Muscle-TRAIT: an integrated platform for storage, annotation and retrieval of data related to muscle transcripts

S. Toppo, N. Cannata, C. Romualdi, P. Fontana, P. Laveder, G. Lanfranchi, G. Valle

CRIBI, Università di Padova

The Genomic Research Group of the CRIBI Center, University of Padua, has been working for several years on the identification and characterization of genes expressed in human skeletal muscle. The original project was based on the discovery of novel genes by sequencing Expressed Sequence Tags (ESTs) from specially designed cDNA libraries restricted to the 400-500 terminal bases at the 3'-end of the transcripts. The research work led to the identification of about 5,000 independent muscle transcripts and to the creation of Muscle-TRAIT, an integrated database of information related to muscle transcripts.

Annotation of large subsets of genes

We have developed a series of bioinformatic tools that help to perform high quality annotation of large subsets of sequences in an efficient way, with the possibility of a continuous automatic updating and with the option of a manual input for adding further information or making changes to the records. Typically the starting point is a set of partial cDNA sequences, such as ESTs or cDNAs that we may want to use to produce a microarray. Several bioinformatic tools have been optimised to search for candidate matching sequences, to validate each pairwise alignment, to attempt multiple alignments and to define a consensus sequence. The results of these processes are stored in a MySQL database as automatic annotation, which can be accessed by means of a user friendly web interface allowing a very easy manual analysis and manual annotation of each transcript. An expert operator can easily annotate more than 100 transcripts per day, finding the best description, possible alternative starting, termination and splicing sites, polymorphisms, gene duplications and more. Sequencing artefacts such as chimeric clones can also be easily identified. This system has been used to annotate the muscle transcripts included in the Muscle-TRAIT database.

Integration of information

The main feature of our system is the integration of information from a variety of sources. In particular each transcript is relationally linked to a series a data including expression level, genomic sequence, intron/exon organisation, chromosomal localisation, protein sequence, protein domains, orthologue genes and other information.

Accessing and retrieval of data from Muscle-TRAIT

The normal access to Muscle-TRAIT is by means of SQL queries, which may be familiar to database programmers but less so to other people. Therefore we have developed a user-friendly interface that can be accessed at http://muscle.cribi.unipd.it. Muscle-TRAIT can be searched either for sequence similarity or for features. In the latter case, it is possible to perform very complex queries in a very simple and efficient way, mixing different arguments. For instance we can select for transcripts involved in the biological process of "actin filament organization", encoded by a gene located on chromosome 4, containing the word "binding" in the description, with an homologous gene in Drosophila but not in Yeast. A further option allows to make queries based on protein features. This option is based on a systematic search for known protein motifs and domains that is done on every transcript of Muscle-TRAIT, including putative transcripts obtained by Genscan predictions.