Scoring Matrices[*]

Susan Cates

In bioinformatics, scoring matrices for computing alignment scores are often based on observed substitution rates, derived from the substitution frequencies seen in multiple alignments of sequences. Every possible identity and substitution is assigned a score based on the observed frequencies of such occurences in alignments of related proteins. The score is calculated from the frequency of occurrence of a match of the two individual amino acids in evolutionarily related sequences, and provides a measure of a chance alignment of the two amino acids. This score will also reflect the frequency that a particular amino acid occurs in nature, as some amino acids are more abundant than others. Higher scores indicate that the probability that those two amino acids aligned by chance is very small, and lower scores indicate a high probability the two amino acids aligned by chance, and are evolutionarily unrelated. Thus, identities are assigned the most positive scores, frequently observed substitutions also receive positive scores, but matches that are unlikely to have been a result of evolution, and are more likely indicative of unrelatedness at that position, are given negative scores. Matrices with scoring schemes based on observed substitution rates are superior to simple identity scores, or scores based solely on sidechain moiety similarity. The two most commonly used types of scoring matrices are the *PAM matrices* [3] and the *BLOSUM matrices* [1].

PAM (Percentage of Acceptable point Mutations per 10^8 years) matrices are based on global alignments of closely related proteins. The PAM 1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Scores are derived from a mutation probability matrix where each element gives the probability of the amino acid in column X mutating to the amino acid in row Y after a particular evolutionary time, for example after 1 PAM, or 1% divergence. A PAM matrix is specific for a particular evolutionary distance, but may be used to generate matrices for greater evolutionary distances by multiplying it repeatedly by itself. However, at large evolutionary distances the information present in the matrix is essentially degenerated. It is rare that a PAM matrix would be used for an evolutionary distance any greater than 256 PAMs.

Whereas the PAM matrices have been developed from global alignments, the BLOSUM (BLOcks SUbstitution Matrix) matrices are based on local multiple alignments of more distantly related sequences. For instance, BLOSUM 62, the default matrix in BLAST, is a matrix calculated from comparisons of sequences with no less than 62% identity. Unlike PAM matrices, new BLOSUM matrices are never extrapolated from existing BLOSUM matrices, but are always based on local multiple alignments. So, the BLOSUM 80 matrix would be derived from a set of sequences having 80% sequence identity.

The level of relatedness of a set of sequences, therefore, directly effects which scoring matrix is most appropriate for aligning the set, whether or not it is a PAM or a BLOSUM matrix. Comparisons of closely related sequences should use BLOSUM matrices with higher numbers and PAM matrices with lower numbers. Conversely, BLOSUM matrices with low numbers and PAM matrices with high numbers are preferable for comparisons of distantly related proteins. Nevertheless, a single matrix may be reasonably efficient over a relatively broad range of evolutionary change. The BLOSUM 62 matrix was chosen as the default for

[*] Version 2.7: Sep 10, 2008 3:31 pm -0500

[†] http://creativecommons.org/licenses/by/1.0

BLAST as a result of an analysis by *Henikoff and Henikoff* [2] wherein BLOSUM 62 detected more distant relationships in a BLAST search, and produced an alignment of diverged proteins more in agreement with three-dimensional structures, than did the corresponding PAM 60 matrix. The BLOSUM series does not include any matrices suitable for very short query sequences, so, in these cases, the PAM matrices may be used instead. Berkeley has a Matrix Information [1] website with a provisional table of recommended substitution matrices and gap costs for shorter sequences.

Now, take a look at some scoring matrices. A PAM Matrix[2] website sponsored by Wageningen University, in the Netherlands, allows online computation of PAM matrices. The default value is a PAM 250 matrix; calculate this matrix and look at the results. This PAM 250 matrix has a built-in gap penalty of -8, as seen in the * column. There are 24 rows and 24 columns. Of course, the first 20 are the amino acids, represented by the one letter code. B represents the case where there is ambiguity between aspartate or asparigine, and Z is the case where there is ambiguity between glutamate or glutamine. X represents an unknown, or nonstandard amino acid.

Exercise 1

In the PAM 250 matrix, where can the highest scores for each amino acid be found? Why?

Exercise 2

Would this be true for any scoring matrix?

Exercise 3

What row and column combination gives the highest score? (Specify the score value.)

Exercise 4

What is the second highest score? (Specify the score value.)

Exercise 5

Why are some scores for amino acid identities higher than others?

Exercise 6

Use the back button on the browser, and calculate a PAM 100 matrix. Are the two highest scoring matches the same combination of row and column as in the PAM 250 matrix? (Discuss with a sentence or two.)

Exercise 7

What is the gap penalty?

Exercise 8

Explain any differences in the gap penalties of the PAM 250 matrix versus the PAM 100 matrix.

To get an idea how the scoring matrix influences an alignment, perform the following exercise using the Biology Workbench[3] . The Workbench will require a password (it's free), but it will grant entrance immediately upon registration of a password. Enter the site, and scroll down the page until the five menu buttons are visible. The "Session Tools" button allows the naming of a session, so that different jobs in progress can be saved under distinct sessions. Select "Session Tools", then select "Start New Session" and click on "Run" to change the name of "Default Session" to a new name. Once the workbench has been exited, the session will remain. Subsequently, clicking on the dot to the left of the session name under the "Session Tools" menu, and then selecting "Resume Session", will recall the session. The Workbench policy at the time of this writing is that old jobs are deleted only when an account has not been accessed for 6 months. This tutorial will use sequences of hemoglobins (Hbs) from different organisms to illustrate the properties of scoring matrices. Choose the "Protein Tools" menu button, then choose the "Ndjinn Multiple Database Search" from the menu at the bottom of the page. Biology Workbench has a large number of databases to search, for this exercise, click in the box to left of the database description to choose the "PDBFINDER" database. Search the PDBFINDER database by typing in the PDB ID codes below into the search box at

[1] http://mcb.berkeley.edu/labs/king/blast/docs/matrix_info.html

[2] http://www.bioinformatics.nl/tools/pam.html

[3] http://workbench.sdsc.edu/

the top of the page. Import the sequences with the following PDB ID codes (use the OR operator between each PDB ID code to search for all of the records in the same search):

1. 1T1N from *trematomus newnesi* (antarctic fish)
2. 1SPG from *leiostomus xanthurus* (spot croaker)
3. 1QSI from *homo sapiens* (human)
4. 1IWH from *equus cabullus* (horse)
5. 1HV4 from *anser indicus* (goose)
6. 1HBR from *gallus gallus* (chicken)
7. 1H97 from *paramphistomum epiclitum* (trematode)
8. 1GVH from *escherichia coli* (enterobacteria)

The import function in the Workbench requires checking the boxes for all the PDB ID codes that were returned, then hitting the import button. There will be several subunits returned with most of these sequences, and some are duplicate sequences, so delete the following chains by clicking the box on the left of the ID code and selecting "Delete Protein Sequence(s)" from the pull-down menu at the bottom of the page:

1. 1HV4_C
2. 1HV4_D
3. 1HV4_E
4. 1HV4_F
5. 1HV4_G
6. 1HV4_H
7. 1HBR_C
8. 1HBR_D
9. 1H97_B
10. 1QSI_C
11. 1QSI_D

After the above sequences have been deleted, choose "Select All Sequence(s)" from the pull-down menu. Analyze the relatedness of this group of sequences by selecting "ClustalW" from the pull-down menu to perform a multiple alignment and draw a rooted cladogram. When the ClustalW page appears, before submitting the alignment, scroll down the page and change the "Guide tree display:" to "Rooted".

When the ClustalW results appear, first scroll down to the cladogram and observe which of these sequences are most closely related versus the more distant sequences. Notice there are three separate clusters of branches descending from the root. The two largest clusters are separated as a direct result of a structural characteristic of hemoglobins.

Exercise 9

What do each of these two clusters represent? (If the answer is not immediately clear, read this description of Hemoglobin[4] from the University of Brescia's on-line Biochemistry Course.)

Exercise 10

According to this cladogram, what is the sequence that is most closely related to human hemoglobin, ID code 1QSI?

Exercise 11

According to this cladogram, what is the sequence that is most closely related to the *E. coli* flavoHb, ID code 1GVH?

Exercise 12

According to this cladogram, what sequence is most closely related to the spot croaker Hb, ID code 1SPG?

[4] http://www.med.unibs.it/~marchesi/hemoglob.html#hemoglobin

It is not as clear from the cladogram which sequences are the most distantly related. However, scroll down past the cladogram to view the ClustalW pairwise alignment scores.

Exercise 13

Which two sequences yield the lowest pairwise alignment score?

At the very bottom of the alignment page, select "Import Alignments", to save this information for later reference, should that be necessary. The imported alignments can only be viewed through the "Alignment Tools" menu.

Apply the information elucidated by the multiple sequence alignment to test the impact of varying the scoring matrices in pairwise alignments. Return to "Protein Tools". Start with two sequences that are known to be closely related, the human Hb chain B, 1QSI_B, and the horse Hb chain B, 1IWH_B, by checking the box to the left of each of their codes. Choose "LALIGN" from the pull-down menu at the bottom of the page to compare two protein sequences to each other with BLAST. When the LALIGN page appears, next to select scoring matrix, choose PAM250 and run the alignment.

Exercise 14

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Now, return to "Protein Tools" and run LALIGN again with the same two sequences, 1QSI_B and 1IWH_B, except choose the "PAM120" matrix this time.

Exercise 15

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 16

(a) Which scoring matrix yielded the highest score for the alignment, and why is this matrix the best choice for this alignment? (b) List any regions where the two alignments differ.

Return to "Protein Tools", this time selecting the 1HV4_A and the 1TIN_B sequences, by checking the box next to their codes. Again, choose "LALIGN", and perform an alignment with the default PAM250 matrix.

Exercise 17

What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 18

Run LALIGN again on the same two sequences, using the PAM120 matrix. What is the (a) score of the alignment, (b) the length of the alignment, and (c) the percent identity?

Exercise 19

(a) Which scoring matrix yielded the highest score for the alignment, and why is this matrix the best choice for this alignment? (b) List any regions where the two alignments differ.

Exercise 20

Do the two different matrices always calculate the same value for percent identity when the same 2 sequences are being compared using each matrix? Why or why not?

Most bioinformatics tools available on the web have selected default scoring matrices that are based on a relatively exhaustive analysis of which scoring schemes work best over a wide range of query sequence characteristics. However, it is important to not only know which scoring matrix is used for a given alignment, but to consider the appropriateness of the default matrix for a given query as well. It is a recurring theme of bioinformatics that these computational tools should not be treated as "black boxes" where one can ignore the internal workings of the software, but instead require thoughtful interaction on the part of the user.

References

[1] Henikoff JG. Henikoff S. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.*, pages 89(22):10915–9, 1992.

[2] Henikoff JG. Henikoff S. Performance evaluation of amino acid substitution matrices. *Proteins*, pages 17(1):49–61, 1993.

[3] Dayhoff MO Schwartz RM. *Atlas of Protein Sequence and Structure, 5 suppl.*, volume 3:353-358. Nat. Biomed. Res. Found., Washington D.C., 978.