3 billion
- Chicken ~ 1 billion
- Fish ~ 400 million
- Tuberculosis (bacteria) ~ 4 million
“About the Human Genome Project”

Question 1: How many genes are found in the human genome?
~ 20,000 - 25,000

Question 2: How many DNA base pairs make up the human genome?
~ 3 billion

Question 3: Name 2 project goals that will require the help of computers.
1. store this information in databases
2. tools for data analysis
Is the Houston Medical Center involved in genomics?

http://www.hgsc.bcm.tmc.edu/

Chimpanzee Genomic Analysis

About the project

The Human Genome Sequencing Center has sequence chimpanzee, Pan troglodytes. This data is available from:

The NHGRI has funded a chimpanzee genome project and WWSC.

International Genome Consortium Database

Conditions for use
• How many genome projects are being sequenced for different organisms at the Human Genome Sequencing Center, Baylor College of Medicine?

• How many primate genome projects are listed?

• Why do you think so many primate genomes are being sequenced?

• Why is it important to humans to learn about bovine genomes?

• Why is it important to humans to learn about microbial genomes?

www.cellsalive.com

Red Blood Cells

A human red cell has no nucleus and is responsible for carrying oxygen to the tissues. Read more about red cells and other images in the gallery.
The student can look up the answers to related cell biology questions at cellsalive.com, example: Where are genes located?

Molecular chains of Deoxyribonucleic Acid (DNA) inside each cell encode the organism’s genes. They are the hereditary information that determines what characteristics each cell, and, in a bigger sense, each organism will have.

**Sequence comparison**

**Healthy vs. diseased**
- Identify genes involved in diseases

**One organism vs. another**
- How closely related are two organisms

**Unknown function vs. known**
- Lots of genes are not understood
Proteins involved in genetic diseases

Lactase - digests milk sugar.
Insulin receptor - mediates the proper response to glucose.
P53 protein - tumor suppressor.

Exercise in Sequence Alignment:
Our example is HbB vs. HbS
Type the following web site into your browser:
Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.
The record for hemoglobin S should be returned. Hemoglobin is the protein in our blood cells that carries oxygen. Click on the link entitled “1HBSB”.

Next to the word Display in the grey region at the top of the file, change “GenPept” to “FASTA”.

[Diagram of protein search interface with a highlighted link and a change in display format]
This will display the amino acid sequence for hemoglobin S in FASTA format.

Hold down the left mouse button while you move the mouse over the sequence. This should highlight the amino acid sequence in blue. Now choose “Edit:Copy” from the browser window, or hit the buttons “Ctrl” and “C” to copy.
Now, click on the NCBI logo in the upper left corner of the web page to return to the main page.

In the dark blue menu bar at the top of the page, click on the word “BLAST”.
In the box of Protein options, click on the link entitled “Protein-protein BLAST (blastp)”.

Click in the Search box and choose “Edit: paste” from the browser menu or hit the “Ctrl” and “P” keys to paste the sequence into the search box.
Change the “nr” database to “swissprot”, then click the BLAST! button.

Click the Format! button.

Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1140215936–8025–124841773514:BlastQ4

A new window will open containing our sequence alignments.
Under the graph indicating the length of the top alignments, there will be a list of aligning sequences in order of decreasing alignment scores. Click on the score of the first item in the list, which is the highest scoring alignment. This will take you to the section of the file where you can view the alignment.

<table>
<thead>
<tr>
<th>Sequences producing significant alignments:</th>
<th>Score (Bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>56749858</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>56749871</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>56749856</td>
<td>sp</td>
</tr>
</tbody>
</table>

Identify the differences in the sequence of Query 1 and Subject 2

Query 1: ETTFVQAOAYKQVAGVANALAKYH

Subject 2: ETTFVQAOAYKQVAGVANALAKYH

A dissimilar substitution occurs at amino acid number 6.
The sickle cell mutation in Hemoglobin.

Sickle cell anemia is a blood condition seen most commonly in people of African ancestry and in the tribal peoples of India.

The individual must have two copies of the mutant hemoglobin gene to exhibit the sickle-shaped cells indicative of the condition.

The hemoglobin S beta subunit has the amino acid valine at position 6 instead of the glutamic acid that is normally present. This alteration is the basis of all the problems that occur in people with sickle cell disease.

Exercises in Multiple Sequence Alignment.

ClustalW is a multiple sequence alignment routine available online at the EBI website: [http://www.ebi.ac.uk/clustalw/](http://www.ebi.ac.uk/clustalw/)
Exercise in
Multiple Sequence Alignment:
Our example is non-alpha versus alpha Hb

The Biology Workbench
http://workbench.sdsc.edu/

The Biology Workbench is one of my favorite teaching tools, because the student can do a complete Bioinformatics project with the Workbench, from retrieving the sequences to performing multiple alignments and creating phylogenetic tree diagrams.

1. Retrieve sequences
   a. Be careful, there are many databases - too much information - too many results from a query confuses the student
   b. GenPept - Genbank gene products - full release
   c. SwissProt - manually curated European database
Genpept search of “fetal hemoglobin”:
7 results in over 1 minute
SwissProt search of “fetal hemoglobin”: 59 results in ~ 20 seconds

Databases selected: SWISSPROT

Matches (0 to 10) / 59

RESULTS OF (fetal AND hemoglobin)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Matching Database Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
<td>&quot;SWISSPROT:HGB2_HUMAN&quot; Hemoglobin gamma-2 subunit (Hemoglobin gamma-2 chain) (Gamma-2-globin) (Hemoglobin gamma-G chain) (Hb F Gamma) [Homo sapiens (Human)]</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>&quot;SWISSPROT:HGB1_HUMAN&quot; Hemoglobin gamma-1 subunit (Hemoglobin gamma-1 chain) (Gamma-1-globin) (Hemoglobin gamma-A chain) (Hb F Alpha) [Homo sapiens (Human)]</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>Hemoglobin beta subunit, fetal (Hemoglobin beta chain, fetal) (Beta-globin, fetal) (Hemoglobin gamma chain) [Ovis aries (Sheep)]</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Papio cynocephalus (Yellow baboon)]</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Macaca mulatta (Rhesus macaque)]</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Macaca fuscata fuscata (Japanese macaque)]</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>Hemoglobin beta subunit, fetal (Hemoglobin beta chain, fetal) (Beta-globin, fetal) (Hemoglobin gamma chain) [Capra hircus (Goat)]</td>
</tr>
</tbody>
</table>

JSO's slide on globin expression during human development

α₂ε₂ Hb embryonic (High O₂ affinity)

α₂γ₂ HbF (moderate O₂ affinity)

ε subunit

γ subunit

(placenta)

(lungs)

α₂β₂ HbA (low O₂ affinity)

β subunit

(low O₂ affinity)
Hb non-alpha subunit alignment

Analysis: HbG1 and HbG2 have an A:G substitution at position 136. HbE has the A at 136, HbB has the G.

Why does the HbS sequence have an N-terminal methionine?

Hb alpha and non-alpha subunit alignment

**Sequence alignment**

**Consensus key** (see documentation for details)
* = single, fully conserved residue
. = conservation of strong groups
| = conservation of weak groups
- = no consensus

**CLUSTAL W (1.81) multiple sequence alignment**
Analysis:

Highest scoring pairwise alignments:
1. HbG1 and HbG2
2. HbS and HbB
3. HbE with HbG1, HbG2

Lowest scoring pairwise alignments:
1. HbA and HbE
2. HbA and HbG1, HbG2
3. HbA and HbB, HbS