

Class 10: Sequence Assembly (1)

Held: Tuesday, 27 September 2011

Summary: We consider the various techniques that are used to build the large DNA databases that we rely upon.

Related Pages:

- EBoard.
- Reading: Chapter 5.

Notes:

- Sam is back. Sorry for the extended absence, but citizenship requires it. He is likely to remain discombobulated for a few days.
- Thursday, we will meet downstairs in the Molecular Lab, Science 1104.
- Project 2.5 returned. General comments via email. BLAST papers will take a little more time.
- The lab instructions and protocol for the biology lab are now available.
- Programming assignment for chapter 5 (due Tuesday): Write a procedure, `generate_fragments(sequence, minlen, maxlen, coverage)` that generates a bunch of random fragments from the sequence, with fragments between `minlen` and `maxlen`. You should generate enough fragments to get the average coverage specified. Then, write a procedure, `determine_coverage(sequence, fragments)`, that aligns each fragment to the sequence and indicates how much each position is covered. Make sure to include sample runs of your program.
- CS Picnic, Friday, October 7. Sam should have the signup tickets on Thursday.
- EC: From Eternity to Here: Shrinkage in American Thinking about Higher Education. Today, 4:15, JRC 101.
- EC: Thursday Extra; Max Kaufmann '12 on generating parallel corpora. Thursday, 4:30, SCI 3821.
- EC: Biology Seminar Friday at noon. Unknown topic.
- EC: Men's Soccer vs. St. Norbert, Saturday at 2:00 p.m.

Overview:

- Similarity Matrices.
- Lab.
- Sequencing DNA.

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 - closely related species
 - 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.
 - And M(i,j) is the probability that amino acid i transforms into amino acid j
- 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. So PAM250 is PAM1 multiplied by itself 250 times.
- Problems with PAM? Here are some of my answers
 - 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
 - ...

Lab

- Do Project 4.5.

Sequencing DNA

- It's time to step back a bit and look at the biological (and bioinformatical) production of data.
- How do we get all the wonderful sequence data that we've been using, at least for DNA sequences?
- The Sanger method is used for short DNA sequences.
 - The book tells us it's really the only strategy used.
 - We'll look at how it works and some of the data it produces
 - Unlike the book, we'll use Ridom TraceEdit to explore data (TraceEdit is available for all three platforms.)
- Note: You may have wondered why the official FASTA format permits more than A, C, G, T, U, and N for DNA/RNA sequences. Hopefully, the sequencing data we looked at gives you a sense as to why.

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.