Sequence Alignment CH364C/CH391L Bioinformatics Marcotte Spring 2013

In general, biology can be thought of as the study of *variations on a theme*. Every modern organism inherited its traits, with minor random changes, from its parents.

At the molecular level, this corresponds to sequences of DNA, RNA, and proteins that are very similar, but not identical, between the parents and progeny.

This process of gene evolution can be modeled as a stochastic process of gene mutation followed by a "selection" process for those sequences still capable of performing their given role in the cell. Over enough time, as new species evolve & diverge from related species, this has the result of producing families of related gene sequences, more similar in regions where that particular sequence is critical for the function of the molecule, and less similar in regions less critical for the molecule's function.

Frequently, we observe only the products of millions of years of this process. Given a set of molecules (DNA, RNA or protein sequences), we would like to compare them, decide how they are similar, and decide if they are similar enough to be considered part of the same family or if the observed similarity is just present by random chance.

Recommended reading:

Durbin, Eddy, Krogh, Mitchison, Biological Sequence Analysis

This next couple of lectures will be largely drawn from Chapters 1-2 of Durbin *et al*..

First, an example of aligned protein sequences. Shown are 3 pairs of sequences, showing aligned sequences of proteins named FlgA1, FlgA2, and HvcPP. Between each pair the perfect matches and close matches (shown by + symbols, indicating chemically similar amino acids) are written.

Two biologically related proteins with similar sequences:

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
        ++K+K+GRLDTLPP  +L+ N    A+SLR ++  QP+     R+ W +KAGQ V V+AG++
FlgA2 TLQDIKMKQGRLDTLPPGALLEPNFAQGAVSLRQINAGQPLTRNMLRRLWIIKAGQDVQVLALGE
```

Also biologically related (& fold up into the same 3D protein structure):

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
       A   +         P        +L   I+ R L P + I     R+AW V+ G  V V
FlgA3 LAALKQVTLIAGKHKPDAMATHAEELQGKIAKRTLLPGRYIPTAAIREAWLVEQGAAVQVFFIAG
```

But these are biologically unrelated (& fold up into unrelated structures):

```
FlgA1 AGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQA-WRVKAGQRVNVIASGD
      AG+V  K G +  + PRT  ++      I+     P  PI    +++A WRV A + V V+  GD
HvcPP AGHV--KNGTMRIVGPRTCSNVWNGTFPINATTTGPSIPIPAPNYKKALWRVSATEYVEVVRVGD
```

The problem we face is how to distinguish the biologically meaningless match (FlgA1-HvcPP) from the biologically meaningful ones (FlgA1-FlgA2 and FlgA1-FlgA3). This won't always be possible, but we can (1) establish some objective criteria for scoring the quality of sequence alignments, (2) establish a random model for which scores are expected just by chance, then (3) decide if our observed alignments are significantly

better than we would expect from our random model. This last step will allow us to identify most of the biologically significant matches.

Creating a scoring scheme

We're going to assume that mutations in these sequences are independent from position to position. This turns out to be reasonable for DNA and protein sequences, and often inadequate for RNA sequences, which can show strong correlations between nucleotides at different positions. However, treating mutations as independent events allows us to create an *additive scoring scheme*, where the score for a sequence alignment is the sum of the scores for aligning the individual positions in two sequences. We'll also assume that sequence changes come in 3 flavors: *substitutions*, *insertions* and *deletions*. Insertions and deletions can be treated as equivalent events, simply by considering one or the other sequence as the reference, and are usually called *gaps*.

Now, we'll consider a pair of aligned protein sequences which we'll call *x* and *y*, although all of the arguments apply equally to DNA sequences. Following Durbin *et al.*, we'll refer to the amino acid at position *i* in sequence *x* as x_i and the amino acid at position *j* in *y* as y_j. To create our scoring scheme to account for the substitutions observed between *x* and *y*, we'll consider 2 possible models:

A *random* model (let's call it *R*), where amino acids in the sequences occur independently at some given frequencies. Here, the probability of observing an alignment between *x* and *y* is just equal to the product of the frequencies with which we find each amino acid. We'll call these frequencies *q*. So, the probability of seeing the alignment of *x* and *y* under the random model *R*, written $P(x,y|R)$, is:

$$P(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

A *match* model (let's call it *M*), where amino acids at a given position in the alignment arise from some common ancestor with a probability given by the joint probability p_{ab}. So, under this model, the probability of the whole alignment is the product of the probabilities of seeing the individual amino acids aligned:

$$P(x, y \mid M) = \prod_i p_{x_i y_i}$$

To decide which model better describes the alignment, we'll take the ratio:

$$\frac{P(x, y \mid M)}{P(x, y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

Such a ratio of probabilities under 2 different models is often called the *odds ratio*. If the ratio is greater than one, model *M* is more probable; if less than one, model *R* is more

probable. Now, to convert this to an additive score *S*, we can simply take the logarithm of the odds ratio (called the *log odds ratio*):

$$S = \sum_i s(x_i, y_i)$$

Here, $s(x_i,y_i)$ is the score for aligning one amino acid with another amino acid:

$$s(a,b) = \log\left(\frac{p_{ab}}{p_a p_b}\right)$$

We've written *a* and *b* rather than x_i and y_i to emphasize that this score reflects the *inherent* preference of the two amino acids (*a* and *b*) to be aligned.

So, we've used 2 simple tricks to create our additive score. From basic probability, the probability of a set of independent events occurring is equal to the product of their individual probabilities. From basic logarithms, it's often easier to calculate the product of a set of numbers by instead calculating the sum of the log of those numbers. Combining these two processes gives us our additive score.

From a set of correct alignments, we can learn the pairwise amino acid substitution scores (the $s(a,b)$'s) and arrange them into a 20x20 matrix called an amino acid *substitution matrix*, such as the BLOSUM50 matrix (this particular one has had the data scaled & rounded off to the nearest integer):

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

Using this matrix, we can score any alignment as the sum of scores of individual pairs of amino acids. For example, the top alignment in our earlier example receives the score:

S(FlgA1,FlgA2) = – 1 – 2 – 2 + 2 + 4 + 6 + ... = 186

and the incorrect alignment has the score S(FlgA1,HvcPP) = (best left to the reader to calculate!)

Now on to Gap Penalties

Practically any two sufficiently long sequences could be aligned if we allowed unlimited gaps to be introduced. To limit this abuse of gaps, we penalize their introduction into alignments, typically in one of two manners. We may have a simple penalty where the "cost" of introducing a gap is related to the length g of the gap in a linear fashion:

$$\gamma(g) = -gd,$$

where d is the cost of initiating a gap.

Second, we may have what's referred to as an *affine* score, where the penalty d for opening a gap is large, but then the gap may be extended with a lower penalty e:

$$\gamma(g) = -d - (g-1)e$$

The affine penalty turns out to be reasonable for many biological sequences, especially where a short or long insertion may be similarly probable.

As with substitutions, the gap penalties can be expressed as a ratio of probabilities of two models (the random and legitimate gap). Provided the frequencies of finding amino acids in the gap are the same a the frequencies of finding amino acids elsewhere in the sequence, the gap penalty, simplified to a log odds ratio, becomes $\gamma(g) = \log(f(g))$, where $f(g)$ is a function of the length of the gap.

Now, we have an objective system for scoring an alignment we produce.
Next, on to algorithms for finding the alignments...