An integrated knowledge-base of gene expression in human skeletal muscle

Nicola Cannata, Rosario Dioguardi, Paolo Fontana, Paolo Scannapieco, Stefano Toppo, Gerolamo Lanfranchi, Giorgio Valle

CRIBI Biotech Centre - Università di Padova

We have build a solid scaffolding that can hold and connect muscle transcript sequencing data to functional data, expression profiles, genomic sequences and genetic diseases. The starting point is the wide collection of skeletal muscle ESTs produced at CRIBI, which are automatically analysed, filtered and stored in a SQL table (HSPD-EST). A schematic view of the organization of the data is shown in the figure.

ESTs are assembled into clusters (HSPD-CLUSTER table), which are very transitory entities as they may change at every new assembly depending on the order that the ESTs were merged or on the presence of new variant isoforms determined by alternative splicing or paralogue genes. On the other hand, many transcripts have now been well characterised and therefore should be considered as stable entities. Therefore, we decided to implement a Transcript Integrated Table (TRAIT) of human skeletal muscle, that includes some of the established information that is already available.

As can be seen in the figure, we have also implemented a Single-Transcript Integrated Table (STRAIT), where different transcripts are stored in different records, even if they come from the same gene, for instance after alternative splicing. Therefore, every single transcript is recorded in STRAIT, while TRAIT is used to link together those transcripts that originated from the same gene. When a new cluster is discovered, then a provisional STRAIT record is automatically created. Records become permanent after the addition of further information such as full length sequencing, functional studies and high density hybridisation experiments, which are currently performed in our laboratory. All the above information is organised under an SQL database management system, in a protected intranet environment, currently including more than 4,000 STRAIT records. All the tables are periodically translated into SRS databases and are accessible on the web at HYPERLINK "http://grup.bio.unipd.it/" .

The full implementation of the other databases (shown in the figure in light blue) is currently under way. In particular, a series of scripts and automatic procedures have been developed, linking full and partial transcripts to genomic sequences in view of the release of the entire human genome sequence. Our scripts make use of programs such as Blast, GeneFinder and Sim4, to perform this analysis systematically on every transcript of our database. The identification of the genomic sequence allows a simple and exact localisation of the genes and gives an indication of the full length sequence, introns, exons, alternative splicing and promoter region. Similar systematic procedures are also under way to link our muscle transcripts to sequences from model organisms such as yeast, C. elegans, Drosophila and mouse.