Matrices of 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 two popular 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. (At the end of the discussion, we turn to 4-by-4 DNA substitution matrices that are appropriate for comparing human and mouse genomic DNA sequences.)

One approach is to base the scores on one or several physical or chemical properties of amino acids that affect the ease with which one can be substituted for another. For instance, substituting a large amino acid for a small one, or a hydrophobic amino acid for a hydrophilic one, is likely to adversely affect the protein's structure. Thus, we can expect such substitutions to be rarer in correct alignments of homologous sequences than are substitutions of "similar" amino acids, and hence two amino acids might be given a high score for having similar size and hydrophobicity, and a lower score otherwise. However, such approaches have not been as successful as the "log odds" (also called "log odds ratio") scoring matrices, which we now describe.

To help motivate these ideas, ask yourself, "What should be the relationship between the score for aligning two As as opposed to aligning two Ws?" 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 less frequently in correct alignments than does A-over-A.

Accordingly, we use two statistical models of an alignment: one reflects biologically correct alignments and the other reflects chance alignments of unrelated sequences. 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.

It requires careful thought to assign proper frequencies $p(x, y)$ to aligned pairs so as to reflect their frequency in "correct" alignments. One problem is to identify a large and "random" set of correct alignments. An even more basic problem is that no one matrix of scores is appropriate for every pair of homologous sequences. 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 Ls 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 aligments.

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 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.)

DNA Substitution Matrices Used by PipMaker. For PipMaker, we wanted scoring matrices customized for comparing genomic DNA sequences from human and mouse. We determined several alignment scoring matrices, each of which assigns a numerical score to each of the 16 nucleotide pairs. Given one of these matrices, an alignment is scored by adding the columns' scores and subtracting certain penalties for gaps. To define log odds scores, somewhat analogous to the PAM and BLOSUM matrices for protein sequences, we chose $p(x, y)$ as the probability that a homologous pair of nucleotides in a neutrally evolving genomic region consists of (x, y). This choice of scores is appropriate insofar as the goal of PipMaker is to identify precisely the orthologous nucleotide pairs, regardless of their function (if any).

The motivation for using more than one substitution matrix in PipMaker differs from that behind the PAM and BLOSUM families, which need to accommodate sequence pairs at a variety of evolutionary distances. When PipMaker compares human and mouse sequences, the evolutionary distance is constant (roughly 100 million years since separation). However, the proportion of G and C nucleotides varies considerably across the genome, which affects both the frequencies, $q(.)$, at which the various nucleotides appear and the substitution patterns governing values $p(.,.)$. We account for these phenomena by using three different substitution matrices, depending on the GC content of the sequences being aligned.

Currently, the volume of trustworthy human/mouse alignments in freely evolving genomic regions is inadequate to directly estimate the values $p(.,.)$ (in part because of the difficultly in verifying that a region is non-functional). Hence, we chose to extrapolate from the mutation biases observed in a class of alignments at a lower evolutionary distance, in somewhat the same spirit as used to derive the PAM matrices. We did this by tabulating approximately 80,000 substitutions in alignments of 600 kb of human DNA transposon fossils to their well defined consensus sequences (A.F.A. Smit, A. Kas and P. Green, personal communication). [Due to the nature of DNA transposons, almost all of the observed substitutions have accumulated without selection. This cannot be said about retrotransposons, since they have a long lifespan in the genome and evolve their sequence over this time, i.e. some mutations are due to selection for transposability.]

These results allowed us to estimate the relative rates of nucleotide substitutions (e.g., the ratio of G→A to G→C replacements), which we assume have remained constant (given a fixed GC level) over recent evolutionary time. Thus, we focus on the mutation biases from the ancestral to the modern sequence, unlike the symmetric relationships involving two modern sequences that are estimated for deriving the PAM and BLOSUM matrices. Because of a lack of DNA transposon copies yet identified in mouse, we have assumed that the substitution biases were similar in rodent and primate evolution. Moreover, for both sequences we use the mutation biases appropriate for the human GC level, even if the mouse GC level is different.

Given the relative rates of each kind of substitution, we estimated their absolute likelihoods by assuming particular levels of sequence divergence since the rodent-primate split, based in part on results summarized by Li *et al.* (1996). For instance, for regions with an average GC level (43%), we take the divergence from the common ancestral sequence to be 18% in human DNA (approximately 16% observed substitutions) and 35% (approximately 28% observed substitutions) in mouse. The substitution rate is a bit higher in AT rich DNA and lower in GC rich regions.

To approximate, e.g., the likelihood that a homologous pair consists of an A in human and a C in mouse, we computed

$$\begin{aligned} p(A,C) &= p_{human}(A \to A) \times q_{anc}(A) \times p_{mouse}(A \to C) \\ &+ p_{human}(G \to A) \times q_{anc}(G) \times p_{mouse}(G \to C) \\ &+ p_{human}(C \to A) \times q_{anc}(C) \times p_{mouse}(C \to C) \\ &+ p_{human}(T \to A) \times q_{anc}(T) \times p_{mouse}(T \to C) \end{aligned}$$

where $q_{anc}(X)$ denotes the frequency of nucleotide X in the ancestral locus, as estimated from the current frequencies in human and the human substitution data. The numbers in the following matrices are:

$$7.61 \log \left(\frac{p(X,Y)}{q(X)q(Y)} \right)$$

where the factor 7.61 has obscure historical roots. The numbers are given to several digits here, to permit their rescaling for other uses.

Because of the higher substitution level in rodents, scoring matrices are not symmetric around the main diagonal, e.g., aligning a human A and a mouse C has a different score from aligning a mouse A with a human C. On the other hand, we have used minor adjustments to "symmetrize" the matrices, in the sense making "opposite strand" comparisons yield the same score. For instance, the score for aligning human A to mouse C is the same as for aligning human T to mouse G.

The following substitution matrices were computed, as described above, for GC levels of 37%, 43% and 50%. An appropriate gap initiation penalty would be 25 (maybe 24) and the gap extension penalty should be 6 (maybe 5.5). A cutoff score of 200 avoids almost all false matches; with 180 you'll get some noise, but most of that can be eliminated by forcing matches to be in order (as in PipMaker's Chaining option).

37% GC level:			mouse:		
		A	G	C	T
	A	6.329	−1.569	−8.436	−9.957
human:	G	−0.091	8.320	−7.510	−7.657
	C	−7.657	−7.510	8.320	−0.091
	T	−9.957	−8.436	−1.569	6.329

43% GC level:			mouse:		
		A	G	C	T
	A	6.920	−1.727	−8.642	−9.946
human:	G	−0.417	7.288	−8.790	−8.324
	C	−8.324	−8.790	7.288	−0.417
	T	−9.946	−8.642	−1.727	6.920

50% GC level:			mouse:		
		A	G	C	T
	A	7.403	−1.531	−8.231	−9.098
human:	G	−0.699	6.533	−9.121	−8.628
	C	−8.628	−9.121	6.533	−0.699
	T	−9.098	−8.231	−1.531	7.403

Reference: Li, W.H., D. Ellsworth, J. Krushkal, B. Chang and D. Hewett-Emmett (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet. Evol.* **5**, 182-187.