

Class 08: Protein Alignments (1)

Held: Tuesday, 20 September 2011

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

Related Pages:

- EBoard.
- Reading: Chapter 4.

Notes:

- Due Thursday: Response paper on the HIV paper.
- Due Thursday: HIV assignment.
- EC for Men's Soccer at Simpson Tuesday at 7 p.m.
- EC for Tutorial Debate Wednesday at 7:45 p.m.
- EC for Kington's Town Hall Meeting on Wednesday. (You shouldn't need extra credit to go - it's interesting and important, but we're giving it anyway.)
- EC for Volleyball Thursday at 7:00 p.m.
- EC for Family Weekend Poster Session (9:00-11:00 Saturday), provided you visit your classmates' posters.
- EC for Volleyball Saturday at 11 a.m.

Overview:

- Overview of Chapter 4.
- Bio-Concept Questions from Chapter 4.
- From Aligning Nucleotides to Aligning Amino Acids.
- Building PAM Matrices.
- Lab.

Notes on Proteins or "Weeks of Biochemistry in Ten Minutes"

- Side note: If you can come up with your own examples, we'd appreciate it.
- We assume that you've already looked at the chapter
- Chapter 3 was really about aligning DNA
 - But not all bases are created equal
- Central dogma review
 - DNA goes to mRNA goes to protein
 - Structure of DNA corresponds well to protein
 - In viruses and bacteria there's a direct correlation

- In Eukaryotes the DNA has a regulatory sequence, exons, and introns.
 - The exons are "fused" to build the mRNA
 - So there's still a correlation, but it's a bit different
- Important issue (that we've been ignoring): We should be looking at alignment somewhat differently in different categories.
- DNA is simple, proteins are more complex
- What's a protein?
 - A chain of amino acids
 - That folds into an interesting shape that lets it doing things
- Folding is affected by
 - Size of amino acids
 - Hydrophobicity/Hydrophilicity
 - Charge
 - Some funky things that happen in a few cases
 - Chaperones
 - Cleavage
 - And more ...
- Two classes of folding are highly predictable
 - alpha helix
 - beta sheet
- Folding of alpha helices is essentially instantaneous
- Structure levels:
 - Primary: Sequence
 - Secondary: Basic folds, such as alpha helix and beta sheet
 - Tertiary:
 - Quaternary: Complex of proteins
- These secondary structures fold into a tertiary structure
 - The tertiary structures combine into complexes
- Homodimers have identical subunits
- Heterodimers have different subunits
- Proteins have lots of functions
 - Enzymes
 - Structure
 - Regulation ('allosteric')
 - And more ...
- Cool thing: Odor recognition genes all produces proteins that have seven alphahelices. (Transmembrane domains.)
 - Concept from example: Mutations that 'mess up' one of these alpha helices are unlikely to 'survive'.
- What can happen
 - Deletion
 - Insertion
 - Substitution
- DNA mutation can create

- Amino acid substitution
- Premature stop (truncated)
- Cryptic starts (elongated)
- What can happen if you have an amino-acid substitution in a protein? (What changes can result?)
 - Alter active site
 - Change charge
 - Change function of protein
 - E.g., green fluorescent protein to red fluorescent protein is a single amino-acid change
 - No effect! Protein keep same activity
 - Polarity/hydrophobicity
 - And more ...
- Interesting example of protein folding issues from the chapter: Repeats
 - Huntington's disease (neurological)
 - Autosomal dominant
 - "Autosomal" - not on X or Y; not sex-linked
 - "Dominant"- only need one copy
 - Gene is Huntingtons
- Part of the gene has a large set of repeats
- Age of onset and severity correlates with number of copies
- This stretch of DNA is unstable - the number of copies you have is not necessarily the same as your parents. You can grow or shrink the number of repeats from one generation to the next.
- We don't know why this happens
- Project idea: Searching for repeats!

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
$$\log((f(j)*M(i,j))/(f(i)*f(j)))$$
$$= \log(M(i,j)/f(i))$$
 - 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 © 2009-2011 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.