TRAIT: a database of transcripts expressed in human skeletal muscle

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

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

The project of discovering and cataloguing the genes expressed in human skeletal muscle has been the main effort of our group during the past years. The experimental approach is based on systematic sequencing of muscle Expressed Sequence Tags (ESTs), obtained from cDNA libraries that have been specifically designed to collect the 3' end of each transcript. At the bioinformatic level the project can be divided into four parts: 1) Development and maintenance of an integrated relational database (muscle-TRAIT) of annotated transcripts and integration of the information into a suitable retrieval system based, at the moment, on a SQL engine; 2) Implementation of tools and database structures to annotate transcripts and to align them on human genomic sequences; 3) Protein domain searches over known and predicted transcripts; 4) searches for putative ortholog genes.

The management of the annotation is quite complex since the information may be very different: from published work to predicted data obtained automatically by bioinformatic procedures. In order to deal with this problem, we have implemented a feature to control the mode of automatic updating. Therefore, the transcripts that were generated after automatic processing can be recognized and will be updated automatically by the system, while the other transcripts will need different levels of manual verification.

Currently, from over 30,000 ESTs we have obtained about 5,000 independent clusters, each identifying an individual transcript. The sequence of each cluster can be analysed by means of automatic procedures involving similarity searches, clustering and multiple alignment algorithms. Currently, over 1,500 transcripts have been identified with these procedures; but, the availability of the entire human genomic sequence will give us the possibility to extend our analysis and to retrieve the whole gene structure of the unknown transcripts expressed in human skeletal muscle.

One of the most recent studies that we have implemented in our database has been carried out on the amino acid sequences obtained from the prediction analysis described above, on 200 putative proteins with unknown function and on 939 known protein sequences from our annotated database. The amino acid sequences were searched for the presence of known domains with the program HMMER on PFAM5.5, using a threshold value of e-5. See Table.