Making scoring matrices: alignment matrices on a far away planet
Eliot Bush

Star-faring scientists discover life on a planet in a nearby solar system. Though chemically different from life on earth, it has a number of similarities with terrestrial life. One of these is the use of polymers as information bearing molecules and catalysts.

Catalysis is handled by polymers (analogous to proteins) with are constructed from 8 different types of monomers. These can be represented with an 8 letter alphabet (we'll use the letters A-H). Life on this planet has a long history, and as on earth organisms have evolved and speciated for millions of years. One can identify homologous polymers from different species, and construct alignments between them. As is true with amino acids on Earth, some of these monomers are more chemically similar to each other than others.

Familiar as you are with methods for creating protein alignments, you decide that the best way to align these alien polymers is to take into account their similarities. You have identified two closely related organisms, and created sets of hand-made alignments between homologous pairs of polymers from them (exoBio.fa). You are now ready to use these alignments to construct a substitution matrix.

A. Use the alignments in exoBio.fa to make several PAM style substitution matrices. The file makeMat.py provides a starting point. It loads in the sequences in exoBio.fa, calculates the target frequencies for different alignment columns, and calculates the frequencies of individual letters. The sequences in exoBio.fa are 1% diverged. Your final program should be able to make any EXO_N matrix, where N is the amount of divergence which the matrix is suited for. For example we would use an EXO40 matrix to align sequences which are approximately 40% diverged.

The program makeMat.py takes the following parameters: a sequence file, a value for lambda, and the N number the matrix should be made for. You use it like this:

```
python makeMat_s.py exoBio.fa .33333333 40 > EXO40.mat
```

At present, makeMat.py loads in the file exoBio.fa and gets the frequencies of alignment columns and of letters. Your task is to modify it so that it produces a score matrix

The following python notes may be helpful.

If you have a Numeric array, gAr, and you want to multiply it against itself 3 times, you could do the following.

```
gAr2=gAr.copy()
for i in xrange(3):
    gAr2=Numeric.dot(gAr2,gAr)
```

To get log base 2 of a number m you could do

```
math.log(m,2)
```

To make sure you're getting the right result, you can compare your own EXO40.mat with a matrix we made, EXO40_example.mat.

When you have your program working, use it to make an EXO20 and an EXO200 matrix.

B. Now lets use these matrices to do something. The sequences for two additional alien species are located in spC.fa and spD.fa. The program align_w_matrix.py will create a local alignment between these two sequenes, given a scoring matrix. You can now run it and compare the results.

python align_w_matrix.py -10 EXO20.mat spC.fa spD.fa
python align_w_matrix.py -10 EXO200.mat spC.fa spD.fa

Note that the file trueSpCD_align.fa contains the “true” alignment between these two sequences. Which matrix gives a better alignment? Are species C and D closely or distantly related to each other?