

## **Class 10: Protein Alignments (1)**

**Held:** Tuesday, 29 September 2009

**Summary:** We consider how our techniques for aligning DNA sequences might change as we think about aligning protein sequences.

### **Related Pages:**

- EBoard.
- Reading: Chapter 4.

### **Notes:**

- Due Thursday: Analysis of HIV env sequences; HIV env paper. Are there questions?
- Sam will be unavailable for office hours on Wednesday.
- Friday noon CS Talk: History of Programming Languages. Free pizza. Any interest?
- Discussion of possible visit by Iowa State Bioinformaticist.

### **Overview:**

- Overview of chapter 4.
- From aligning amino acids to aligning proteins.
- Building PAM Matrices.
- Lab.

### Notes on Proteins

- Notes from Vida on the biological side of things.

## **BioConcept Questions from Chapter 4**

Since the chapter seems fairly straightforward, we'll go over the BioConcept questions. [From St. Clair and Visick, p. 90.]

1. Suppose a mutation changes a codon in a gene from GUA to GAA. What is the corresponding amino-acid change?
2. What are two ways in which this small change in DNA can produce a drastic change in the function of the protein encoded by this gene?
3. Even though this mutation changes only a single nucleotide, it is rarely observed when comparing actual genes from different organisms. Why isn't it more common?

4. The enzyme *lactase* is found in your small intestine and converts lactose from dairy products into two simple sugars. The **active site** of this protein, where the enzyme binds and breaks the lactose, is made up of several amino acids, and as you would expect, mutations that change these amino acids often affect the function of the enzyme. But, some mutations that change amino acids far from the active site also drastically affect enzyme function. What could explain the effect of these mutations?

5. Some mutation in *HBB* produce beta-globin proteins that appear to have exactly the same three-dimensional conformation as normal beta-globins. Yet, these mutations produce hemoglobin molecules that do not function properly. Can you think of a possible explanation?

6. Suppose a gene's coding sequence begins with ATGCTCCGGCAAAGG.... A gene in another organism begins with the sequence ATGTTAAGAAACCGT..., so there does not seem to be much sequence similarity. Would our conclusion be different if this were a protein alignment? (Hint: Translate the two sequences before answering the question.)

## Sequence Alignment: From Nucleotides to Amino Acids

- We've seen how to align sequences of nucleotides with Needleman-Wunsch, a prototypical dynamic-programming alignment algorithm.
- But Needleman-Wunsch, as we've used it, assumes that all mutations have equal cost.
- That's clearly not the case for amino acids.
- Hence, we build a function that gives the cost of substituting one protein for another
  - The function can be based on characteristics of the amino acids (e.g., we can assume that mutations that change hydrophobic amino acids to other hydrophobic amino acids are more likely to be accepted). = We can use a matrix that assigns a value to every mutation.
- Two common families of matrices are the PAM and BLOSUM matrices.
- How are they computed? (Does the book tell us?)
  - Generally by working from existing data (i.e., known sequences)
- Why does Figure 4.7 only show half a matrix?
- If PAM1 represents at 1% mutation rate, what does PAM250 represent?

## PAM Matrices

*Disclaimer: I could not track down the original PAM paper, and the variety of online resources are surprisingly inconsistent in their descriptions.*

- One of the standard (and perhaps oldest) substitution matrices.
- PAM1
  - 1% mutation
  - Fill in basic matrix with frequencies (e.g., the position indexed by (A,R) is the probability of seeing R in the mutant given A in the wild type)
  - Scale!
    - Convert to probabilities
    - Take log
    - Do other funky stuff

- That is, the value at position (i,j), representing a mutation from amino acid i to amino acid j is something like
 
$$\log\left(\frac{f(j) \cdot M(i,j)}{f(i) \cdot f(j)}\right)$$

$$= \log(M(i,j)/f(i))$$
  - Where f(i) is the frequency of amino acid i occurring.
- No, we won't do it by hand, but we'll talk about the design of the formula.
- To handle more than 1% mutation, we multiply the base matrix by itself k times.
- Problems with PAM?
  - No indels used in analysis.
  - Every position treated as equally likely. In practice, mutations seems more likely at some positions than others.
  - Choice of PAM250 (or whatever) is primarily heuristic
  - ...

---

Copyright © 2009 Vida Praitis and Samuel A. Rebelsky. This work is licensed under a Creative Commons Attribution-NonCommercial 2.5 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.