Relationships in rank and file: Better sequence searches of genes and proteins

February 23 2009

Since the sequencing of the human genome eight years ago, enormous progress has been made in analyzing and understanding it. Nevertheless, the function of most human genes is still barely understood. An important first step in determining the function of a gene or protein is to compare its sequence with the sequences of hundreds of other organisms that are experimentally easier to investigate. From the functions of related genes or proteins identified in these database searches, the researchers can often infer the unknown functions of human or animal genes.

Now, computational biologists Johannes Soeding and Andreas Biegert of the Gene Center of LMU Munich, Germany, have successfully developed a method that makes database searches significantly more sensitive, while being just as quick. Instead of comparing sequences letter by letter, their idea is to take the sequence neighbors surrounding each letter into account during the comparison. This idea should be generally applicable in other areas of sequence searching and sequence analysis.

The rule for both genes and proteins is: their function is primarily based on the sequence of their DNA or amino acid components. Genes with similar sequences frequently have a similar function. The same goes for proteins, although for them, the three-dimensional structure into which they fold, and which cannot be predicted offhand from their sequence, is equally important. Still, similar protein sequences suggest relatedness or, in other words, the descent from a common ancestral protein, and with it

a similar function.

Accordingly, the sequences and functions of genes and proteins all get stored away into databases, which scientists around the world use for comparing their new data against. But even the best and most frequently used algorithms such as BLAST (Basic Local Alignment Search Tool) have to make use of certain simplifications in order to make efficient searching in the gigantic databases possible at all. After all, the researchers expect BLAST to compare a given sequence - the letter code describing the sequence of DNA components or amino acids - with all sequences in the database in just a few minutes.

Search engines like BLAST evaluate the similarity between a pair of sequences by aligning them underneath each other in such a way that similar amino acids come to lie in the same columns. The sequence similarity is then calculated by adding the similarities of all aligned amino acids. Here, the similarity between amino acids is measured by how often they mutate into each other without adverse effects, a measure that largely coincides with how similar their sizes and other biophysical properties are.

BLAST has been the most important method for sequence searching since its development in 1990. It is called up around 500,000 times a day from all around the world. Yet this tried and true program is far from perfect. When evaluating the similarity of two amino acids, it ignores their neighboring amino acids, their sequence context. Johannes Soeding and Andreas Biegert of the Gene Center Munich and the cluster of excellence "Center for Integrated Protein Science Munich (CIPSM)" of LMU Munich have now developed a method that significantly improves similarity searches: Their "context-specific" BLAST, or CS-BLAST, can sniff out twice as many distant "relatives" of proteins as BLAST.

When determining the similarity of an amino acid to the reference

sequence, CS-BLAST includes the sequence context of every amino acid, namely its six left and six right sequence neighbors, in the analysis. "The idea is that the context says much more about how likely two amino acids are to mutate into each other", explains Soeding, who heads the group for "Protein Bioinformatics and Computational Biology" at the Gene Center Munich. "Take as an example folded and unfolded regions in proteins. In an unfolded region, the amino acid valine, for example, can usually mutate into any of the other 19 amino acids without any adverse effect. In a folded region on the other hand, it will mutate with high probability into other hydrophobic, or water-repelling, amino acids."

The program is based on a very general idea that can be applied to every kind of sequence search and alignment method. The researchers have demonstrated this at the example of PSI-BLAST, an algorithm in which the related sequences already found are aligned one under the other into a so-called multiple alignment. This makes it possible, for example, to identify positions at which only certain amino acids can occur, which improves PSI-BLAST's ability to distinguish related from unrelated proteins. "We managed to increase PSI-BLAST's sensitivity significantly by making use of the sequence context. That way, two consecutive searches using our context-specific version of PSI-BLAST deliver better results than five searches using the conventional engine," says Soeding.

The new method is just as fast despite its better sensitivity, explains the researcher, because the sequence search takes place in two steps: "Both in conventional BLAST and in our method, a search matrix is first calculated," Soeding continues. "This step is more complicated when you do it our way, but at one second, it is still very fast. Only the second step, the database search using the search matrix, takes a lot of time - and this step is the same for both approaches."

In the future, the scientists intend to apply the newly developed

algorithm to genomic alignments as well, where not only individual genes, but rather entire segments of DNA are compared. "As with proteins, there are certain key regions in DNA that fulfill crucial regulatory functions," explains Soeding. "You can identify these regulatory regions, which are important for a deeper understanding of many diseases, by comparing the human genome with those of other mammals." Using a context-specific method, the LMU researchers intend to substantially improve the quality of such genomic alignments and, with it, the identification of regulatory regions. "We believe context-specific methods could become standard throughout the entire field of biological sequence analysis," concludes Soeding.

Source: Ludwig-Maximilians-Universität München