Evaluating Methods for Analyzing Amino Acid Sequence from Database

Geliş Tarihi: 18.12.2017
Kabul Tarihi: 25.12.2017

Özden ÇOBANOĞLU

Abstract: This paper is about a book review related with how to analyze amino acid sequences derived from URFs (unidentified reading frames) and ORFs (open reading frames) through the genome written by R. F. Doolittle.1

Key Words: URF, ORF, amino acid, nucleotide sequence, and phylogenetic tree.

Introduction

Sequencing genetic material is really critical point to find out coding sequence that encodes a specific protein. At this purpose, Maxam and Gilbert developed the technique that allows cleaving DNA bases by chemical reagent2. In addition, Sanger’s dideoxy technique made huge amount of contribution in this area of sequencing4. Today a hundred thousands of nucleotide even whole genome of some organisms can be easily sequenced by computer-based technology. However, a reading DNA fragment from gel electrophoresis is not the end of the game, instead it is just the beginning. The real question arises about whether the derived sequence codify any known functional protein. Therefore in this review, it will be discussed a computer approach that play a big role to uncover if the open reading frames (ORFs) are from unidentified reading frame (URFs) and also to touch how the derived protein would be recognized if it really exist.

Search for Databank

Today’s electronic technology allows people to search easily through any publicly open databases around the world just by press a button from their personal computers. So if the sequence that researchers are looking for were already being identified, people would get the results that could be high enough to conclude that is a real match with the inquiry. However there are some pitfalls that the researchers need to aware of. Even when the researchers quest for random sequence in any database, they are still able to get so many responses that resemble somehow what they asked for. Thus the real job is beginning at this point, which means that the researchers need to carefully evaluate those

1 Uludag University, Faculty of Veterinary Medicine, Department of Genetics, 16059 Bursa, Turkey.
ocobanoglu@uludag.edu.tr
results whether they make sense biologically. If they believe the alignment is biologically relevant, then those responses should be considered seriously in later steps.

The first issue is to identify what type of source or database that is needed to scan. Because the searching prokaryotic gene sequence in eukaryotic gene sequence database will be a problem. Furthermore, it is important to know how often the databases are updated. The other issue is that most of the information about a known gene or sequence is really redundant in many databanks. However what they should search is small and nonredundant, but including most of the collections fair enough since some of the studies are definitely needed to be performed in nonredundant circumstances. This way, the researchers might be able to reduce central processing unit time (cpu) which deal with all instructions received from and software and hardware on the computer significantly and as well as get response quickly. It must be taken into consideration what kind of sequence either nucleotide or amino acid would be searched as a query. If the researchers look for DNA sequence by allowing gaps, their chance to find a similarity between two unrelated sequences would be higher than 50%. Hence this could cause to unidentified the existed relationship between the sequences. On the other hand, when nucleotide sequence were translated into amino acid sequence, a signal to noise ratio could be increased since twenty different amino acid codons are existed. A single letter code for each amino acid were also used to save computer space.

The aim is to find out the relationship between the sequences or at least part of the sequences. But the gap in the alignment due to indels (an insertion or deletion of bases) could make difficulty to get real number of matches or similarities between two related sequences. Therefore by examining overlapping segments, they may control the problems due to not only the rearrangement of the segments but also insertion or deletion of the residues. Moreover, using an appropriate scoring table and defining legitimate cutoff values make a great effect on discovering real alignment. By this way, people can give a full credit for exact matches as well as partial credit for similarities among residues by using weighted scale. Wilbur and Lipman (1983) and Lipman and Pearson in Fastp (1985) used algorithms based on k-tuple matching to find the match due to a homology as well as an analogy. This application works very well unless the sequence has repeat elements or the sequence is really short fragment. Computers could give 10-20% identity between two randomly chosen sequences even if the researchers do not use gap-scoring system. Thus they need to decide penalty for the gap and gap length in order to increase an optimum alignment score. In addition the degree of confidence is defined as if two random sequence (larger than 100 base pairs (bp) share more than 25% identity, they can conclude that those two sequences are highly related, after applying suitable gap penalty. But if two sequences share less than 15% identity, then it can be interpreted as they are unrelated fragments, so there is no point for going further analyses. On the other hand, if identity score between two sequences is about 15-25%, then they definitely need to make an additional statistical analyses about non-randomness of the fragments. Hereby when the researchers compare two fragments they need to consider the length of sequence and also a composition of sequence.

To correct those parameters, they can use a randomizing (jumbling) approach. What they need to do is to compare fragments by randomizing both sequences repeatedly. Then the alignment score is also compared with randomized set means. Finally they can estimate how significant an alignment it is. Overall if they have unidentified reading frames (URFs) and they want to find out which one of the sequences is the real open reading frames (ORFs), then they need to scan genome databank with most an appropriate searching algorithm program such as blast or fasta as a first step. Once they get the results, they need to an evaluate sequence alignment, as either it is real or just matching by chance.

Finding Real Match

The query sequence may perfectly align up with data base sequence. This can be an achievement if the goal is just to find out any exact match with previously reported sequence. On the other, if the sequence does not get any hit from target database that will not really be undesirable findings either. Even the nucleotide sequence could be a unique piece, thus people need to spend more time and do more analyses to understand whether this is a real sequence that encodes genuine protein. They have to start comparison with the sequences that are biologically relevant. Basically there are two alignment
algorithms to determine similarities between the sequences. One of them is a Needleman-Wunsch global alignment program. The key point of this is to construct comparative matrix by rewarding matches with a positive score. The scores are assigned to consecutive matches according to scoring table that were chosen in advance. They also have to penalize the gaps if the paths move out off diagonal. The overall goal is to find best end to end score mathematically between two candidate sequences by tracking back from highest scores at the outer edge. The other choice is to find a single best similar segment between two compared sequences, which is a local alignment or bestfit program. Unlike the global alignment, not only the gap and the gap length but also mismatches are penalized within a seller matrix in bestfit. In this case, they will find out a highest score within matrix, thus they are able to trace back an alignment from a highest internal score.

How efficient the result from those algorithms really depends upon the scoring table that is occupied and also the gap and penalty for the gap length that is assigned. But the researchers must always remember that the computers do what they are asked to do. In other words, a finding high optimal alignment score does not really prove that those sequences are indeed related biologically. Hence they just not only search and align the sequences but verify the results also. To accomplish that objective, people can do jumbling test by randomizing two sequences and comparing those various segments. Then they can carefully interpret the results whether the findings make sense or not. The other way is to find out if the query sequence is related functionally or structurally with the other sequences, they can utilize dot-matrix plot by putting a ‘dot’ for each match in the matrix, thus the researchers can observe the longest diagonal with most dots and fewest gaps, which would be most probably the best diagonal between two sequences.

**Constructing Tree**

Once the researchers really get significant hits from database, the next step is to find how the sequences are related in evolution or when the divergence started in evolutionary pathway. It is certain that if the sequences that are compared share high similarities, then they most probably related in evolution. The studies displayed even different parts of the protein tend to alter in several ways. On the other hand, some parts of the protein such as catalytic units and binding sites are found to be most concerned regions among protein structures. Therefore people want to estimate phylogenic relationships amongst the related organisms. How the organisms are different from each other really depends on how fast or slowly their protein sequence change during evolution.

Constructing phylogenic tree are also give a chance to group the related proteins into big families by utilizing protein sequence alignment. The length of the branch in the phylogenic tree represents the evolutionary distance between protein sequences. The building tree based on sequence information is easily accomplished. But the researchers need more complicated approach in order to construct tree for more extended set of sequences. Once they get convenient difference matrix, there are many programs that allow a finding topology and calculating the length of branch. The main point when constructing good tree is that the values in different matrix must be well defined, thus the good matrix really depends upon how consistence the sequence alignments are. Today’s computer based technology makes to derive multiple alignments very quickly. The point that we come up with the researchers should always keep in mind about time consuming (cpu) and space usage when they run the huge amount of data set simultaneously.

**Prediction Structure and Function**

Prediction of protein structure can be achieved by three different levels. People can predict a primary structure of protein from amino acid sequences. However prediction of secondary or tertiary structures are not always easy or also possible. But they can estimate protein structure by using modeling such as a homology, statistical and stereochemical. If the sequence looks alike to any known protein in databank, they can predict the structure of protein by using related protein which is already crystallized, as a model. They can also utilize an information about common functions that proteins share together. They can find out atomic coordinates of the specific proteins which three dimensional structures are known from protein database. Even if proteins are more distantly related, they will give good results by employing this method. Today both the recombinant DNA technology that have great effects on purifying proteins and X-ray techniques that even allows for crystallizing membrane proteins ma-
ke a great contribution for collecting X-ray data. Many proteins have repeat sequences from lower limit of two to hundreds residues. Since repeated sequences have structural or functional connotations, the researchers need to weight them properly. They also have to take into account of tandem duplications. Because if the long protein sequence does not give any signal for internal duplications, they can be accepted either their sequences alter quickly or the protein is very ancient. There is another tool, which is hydropathy plot that can be utilized to measure the polarity. In other words, it gives an idea about what parts of sequence would like to be in inside or on outside of proteins. But the researchers need to also consider membrane-spaning segments from the plot.

There are several methods available to predict secondary structure of protein. Chou-Fasman methods allow calculating probability for being α helix, β-sheet or a turn-coil for a given segment of sequence from existing crystal structures with 66-85% accuracy. If the protein sequence shares about 50% identity with another protein, the most probable protein structure could be easily modeled.

**Conclusion**

If the researchers do not have all the necessary informations about homologous protein from every known structures and also all of the proteins may not be crystallized yet, three-dimensional modeling will not be easily completed. Conversely, it does not mean that it is not possible. Actually it will be possible in near future but it will certainly need more efforts and more improvement in structure analyzing technology, and much more powerful computers, as well. But if they are still not confident about the results, the best way to proceed is to complete their analyses by supporting with laboratory experiments which can actually prove the existence of candidate protein and gives an information whether it is derived from ORF or URF.

**References**