About the Course

• Dr. Tommy Kaplan
  
  Mail  tommy@cs.huji.ac.il
  
  Office  Rothberg B-529
  
  Office Hour  By appointment

• Ms. Noa Moriel (TA)
  
  Mail  noa.moriel@mail.huji.ac.il
  
  Office Hour  By appointment

Website  moodle2.cs.huji.ac.il/nu16/course/view.php?id=76558

Course Book  *Biological Sequence Analysis* by Richard Durbin and Sean Eddy

Requirements

• Attendance

• HW assignments
  
  – There will be between 3 to 5 assignments.
  
  – Each HW assignment will have a theoretical part and a practical part.
    
    * The practical part will be a programming assignment which should be done in Python or Matlab.
* The theoretical part should be computer-typed and submitted as pdf. Explaining your logical inferences in crucial.

- Scribes
  Scribes are to be written in English using LaTeX and submitted as pdf. If you haven’t set a date to take the lesson notes please contact Dana at dana.perez@mail.huji.ac.il.

- Home exam
  - The home exam is an individual exam
  - Two days will be given for completing the home exam.
  - The exam will consist of 3 questions; one will be similar to something we have seen in class, and the other two will be in a higher level.
  - The exam will be open source.
  - Answers will be computer printed.

**Grade Composition**

- 50% HW assignments
- 20% Scribes
- 25% Home exam
- 5% Active participation

**Algorithms in Computational Biology**

**The Scientific Method**

We would like to learn how to translate a research question into relevant tests (e.g. are pig neurons different than human fetal neurons?). In order to do so we will learn to manipulate data, interpret it and decide whether the results are statistically significant.
For this purpose we’ll talk about models, learn how to build models (especially probabilistic ones) and how to learn their parameters. These models will then be put in a biological context.

Given a protein \( S \) and its sequence \( s \), we would like to know \( S \)'s function. In CS words, given a word \( s \) of a hundred letters above an alphabet of 20 letters, we’d like to find its meaning. Let’s assume that if another protein \( T \) has a sequence \( t \) that is similar to \( s \). Then, the protein \( S \) probably has a function that is similar to \( T \)'s. We might want to check the existing data for other proteins which have a known function and that are similar in sequence to \( s \).

We could use a scoring system based on sequence similarity, i.e

\[
\text{score}(s, t) = \text{the degree of imagination between } s \text{ and } t
\]

Using this function we could look for the best matching sequence \( t \):

\[
\text{outcome} = \arg \max_{t \in \text{all proteins with known function}} \text{score}(s, t)
\]

If we get a satisfying outcome, we could conclude that \( S \)'s function is similar to \( T \)'s function. What if we find no such protein \( T \) with a sequence that is similar to \( s \)? We could suggest another scoring system.

Domains are areas in a protein that act as an independent unit (they evolve and function independently of their context). Knowing this, we could suggest a scoring system with a scoring criteria of the degree of similarity of two protein sequences in specific domains:

\[
\text{score}(\text{sub}(s), \text{sub}(t))
\]

In CS words, in order to find a similar word to \( s \), you might not want to demand that each and every letter of the examined word \( t \) will be identical to the target word \( s \), but perhaps take important sub-strings and score similarity by finding a word \( t \) that contains similar sub-strings.

It’s all a matter of how you would like to define similarity.

By the theory of evolution, all organisms have evolved from the same origin, and have speciated over the course of time. This means that many organisms that are in the same “subgroups” (e.g vertebrates) share some DNA sequences. An evolutionary conserved sequence is one that originated from a certain species and appears in two (or more) sequences, across species or has multiple appearances within the same organism. Conservation across species indicates that a sequence is maintained by evolution. A highly conserved sequence is one that has remained unchanged far back in geological time. The farther back we go in
evolutionary time, the less we are likely to determine $S$ and $T$ are related in function. So the higher the estimated conservation level of these proteins needs to be.

Coding areas in the DNA are more likely to be evolutionary conserved than non-coding areas, and this is why we are more likely to find an evolutionary conserved sequence in coding regions of homo sapiens and the onion, than in non-coding regions. As the selection against mutations in non-coding areas is much milder, we would have to look for these sequences in evolutionarily closer species.

This leads us to the next suggestion, which is to score by sequence similarity in evolutionary conserved parts of the sequence. This is done under the assumption that protein function or structure is the cause for high evolutionary conservation. That is, we would like to know if there is a protein $T$ that is similar (not necessarily identical) to our protein $S$ in its evolutionary conserved sequences. To conclude that sub-sequences of $s$ is the same evolutionary conserved sequence as $t$, we want to check the hypothesis that they are derived from a common ancestor. We determine this by measuring the similarity between sequences $s$ and $t$.

**Dealing with a Scientific Question**

1. Formulating the biological question
   - e.g. do proteins $S$ and $T$ have a common ancestor?

2. Gathering data and knowledge
   - We need to gather both practical and theoretical knowledge.
   - The best practice is to look at the raw data before proceeding, and specifically to look for our objects of interest. Sometimes this is the best tool for defining or refining our hypothesis.

3. Mathematical formulation and modeling
   - Defining the question as a computer science problem (optimization, decision, etc.).
   - In our example, we actually have an optimization problem. We would like to define a distance function between two sequences.
An example of a distance function: Given two proteins, finding the optimal alignment between their two sequences. If $s = AGT$, $t = AACT$, we can align:

\[
\begin{align*}
\text{A} & \_ \text{GT} \\
\text{AACT}
\end{align*}
\]

or:

\[
\begin{align*}
\text{AG} & \_ \text{T} \\
\text{AACT}
\end{align*}
\]

The score is set to be the percent of identity between the sequences. Another example of a distance function: Editing distance. Editing distance is a scoring system by which the score is set to be the minimal number of single mutations (one base changes) that are needed in order to get $t$ from $s$. We could also set the scoring function to be the number of mutations needed to get from the sequence of the examined common ancestor to $s$ and $t$.

- We choose the distance function that goes along best with the step of gathering data and knowledge.

- We need to choose whether we want to plan a simple model or a complex one, and whether it will have parameters or nor. To get this done right, we need to use our biological knowledge. Suppose we have the sequences $s$ and $t$ as mentioned above. A lot of questions arise: How would we like to define the best alignment for them? Should we just choose the highest score our algorithm returns for $\sigma(G, A) > \sigma(G, C)$, or also consider adding a penalty for inserting a gap? Who says the score should even be additive? Meaning, there is no certainty that $\sigma(G, A) + \sigma(T, G) = \sigma(GT, AG)$ (in fact, it’s certainly not true for most cases).

4. Applying an algorithm

- We would like to formalize an algorithm for finding the score that we have previously decided upon.

5. Reevaluating parameters of the model

- After applying our algorithm we might want to adjust the values of our parameters. In the context of our example, we might want to
change the matrix values, or the gap penalty, according to the results of the alignment.

6. Estimating the statistical significance

- We want to know whether the score we got for two sequences is significant or not

- **P-value** The probability to receive our score or better than it. Or: how “surprising” is our score?

  A general example: Given a fair die, what is the probability to get 80 “sixes” in 100 throws?

  \[
p = \binom{100}{80} \left( \frac{1}{6} \right)^{80} \left( \frac{5}{6} \right)^{20} = 7.8239554 \times 10^{-44}
  \]

- In our case, \( H_0 \)- the proteins \( S \) and \( T \) are not derived from a common ancestor, and \( H_1 \)- \( S, T \) are derived from the same common ancestor. Suppose that the choice of nucleotides is distributed uniformly, and that our genome is \( 3 \times 10^9 \) bases long. Is it surprising that our genome has a common sequence of 20 nucleotides with the onion genome, which is \( 17 \times 10^9 \) bases long?

  The probability to sample a specific sequence of length 20 is

  \[
  \left( \frac{1}{4} \right)^{20} \approx \left( \frac{1}{1000} \right)^{4} = 10^{-12}
  \]

  So, the chances to find a specific sequence of length 20 in the human genome is

  \[
p = 10^{12} \times 3 \times 10^9 = \frac{1}{300}
  \]

- The previous p-value is the probability of finding a certain sequence of 20 nucleotides in the human genome. However, we want to know the chances for finding this sequence in the human genome, given that it appears in the onion genome. This means we will need to sample nucleotide sequences that are 20 bases long, \( 17 \times 10^9 \) times, and check if the sequences appear in the human genome. We would, in this case, demand at least one success out of \( 17 \times 10^9 \) trys. If we do this, we would probably wind up sampling two identical sequences by pure random chance. As it turns out, we shouldn’t be surprised at all if we find a common 20 nucleotide sequence between the onion
genome and the human genome, even if the genome sequences are distributed uniformly (which they are not, because generally speaking genomic sequences tend to repeat). The more experiments made, the less surprising the outcome is. This is a detail that researchers sometimes tend to “overlook” a bit when they repeat their experiment several times, out of the wishful thinking that their hypothesis must be correct. So, it is crucial to examine the significance of our score in consideration of the number of experiments done.

7. Conclusions

• This is the time for refining biological insights. If the score is significant enough, and we have done the background check appropriately, we could say that our results are significant.

• We could present the score of our hypothesis as it is, or the score of our hypothesis relative to the null hypothesis.

• Visualization is important. In our example, we could produce a histogram of the scores for two random sequences, one from human and one from onion. The distribution that we expect to see is close to a normal distribution, but a bit tilted toward high similarity, as we have chosen the optimal alignment for each sequence. On this histogram we will also present the degree of similarity between $s$ and $t$. 

![Graph showing similarity between two sequences](image)