BioE/MCB/PMB C146/246, Spring 2005
Problem Set 3

3. Sub-additive gap penalties lower the time required for computing gap penalties. In the special case of affine gaps, they reduce the gap computation to constant time, instead of O(n) time.

Sub-additive gap penalties provide a better biological model for gaps than per position gap penalties. Insertions or deletions are difficult to open in protein sequences, because of constraints on protein structure. However, once an indel is allowed in a structurally flexible region, it should be much easier to extend it to multiple positions. (Also, the original insertion or deletion might have involved a few adjacent nucleotides at once.) Therefore, it makes sense to penalize the initiation of a gap more heavily.

4. The construction of a PAM matrix begins with a group of aligned sequences that are 99% similar. A phylogenetic tree for the sequences is inferred. PAM assumes a Markovian mutational process, where each subsequent mutation of a residue is not affected by any previous mutations. From the alignments, the relative mutability of each residue (normalized to 1%) is calculated by dividing the number of accepted changes by the exposure to mutation. The number of accepted changes can be calculated by the number of mutations in the extant sequence compared to the inferred ancestor. Mutability between two residues a and b is calculated as follows: $\mathrm{M}_{ab} = \mathrm{freq}(ab) \times \mathrm{mutability}(a)/\mathrm{freq}(a)$. PAM^n matrices are generated by matrix multiplication of the M matrices, n times. The higher the number of the PAM matrices, the longer the evolutionary time and the less similar the sequences.

There are several problems with PAM matrices. First, they assume Markovian mutational processes. This may not be accurate because certain residues may be hypermutable (eg. regions of protein that are unstructured) and others hypo-mutable (eg. the catalytic core), which would not be accounted for in the PAM matrix. Second, it assumes that residues are independent, which isnt necessarily true for proteins. This information can't be incorporated into any scoring matrix, though, without detailed information about the protein's structure, which isn't feasible. Third, all calculations are based upon the original phylogeny, which is an approximation. Fifth, the mutations seen in a PAM matrix are biased toward those seen in short evolutionary time since the starting sequences were so similar.

BLOSUM matrices are constructed from groups of ungapped aligned sequences whose percent identity meets a particular threshold similarity. From the alignment, the joint probability of any pair of residues in a column is computed. Dividing by the product of the marginal probabilities for the individual residues, a likelihood ratio can be computed for observing two residues aligned due to common ancestry versus by chance.

BLOSUM matrices have too important advantages over PAM matrices: they work better in practice and they are derived from actual substitutions rather than inferred ancestral sequences. BLOSUM doesn't require any evolutionary model for its construction, which can be a disadvantage since we are attempting to infer evolutionary relationships between sequences when we implement the matrices.

5. Since multiple mutations may occur in the same position, slightly less than 2% of the amino acids will change.

6. The substitution score of a PAM matrix represents a likelihood ratio: the likelihood that two residues have common ancestry over the period of time represented by the PAM matrix, versus the likelihood that the residues are aligned by chance.

The score for the alignment represents the overall likehood ratio for the entire alignment to be descended from a common ancestor over the given PAM distance, divided by an alignment formed by chance. As the sum of the substitution scores for each pair of residues in the alignment, it represents joint likelihoods, assuming that each position in the alignment is independent.

7. Gap parameters are not estimated the same way as substitution score for several reasons. Generalized gap penalties use varying costs for different regions within a gap. Therefore, the assumption of positional independence used to construct substitution matrices does not hold for gaps. In addition, note that a gap doesn't actually exist outside of an alignment; instead, gaps are indicators of insertions or deletions. Calculating the frequency of an alignment between a residue x and a gap is actually measuring the frequency with which x is inserted or deleted from a sequence (these two possibilities are indistinguishable).

8. Marginal probabilities: M = 0.25 I = 0.175 K = 0.325 E = 0.25
Joint probabilities

```
     M     I     K     E
M 0.078
I 0.156 0.056
K 0.178 0.033 0.200
E 0.011 0.050 0.039 0.200
```

Product of marginal probabilities

```
     M     I     K     E
M 0.063
I 0.088 0.031
K 0.163 0.114 0.106
E 0.125 0.088 0.163 0.063
```

BLOSUM matrix: $\log_2$

```
     M     I     K     E
M 0.316
I 0.830 0.859
K 0.130 -1.77 0.921
E -3.49 -0.81 -2.06 1.678
```

Scaled by two and rounded:

```
    M   I   K   E
M   1
I   2   2
K   0  -4   2
E  -7  -2  -4   3
```

9.

	S	T	V
S	970.47	20.64	8.88
T	20.64	979.29	0.06
V	8.88	0.06	991.05