Amino Acid Substitution Scores

Consider the problem of scoring a region of an amino-acid alignment that contains no gaps, such as:

```
GSAQVKGH
GNPKVKAH
```

Here we discuss standard ways of assigning a score to each amino acid pair, i.e., to each possible column of a gap-free pairwise protein alignment. Examples of such scoring matrices include the PAM30, PAM70, BLOSUM80, BLOSUM62 and BLOSUM45 matrices that are available on NCBI's blastp server. Such scores are appropriate for comparing two sequences about which we have no other information (as opposed to position specific scores tailored for a particular protein family). Thus, we seek a 20-by-20 array of numbers for protein sequence comparisons.

Observation 1. There exist independent and reliable means of deciding if a particular scoring matrix gives good results. Given an alignment that is optimal with respect to certain scores, does each non-gap column contain letters that are derived from the same ancestral letter by replacement operations? We can use 3-dimensional structural information to decide the "correct answer." With database searches, we care more about the scores than about the actual alignment, so, basically, the question is: how many of the known homologs of the query sequence score higher than the highest-scoring unrelated sequence? A number of protein families are extremely well studied and can be used to answer such questions. In contrast, for non-coding DNA sequences, it is difficult to determine the "correct alignment."

One approach is to use a knowledge of protein biochemistry to predict which amino-acid pairs are most likely to arise by replacement operations, and thereby obtain scores. However, it works better to simply look at a "training set" of correct alignments and observe frequencies of each kind of column. (Knowledge of biochemistry is of course important for determining the training set.)

To help motivate these ideas, ask yourself, "What should be the relationship between the score for aligning two `A`s as opposed to aligning two `W`s?" The point is that `A` occurs much more frequently than `W` in a typical protein sequence. Most people will say that a `W`-over-`W` column should score higher than an `A`-over-`A` column. Intuitively, the reason is that `W`-over-`W` provides stronger evidence that the alignment is correct since it will occur in a chance alignment of unrelated sequences much less frequently than `A`-over-`A`, despite the fact that `W`-over-`W` appears somewhat less frequently in correct alignments than does `A`-over-`A`.

Observation 2. We use scores that assess the likelihood that the alignment's columns are drawn from the population of correct columns, as compared to the likelihood they are generated by chance. Accordingly, we use two statistical models of an alignment: one reflects biologically correct alignments and the other reflects chance alignments of unrelated sequences. (We'll think of the columns as independent identically distributed (i.i.d.). Markov chains are an alternative, though more complicated.) It is the ratio of the two probabilities that interests us most. For the model of aligning unrelated sequences, we assign probabilities $q(x)$ to amino acids x, reflecting the frequency with which they appear in protein sequences. Thus, the probability that a random alignment of unrelated sequences happens to align the sequences $x_1x_2\ldots x_n$ and $y_1y_2\ldots y_n$ is $\prod_{i=1}^{n}(q(x_i)q(y_i))$. Assuming that the 20 q-values

are positive and sum to 1, these probabilities for all alignments of fixed length n sum to 1. The q-values are estimated from a sample of protein sequences simply by letting $q(x)$ equal the frequency of x in the sample.

The other ingredients for determining scores are the frequencies $p(x, y)$ that a column of a "correct" alignment consists of x and y. To interpret these numbers as probabilities, we require that their sum over all 400 amino-acid pairs equal 1. Thus, the probability of a correct alignment happening to be between $x_1x_2 \ldots x_n$ and $y_1y_2 \ldots y_n$ is $\prod_{i=1}^{n} p(x_i, y_i)$. We are interesting in the "odds ratio" $\prod_{i=1}^{n} p(x_i, y_i) \ / \ \prod_{i=1}^{n}(q(x_i)q(y_i))$. This is the ratio of the odds that an alignment of homologs would take the given form (i.e., involve the particular sequences $x_1x_2 \ldots x_n$ and $y_1y_2 \ldots y_n$) versus the odds that a chance alignment of unrelated sequences would take that form. We assign to the column x-over-y the odds ratio $p(x, y)/(q(x)q(y))$, so that the alignment's odds ratio is the product of its columns' odds ratios. As a convenience, we actually work with the logarithms of these ratios, which allows us to add values instead of multiplying them.

In summary, to each potential aligned pair, x-over-y, we assign the score

$$s(x, y) = \log\left(\frac{p(x, y)}{q(x)q(y)}\right)$$

where $p(x, y)$ is the frequency of the column in a target population of "correct pairwise alignments" and $q(x)$ and $q(y)$ denote the frequencies of amino acids x and y in some appropriate collection of protein sequences. To evaluate the relative likelihood that a gap-free alignment correctly matches two homologs, as opposed to matching unrelated sequences, we add its column scores. As a further practical convenience, we use a score matrix obtained by multiplying the entries of this log odds matrix s by some appropriate constant and rounding to the nearest integer.

Observation 3. The frequency that the column x-over-y appears in a correct alignment depends on the evolutionary distance; in theory we need a unique scoring matrix for each distance. To see this, consider the amino acids denoted `I` and `L`, which are very similar and frequently substituted for one another in highly similar proteins. What specific number should we assign to $p(I, L)$, i.e., for the frequency that a column of a correct pairwise alignments is `I`-over-`L`? The frequency with which `L` replaces `I` depends on how long the two sequences have been separated. Suppose that for sequences separated by 1 million years, the probability of `L` replacing `I` is 1%. Then in 5 million years, the probability of `L` replacing `I` is about 5%. Of course, some small percentage of the `L`s that replaced an `I` in the first million years will have been replaced by something else (possibly `I`), but we can expect that for small evolutionary distances the probability that y replaces x will grow linearly with time (assuming $y \neq x$). In general, we want to pick p-values (or the implied matrix of substitution scores) that best mirrors frequencies of aligned pairs in a "target population" of alignments.

PAM Matrices. Roughly 30 years ago, Margaret Dayhoff and her coworkers defined a family of substitution matrices that ruled the protein-alignment world for years, and which still are useful in certain contexts. *PAM* stands for "point accepted mutation" or "percent accepted mutation", which is a unit of distance between protein sequences. (The word "accepted" in this context

refers to mutations that have become fixed in the population.) One PAM of evolution means that the total number of substitutions (some of which may have been in the same sequence position) is 1% of the sequence length. After 100 PAMs of evolution, not every position will have changed, because some positions will have mutated several times, perhaps returning to their original state. In fact, even after 250 PAMs, proteins are still sufficiently similar that sequence homology can frequently be detected. There is no clear correspondence between PAM distance and evolutionary time, since different protein families evolve at different rates.

In Dayhoff's approach, the relative rates of the 380 possible amino acid substitutions (not counting "substitutions" that leave a position unchanged) were determined by inspecting alignments between protein pairs with at least 85% identity. Considering only very similar sequences allowed the correct alignments to be determined with high certainty. The relative frequencies of the various mutations can then be multiplied by a carefully chosen constant to give an average change in 1% of all positions. The resulting frequencies $p_1(x, y)$, after dividing by $q(x)q(y)$, taking the logarithm, multiplying by a certain constant, and rounding to an integer, give the "PAM1" matrix.

Scoring matrices corresponding to any PAM distance can be determined by extrapolating from 1 PAM. For instance, x mutates to y in two PAMs of evolution if and only if x mutates to some z (possibly $z = x$) in the first PAM and z mutates to y in the second PAM. To formalize this observation it is convenient to think in terms of the probability that a given x will be matched with y in an alignment of sequences differing by i PAMs, i.e., $r_i(x, y) = p_i(x, y)/q(x)$. Then, $r_2(x, y) = \sum_{z=1}^{20}(r_1(x, z)r_1(z, y)$. This relationship can be generalized to arbitrary PAM distances, allowing computation of the PAMi matrix for any positive integer i.

In practice, people frequently use a matrix that is geared to very distant, but still detectable, homologies (e.g., PAM250), figuring that not-so-distant homologies will be detected by almost any scoring matrix. Still, if one wished to bias the search toward very similar sequences and avoid distant matches, one might use another matrix. For instance, a PAM15 or PAM30 matrix might work well for pinpointing human-mouse ortholog pairs, since their average level of differences is around 15%.

BLOSUM Matrices. Steve and Jorja Henikoff took an alternative approach to determining a family of scoring matrices. They used their BLOCKS database, which contains ungapped multiple alignments, called *blocks*, of core regions from hundreds of protein families. (The name BLOSUM stands for BLOcks SUbstitution Matrices). This permitted them to directly tabulate the frequencies $p(x, y)$ for distantly related proteins, instead of needing to extrapolate from observation. The trick is to limit the "observed" alignment columns to fairly divergent sequence pairs, so that the frequency matrix $p(x, y)$ will have much of its total weight off the main diagonal (i.e., a lot of non-identity columns).

The Henikoffs took the following approach to tuning matrices to particular evolutionary distances. For instance, consider BLOSUM62. In each block, rows (sequences) were clustered such that sequences sharing at least 62% amino acid identity were clustered together. Then frequencies of aligned pairs were counted only between sequences in different clusters. Thus, the frequencies $p(x, y)$ that determined the BLOSUM50 matrix were constructed entirely from

fairly distant sequence pairs, while BLOSUM80 also utilizes alignments of fairly similar (up to 80% identical) sequences. Note that for large evolutionary distances, it is appropriate to use a BLOSUMi matrix for *small* i (which is opposite of the rule for PAM matrices).

One technical issue that arises in estimating the frequencies $p(x, y)$ is that raw counts of pairs (x, y) need to be adjusted so that information obtained from a family of proteins that happens to currently have many known members does not drown out that from a family with just a few known examples. For BLOSUM matrices, this was done by weighting each cluster of sequences in a block as a single sequence when counting amino-acid pairs. More precisely, the number of (x, y) pairs observed between sequences in different clusters (from the same protein family) were divided by $m \times n$, where the clusters have m and n sequences, respectively. (Observed values used to determine $p_1(x, y)$ for PAM matrices required a similar adjustment.)