Bioinformatics
Bi1X-2010
Part I:
Public databases
Arbel Tadmor

Overview
• Obtaining sequence data from the internet
• Aligning two sequences
• Using the biologist’s google: BLAST

We’ll use hemoglobin as a case study

Related reading: Ch. 6 of Stryer (Biochemistry, 7th edition)

First stop: wiki

Structure of human hemoglobin
• In adults hemoglobin is a tetramer: α₂β₂
• Each subunit contains a non-protein heme group (that holds an iron)
• The iron binds to an oxygen (shifting absorbance from blue to red)
• Multiple subunits give rise to cooperatively in oxygen binding and unbinding allowing this protein to release more oxygen in the tissues (making it a good transporter).

α subunit
(141 aa)

β subunit
(146 aa)

Iron-containing heme group

Let’s focus on the α₁ subunit of hemoglobin. This gene is called HBA1

On which chromosome is this gene?

Let’s look up this gene

• Go to the PubMed website: google pubmed
• Search for HBA1 under gene category
We’ll start with human hemoglobin

RefSeq accession number XX_#

More about the RefSeq database here…

Genomic context of HBA1

Genomic context of HBA1

What is the length of …

• The genomic sequence (NC_...)?
• The mRNA sequence (NM_...)?
• The protein sequence (NP_...)?

Hint: Double click on gene then Right click→properties…

Genomic vs. mRNA vs. protein sequences

Genomic context
Genomic vs. mRNA vs. protein sequences

**Question 1**
Why is the protein product length 142x3 = 426bp shorter than the protein processed length (429bp)?

The mRNA includes additional sequence between the start and stop codon.

**Question 2**
Why is the mRNA length after splicing (576bp) longer than the protein processed length (429bp)?

The mRNA includes an additional untranslated 5' region and an untranslated 3' region.

Let's look at the sequence

**Homo sapiens chromosome 16, GRCh37 primary reference assembly**

Untranslated 5' region of the mRNA (blue)

ATG Start codon

5' splice site (GU)

Intron (green)

3' splice site (AG)

Total length mRNA = blue red = 576 bp

Total length protein = red = 429 bp
Typical Eukaryotic mRNA

We would like to find similar proteins in nature

Which sequence should we use for the search?

• Genomic?
• mRNA?
• Protein?

OK let’s grab the aa sequence of the protein

and the nt sequence of the protein

Nucleotide versus amino acid sequences

• Which sequence should we use to search with, the amino acid sequence or the nucleotide sequence?
Nucleotide versus amino acid sequences

It depends on your goal, but generally to find homologs, aa sequences is the way to go:

- Selection pressure on amino acid sequence is much stronger than on nt sequence
- Two random nt sequences share 25% (50% with gaps) identical characters whereas amino acids sequences share only 5% (10-15% with gaps), improving the signal to noise considerably.<we’ll see this later…>

Overview

- Sequence alignment - basic concepts
- Local vs. global alignment
- BLAST—a local alignment tool
- Exercise: alignment of random sequences
- Exercise: finding homologs of HBA1 with BLAST

Finding homologues sequences

- Homology≠ similarity

Homology: two sequences are descended from a common ancestor, therefore in an alignment identical residues at a site are identical by descent

Similarity: merely reflects proportion of sites that are identical

- What changes could occur over time in a sequence?

Matches, Mismatches and Indels

Two aligned, identical characters in an alignment are a match.

Two aligned, unequal characters are a mismatch.

A character aligned with a gap, represents an indel (insertion/deletion).

A A C T A C T - C C T A A C A C T - - - - - - C T C C T A C C - - - - - - C T C T T T

the bars show mismatches

10 matches, 2 mismatches, 7 gaps
Total number of characters 10+2+7=19
Percent identity = 10/19 = 53%
Substitution matrix for amino acids

**BLOSUM**: Blocks Substitution Matrix

- Values in matrix are empirical, based on a large sample of verified pairwise alignments
- \( S_{ij} \) element = \( \log \left( \frac{p_{ij}}{q_i q_j} \right) \) log of probability (a_j that amino acid i mutates into amino acid j) in a homologous sequence normalized by the probability of this match by chance given the frequency of the amino acids (q_i q_j) > positive: better than chance, negative: worst than chance

Positive for chemically similar substitutions (more likely than chance)

Common aa have a low score

Rare aa have a high score

We will come back to substitution matrices later when we talk about phylogeny

---

**Local alignment vs. global alignment**

- **Global alignment** – attempts to align every residue in every sequence
- **Local alignment** – find best subsequence alignment (useful for finding similar exons in two genomes, finding similar functional regions in a protein, etc.)

---

**BLAST Scores and Statistics**

- **Percent identity** is the fraction of identical characters
- **Percent similarity** (for amino acids) is the fraction of identical or chemically similar residues
- **Bit score** – A normalized alignment score \( S' \). The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment.
- **E value** – The E-value is a test statistic that gives an indication of the statistical significance of a given pairwise alignment by comparing it to a model of random sequences. It's the average number of sequences with this level of similarity (i.e. raw alignment score S) or better expected to be in the database by chance. Reflects the size of the database and the scoring system. Lower is better. The threshold is usually placed at \( 10^{-3} \).
- **P value** – The probability of finding at least one such sequence in the database by chance = \( 1 - e^{-E} \) (The E value is just the \( \lambda \) parameter in a Poisson distribution \( \lambda e^{-\lambda} / n! \) …)

Let's search now for homologs of human HBA1 in the refseq_protein database

Example of local alignment result

106 identities + 14 "+" = 120 positive hits

Before we align...

• Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?
Before we align...

- Should we align the amino acid sequences or the nucleotide sequences of a protein coding gene?
  
  **Amino acid**
  - Amino acids are more conserved
  - Aligning nts can lead to placing gaps inside codons

Nucleotide! contains more information

- The nucleotide sequence gives us the ability to
  - Detect silent mutations (codon bias)
  - Measure selection pressure
  - Detect frame shifts (rare but can occur in defunct genes)
- No information is lost (but it's always worth comparing hypothetical translation to annotated version if it exists)

Create a FASTA file (must have extension “.fasta”)

- For MEGA
  - Use unique names
  - Use '_' instead of spaces
  - Use only letters, numbers and the characters '.', '_' and '-'
  - Names should be <10 characters
  - Use meaningful names
  - Use the file extension ‘.fasta’

Overview

- Multiple alignment (HBA1 homologs)
- Phylogenetic trees
- Measuring evolutionary distance
- Building a neighbor joining tree with MEGA (HBA1 example)
- The case of rRNA sequences

Bioinformatics

Bi1X-2010

Part III: Multiple alignment and phylogenetic analysis

Arbel Tadmor
Multiple Alignment in MEGA4

- Google MEGA4 and install
- Launch MEGA4
- Drag the FASTA file into MEGA4

ClustalW:
Popular multiple alignment algorithm

Algorithm overview:
- Global alignment on all sequence pairs to find the distance between all pairs of sequences
- Uses distances to create a guide tree
- Align the closest sequences in the guide tree, followed by adding more sequences to the initial alignment

What can we do with a multiple alignment?

- Identify conserved regions within protein
  - Signifies conserved function
  - Useful for primer design
- Identify variable regions within protein
  - Functionally not important
  - Important but under positive selection pressure or rapidly changing
  - Identify non-silent mutations (e.g. leading to disease, due to adaptation, etc.)
- Construct a phylogenetic tree (discuss later)

Alignment using ClustalW

*Hint: be sure to align in the protein pane*

Manual inspection of alignment

- Pay attention to the edges
- Are there any obviously wrong sequences that did not align well?
  - Sequences too divergent? (must have ≥20% aa identity)
  - Reverse complement?
  - Frame shift?
Alignment using ClustalW

Example of residues conserved due to function:
His87 and His58 maintain the heme and oxidation state of the iron

Proximal His: Anchoring of the heme is facilitated by a nitrogen from a histidine that binds to the iron.

Distal His: The bound oxygen can be in two states, dioxygen (bound to Fe(II)) and superoxide (bound to Fe(III)). Oxygen must be released in the former because the latter is both harmful and leaves the iron in a state that cannot bind oxygen. The distal histidine binds more strongly to superoxide and the oxygen is therefore less likely to be released.

Cn3D demo of Homo sapiens hemoglobin subunit α

Histidines are conserved

Conserved His58
Conserved His87

Selection pressure is on the amino acid sequence

Some residues mutate to chemically similar residues

Phylogenetic analysis

What are trees good for?
• Identify close relatives
• Determine evolutionary relationship between sequences
Some tree terminology

- **Internal node** (e.g., inferred sequence of ancestral taxa)
- **External nodes** (e.g., sequence of taxa that exists today) = your data

Branch

Scale bar = number of substitutions per site

- **We will discuss only bifurcating trees**: each node has only two immediate descendant lineages, i.e., we assume evolutionary speciation is a binary process.

Trees have two elements: branch lengths and branching order

- **Topology** = branching pattern
- **To estimate a tree, you need to estimate**
  - Branch lengths (simple problem)
  - Branching order (difficult for many seq)

Branch lengths

- **Evolutionary distance** = accumulated horizontal distance between two external nodes = estimated number of substitution per site that differ between the two sequences = d

Distance between A and B =

Branch order:

- **Note that trees are like mobiles...**
- These two trees are equivalent in every way.

Rooted trees

- **Rooted tree**
  - Root = common ancestor of all taxa in the tree
  - Common ancestor of A and B
  - Clade = all descendants of any particular interior node

- **Rooted trees**
  - If a tree is rooted, root = left most internal node
  - By selecting a node to be a root, you set a time arrow
  - The more recently species share a common ancestor, the more closely related they are (e.g., A and C are more closely related than A and D)
  - Which node is the root? You “break the symmetry” by adding additional information: e.g., you know D is more closely related to the ingroup sequences than the ingroup sequences are related to each other.

- **Rooted tree**

Unrooted trees

- **Unrooted trees**
  - If it is not explicitly said that the tree is rooted, assume it is unrooted
  - Unrooted trees do not specify an evolutionary pathway (who descended from whom) only relationships among taxa

5 ways to root this tree
How to construct a tree?

Algorithmic methods (distance based) | Tree-searching methods (character based)
---|---
Use algorithm to construct a single tree from the data | Construct many trees then use some criterion to decide which is the best tree
1. Multiple alignment | 1. Multiple alignment
2. Calculate distance matrix = matrix of evolutionary distances between all pairs of aligned sequences | 2. Compare characters at each column in the alignment and give each topology a score
3. Calculate tree topology | 3. Choose the topology with the best score
Example: Neighbor Joining (many others) | Examples: • Maximum Likelihood methods
• Maximum Parsimony methods
• Bayesian methods

Measuring evolutionary distance between two sequences

• The evolutionary distance between two sequences \( d \) is the total number of number of aa/nt substitutions per site between the two sequences
• \( d=2rt \)
  \( t=\text{time in years}, \ r=\text{substitution rate per year per site} \)
• Branch length in tree = \( d \)
• How can we estimate \( d \) from the sequences?

Measuring evolutionary distance between two sequences

• \( p \text{ distance} \): \( p=n/n = \text{number of different aa/nt between two aligned sequences of length } n \)
  – Doesn’t account for multiple hits: \( A\rightarrow C \rightarrow T \)
  – Doesn’t account for back mutations: \( A\rightarrow C \rightarrow A \)
  – Doesn’t account for parallel mutations: \( A\rightarrow C; A\rightarrow C \)
  – Underestimates \( d \)
  – Saturates at \( p=0.75 \) (not a good estimate of \( d \) when \( p \) is high)
  – Can result in wrong topology
• 
  **Estimation of \( d \) based on a stochastic model**: aa/nt substitutions are modeled as a stochastic process.
  – Different stochastic models make different assumptions regarding the probability of aa/nt substitutions
  – Different models assume different substitution matrices

Calculating distances in MEGA4:

- nt p distance
Calculating distances in MEGA4: p distance

\[ p \text{ distance} = 100 - \frac{\text{percent identity}}{100} \]

Note that p distances < 0.75

Calculating distances in MEGA4: JC correction

Building a Neighbor joining tree

Close MEGA4 and open your MEGA file.
You should get this window with a nt sequence:

Building a Neighbor joining tree

Now let's build a NJ tree
Building a Neighbor joining tree

- Use JC nt model
- Calculate bootstrap support with 1000 replications

Building a Neighbor joining tree (unrooted nt tree)

This distance corresponds to 0.05 nt substitutions per site

- Human and primates group together
- Fish group together

Bioinformatics

Bi1X-2010

Part IV: Phylogenetic analysis of rRNA sequences - an exercise in class

Arbel Tadmor

How are SSU rRNA sequences different?

- rRNA genes are universal genes that are highly conserved
- Used as phylogenetic markers for species

Some technical points

- No aa sequence
- Selection is directly on the nt sequence
- Alignment should take into account secondary structure of rRNA molecule
- We will therefore use a dedicated website called green genes to align the sequences and analyze the alignment in MEGA

Universal phylogenetic tree based on small-subunit (SSU) rRNA sequences

rRNA secondary structure

357F Forward primer
1492RL2D Reverse primer
Program for today

- Learn more about the rRNA gene and the nature of the amplicon you generated
- Read the green genes website tutorial
- Convert your traces to nt using Sequence Scanner
- Align sequences with green genes
- Check for chimeras using green genes
- Import alignment into MEGA
- Calculate distance matrix
- Build a NJ tree
- Identify closest relatives of your sequences
- Is it likely that you found your phylotypes in the pond?

Example 1: Global alignment of two random 300bp nucleotide sequences

We will generate random sequences of 300 nt in Matlab:

```matlab
>> rand_int = floor(4*rand([1,300]))+1;
>> rand nt = int2nt(rand_int)
```

```
rand nt =
TCGAAGGCGCTCGTAGATACGTGTCCCAACTGTTGCCTAAGCGCGCTACAGTAGGGCGAGGCACGCTACTGTTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATATCCTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT
```

Example 1: Global alignment of two random 300bp nucleotide sequences

>>> Random seq 1
TCGAAGGCGCTCGTAGATACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTGTTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATATCCTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>>> Random seq 2
CCGTAACGTCGTCACAGAAGGCGTGCGAGCACCAATGTCTATCATCGGTCACTTGTGTTAGGTGTACCAAAGCTGAGAGTCGTCTATCTTATCTTCTAATAGTACCTCTATTAAGATTGAGTTGTTGACCCTACAAGCAAATCGTCGTCGCTCCCTCAAACTTGCCTGCTCTATCAACTCTAGAATGTTGTCTAAGCCGACAGGACCGAACGGTCAATGTGGCGTTCACGATTCAGGCATTATACAAGGCCAATCGTGCGGATGCGGCAGGGGCCCTTCTAACGAAGCGGGGTGCTGGATAT

Example 2: Local alignment of a 300bp sequence and an internal fragment containing a single insertion and a single mismatch using BLAST

>>> Random seq 1
TCGAAGGCGCTCGTAGATACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTGTTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATATCCTACGGGCCTTCCCCCAACATAAGGCAACTCATGCGGGGATACACATGCGCCTCGGTCCGATATGATTGCCGCATTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCTGT

>>> Random seq 2
AGTACCTCTCATACAGGATAGAAGAAGAGAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
Example 2: **Local alignment of a 300bp sequence and an internal fragment containing a single insertion and a single mismatch using BLAST**

```
Random seq 1
TCGAAGGCGCTCGGTAGAGTACGTGTCCCAACTGTTGCCTAAGCGCGCGTACAGTAGGGCGAGGCACGCTACTG
TTACGAGATTCCTACCGAAGAAAAGTTAAGCCCCTCGAAAGGTAACCATCGGAGCCCGTGATCTGGCATGAAATA
CTACGGGCCTTCCCCCAACATAAGGCAACTCATG
TGCGGGGATACACATGCGA
CTCG
GTCCGATATGATTGCCGCA
TTTTCACGGTTGCCTCATCAAGCCCGCCAACGGGTTAGTGGAACGAATATGAGGCAGACTCTCACATCGCTATCT
```

```
Example 2: Local alignment of a 300bp sequence and an internal fragment containing a single insertion and a single mismatch using BLAST
```

Example 3: **Global alignment of two random 300 residue amino acid sequences**

We will generate random sequences of 300 aa in Matlab:

```matlab
>> rand_int = floor(20*rand([1,300])+1);
>> rand_aa = int2aa(rand_int);
```

```
EMQSSVHIKTADYYITYFGHHFIVGEWLPNIRFPYFFWIITTDARNDMAIFNCQDETQSKKPSYNSDANNNYQYMWGCDLQEAKVTAMGNLHLWNHRGPRFQKDHACQLCEPHRGITETKRQKIDCSMNPHIPARKHYRGLNYMYMAENMRFIELQETEFQWNVWVNMEMSDGQLMPQYMNDMSME
```

```
Example 3: Global alignment of two random 300 residue amino acid sequences
```

For >100 aa, >25% identity is required to say with almost certainty that an alignment is not the result of chance.

```
Tree challenge
```

Is the frog more closely related to the fish or the human?

Suggested reading: The Tree-Thinking Challenge, Baum et al. Science 2005