Weekly Homework 8

Ana-Cristina Rogoz

Substitution Matrices in Bioinformatics

May 7, 2020

In Bioinformatics, the concept of substitution matrices is used in order to determine how long it will take for one sequence to evolve into another one. These sequences are generally composed either by DNA or by amino acids. One of the easiest examples in which we can use matrix substitutions is in order to compute an alignment score for two different sequences of DNA. If we would take the first sequence as ATGACTGGA and the second sequence as ATATGA and, we could compute afterward the best alignment score. For the identity similarity matrix, we suppose that each nucleobase is similar only to itself and it cannot be transformed into any other one. In our example, we can choose different alignments since the second sequence has a smaller length than the first one. Therefore, an alignment such as ATGACTGGA and AT-A-TG-A would achieve 6 points, the maximum possible score since the second sequence only has 6 characters, which means at most 6 occurrences.

When it comes to amino acids, their sequences alignments scores can be computed much harder, since the alphabet now contains 20 characters, not only 4 as it did in the case of nucleobases. Also, there are several amino-acids that partially match in chemical properties and therefore those substitutions from one to another have higher chances of occurring. Therefore, scores of 0 and 1 are no longer able to clearly illustrate these probabilities. We will now have positive and negative scores with the significance that a positive value is more likely to occur than any random substitution and a negative value on the contrary.

Two of the most famous substitution matrices are BLOSUM and PAM. Both of these matrices are computed by observing substitutions of real-life protein alignments. The first

one, BLOSUM has its abbreviation from the 'Block Substitution Matrix'. For computing its values, we should have a finite number of amino-acids and compute the frequency of each character from the whole table. Afterward, the second step is to compute for each pair of characters the number of occurrences if we would take any two amino-acids from the same column. Having these individual and pairwise frequencies calculated, we are now able to compute the final substitution probability for a given XY amino-acids pair in the following manner: it is equal to the logarithm of the relative value of X frequency multiplied by the relative value of Y frequency, the result multiplied by 2 (because XY and YX have the same substitution probability) and everything divided by relative frequency of XY pair from the second step. In order for this matrices to be reliable and free of bias, each BLOSUM matrix has a threshold sequence identity. The block with no more than 50% of sequence identity is called BLOSUM50 and there are other alternatives as well such as BLOSUM62 and BLOSUM80. The second substitution matrix highly used is PAM, which states for 'Point Accepted Mutation'. The number attached to each PAM matrix, such as PAM1, represents the evolutionary distance between two amino-acids sequences. This evolutionary distance represents how, on average, the first sequence converted 1 in 100 amino-acids into the second sequence's amino acids. In practice, the most used options are PAM30 and PAM70.

Until this point, the text was entirely taken from my homework for last year's subject called 'Introducere in Bioinformatica'. Since I've previously talked more in-depth about computing the BLOSUM matrix, I'll continue by explaining the procedure for computing the PAM matrix. As previously mentioned, PAM matrices usually have a number attached such as PAM1, PAM30, or so on. These attached numbers represent the number of PAM units apart between two sequences that can be compared by that matrix. These PAM units are also called evolutionary units, and they represent the number of different amino-acids between the two compared units. If the two sequences have a different length than they should be realigned (by adding some empty spaces) so that the number of amino acids in common is the highest possible. Anyway, for a given matrix, let's say PAM100, a cell XY has the value equal to Z, meaning that the frequency of X and Y replacing one another in the two sequences which are 100 PAM units apart is equal to Z. These Z values are computed for each number of PAM units apart based on the following steps. Firstly, there are collected

and aligned sequences, afterward it is calculated for each column the number of occurrences of each amino acid and based on them, it is extrapolated the probability that amino acid from a particular column can be converted to another one, that can be also found on that same column. Moving on, for each amino acid X we can calculate the number of times X and another amino acid Y were present on the same column. An observation that can be made is that every PAM matrices are symmetric. Having the probability of amino acid X to appear, the total number of times when X was placed in the same column as another Y amino acid, there is computed a mutability measure which takes the previous two numbers into account, but also the number of PAM units for which we are computing this matrix. In the end, having all these measures computed, one can find the score between two amino acids X and Y in the PAM matrix.

As a conclusion, BLOSUM and PAM have different core principles from which they are computed and their scores should be distinctively interpreted.