From sequence to function using links to ortholog genes

E. Bertocco, N. Cannata, S. Toppo, P. Fontana, P. Scannapieco, G. Valle

CRIBI, Università di Padova, via U. Bassi 58/b, 35131 Padova, http://muscle.cribi.unipd.it

The recent advancement in technology has made possible the sequencing of the complete genome of several eukaryotic model organisms as well as many prokaryotic organisms. Now, the purpose of functional genomic is to assign a role to these genes and to the proteins that they encode. The simplest method to assign a putative function to a novel gene is based on similarity search, either at the nucleic acid or amino acid level. The rationale is that if a similar sequence has already been characterized in the same or in other organisms, then we can usually assume that our sequence may have a related function.

Recently, the approach of making functional links based on sequence similarity is getting even more attention since we can compare a sequence against the entire set of sequences of several model organisms. In this case the information that we can infer goes beyond the simple link between two sequences and allows further conjectures based on the full knowledge of the entire set of genes (for a review see Eisenberg, 2000).

Our group has already carried out a systematic comparative analysis between yeast and human genes that led us to the identification and characterisation or several human sequences similar to yeast (Hussy), not previously known (Stanchi et al., 2001). Here we describe a more extensive study that opens the possibility of linking ortholog genes of model organisms to our muscle-TRAIT (TRAnscript Integrated Table) database. We have set up an automatic procedure for a systematic similarity search based on the BLASTX program, between our muscle-TRAIT transcripts and the full complement of proteins inferred from the genomic sequences of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*, obtained from specialized database at the European Bioinformatic Institute (EBI). All the resulting information has been organized into *SQL* relational tables that allow a complex query system.

For our analysis we tested a set of 4,300 sequences of our muscle-TRAIT database. Of these sequences 1,061 gave a BLASTX similarity match better than e-15 against the SWALL (protein databases of C. elegans, S.cerevisiae and Drosophila) protein database producing the results shown in the table.

The above analysis shows that there are many unknown human muscle transcripts that are significantly similar to sequences of the three model organisms considered in our study. This is particularly surprising since these sequences are limited to the 400-500 bases at the 3' end region of the mRNA, which are often non-coding. We expect that with the completion of the human genome and with the resulting possibility of inferring the full-length sequence of each putative transcript, the number of hits will be much higher. It should be also noticed that the term “known transcript” refers only to the knowledge of the sequence and not to the function. In fact, many “known transcripts” of the above table are absolutely unknown in terms of their biological function.

References

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O.Protein function in the post-genomic era. **Nature** 405, 823-826 (2000).

Stanchi F, Bertocco E, Toppo S, Dioguardi R, Simionati B, Cannata N, Zimbello R, Lanfranchi G, Valle G. Characterization of 16 novel human genes showing high similarity to yeast sequences. **Yeast**. 18, 69-80 (2001).