BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 5 2003, pages 661–662
DOI: 10.1093/bioinformatics/btg045

TRAIT (TRAnscript Integrated Table): a knowledgebase of human skeletal muscle transcripts

Stefano Toppo, Nicola Cannata, Paolo Fontana, Chiara Romualdi, Paolo Laveder, Emanuela Bertocco, Gerolamo Lanfranchi and Giorgio Valle**

CRIBI, Università di Padova, via Ugo Bassi 58/B, Padova, 35131 Italy

Received on September 3, 2002; revised on November 11, 2002; accepted on November 22, 2002

ABSTRACT

Summary: TRAIT is a knowledgebase integrating information on transcripts with related data from genome, proteins, ortholog genes and diseases. It was initially built as a system to manage an EST-based gene discovery project on human skeletal muscle, which yielded over 4500 independent sequence clusters. Transcripts are annotated using automatic as well as manual procedures, linking known transcripts to public databases and unknown transcripts to tables of predicted features. Data are stored in a MySQL database. Complex queries are automatically built by means of a user-friendly web interface that allows the concurrent selection of many fields such as ontology, expression level, map position and protein domains. The results are parsed by the system and returned in a ranked order, in respect to the number of satisfied criteria.

Availability: http://muscle.cribi.unipd.it and http://muscle.cribi.unipd.it/features/querystrait.html

Contact: stefano@cribi.unipd.it; giorgio.valle@unipd.it

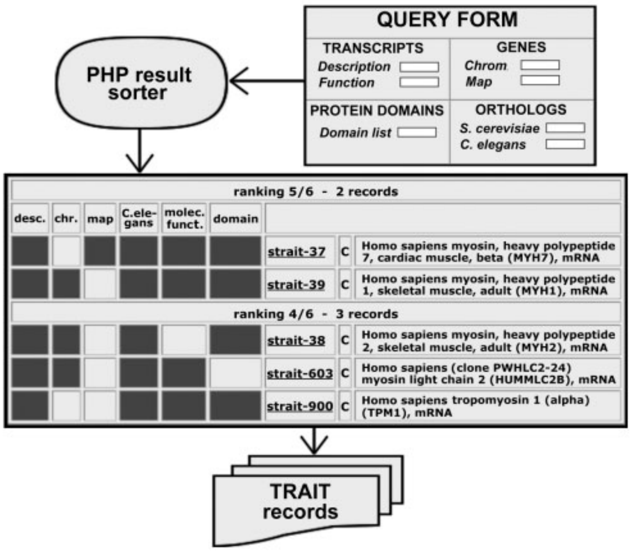
The acquisition and integration of data from the human genome project is having a great impact on biomedical research. Unfortunately, due to the huge amount and wide variety of available information, access to specific knowledge is often difficult without dedicated tools. Current retrieval systems of large databases, based on free text search such as Entrez and SRS still suffer from the difficulty in refining the query in order to lower the background-noise hits. However, if the queries are too complex the frustrating ‘no hits found’ is returned and no suggestions are given on how to modify the search to retrieve useful information. Muscle-TRAIT (TRAnscript Integrated Table) was created to answer these problems.

Our research group is working on the identification of genes expressed in human skeletal muscle. The aim of original project was the discovery of unknown genes by sequencing Expressed Sequence Tags (ESTs) from specifically designed cDNA libraries restricted to 400–500 bases at the 3′-end (Lanfranchi *et al.*, 1996). Our work led to the identification of about 4560 independent muscle transcripts reflecting the expression profile of skeletal muscle.

In order to manage our data and to accomplish high quality annotation of the transcripts, the function of which is in many cases still unknown, we developed several bioinformatic tools performing the following main tasks: (1) sample tracking and automatic quality control based on Phred (Ewing and Green, 1998); (2) construction of consensus sequences from clustered ESTs, using a strategy based on Blast to identify candidate sequences to include in a cluster and Cap3 (Huang and Madan, 1999) to perform the assembly and produce the consensus sequences; (3) annotation of the data and integration with the information available in public databases.

The annotation process is also a multi-step procedure. The first step is the periodic download of new transcripts from LocusLink (Pruitt and Maglott, 2001). The second is the automatic managing of annotations associated with the imported sequences, including Gene Ontology (http://www.geneontology.org), which are stored in internal MySQL tables. The third step is linking TRAIT to LocusLink, which is performed automatically by means of a combination of programs including Megablast, Fasta and Phrap. Finally, the fourth step is carried out by expert annotators, with the aid of a web interface that allows to view and manually refine the alignments and their associated references. TRAIT entries are divided into four classes from ‘A’ to ‘D’ depending on the level of certainty. ‘A’ (Auto) are novel or unknown sequences that have been generated and annotated automatically by our system and are likely to change without notice; ‘B’ (Blocked) are entries that were in class ‘A’, but have been blocked by an authorised annotator; any automatic change will only be possible after authorization from the annotator; ‘C’ (Confirmed) means that the sequence is either manually

*To whom correspondence should be addressed.



ranking 5/6 - 2 records								
desc.	chr.	map	C.ele-gans	molec. funct.	domain			
■		■	■	■	■	strait-37	C	Homo sapiens myosin, heavy polypeptide 7, cardiac muscle, beta (MYH7), mRNA
■	■		■	■	■	strait-39	C	Homo sapiens myosin, heavy polypeptide 1, skeletal muscle, adult (MYH1), mRNA
ranking 4/6 - 3 records								
■	■		■		■	strait-38	C	Homo sapiens myosin, heavy polypeptide 2, skeletal muscle, adult (MYH2), mRNA
■	■		■	■		strait-603	C	Homo sapiens (clone PWHLC2-24) myosin light chain 2 (HUMMLC2B), mRNA
■			■	■	■	strait-900	C	Homo sapiens tropomyosin 1 (alpha) (TPM1), mRNA

Fig. 1. Retrieval process and results from a TRAIT query using the following six criteria: (1) *description* must contain 'myosin'; (2) genes located on *chromosome* '9' or '16' or '17'; (3) *cytogenetic map* '14q12'; (4) putative orthologs in *C. elegans* with a Blast E-value $\leq 10^{-30}$; (5) transcripts involved in *molecular function* of 'muscle motor' or 'structural protein of muscle' (selectable keywords from the menus of Gene Ontology sub-section); (6) *protein domains* 'Myosin_tail' or 'Tropomyosin'. The coloured cells of the summary indicate which searching criteria were satisfied. There are no hits satisfying all six criteria, however, two and three records respectively satisfy five and four criteria.

annotated by our group or matches a sequence annotated in LocusLink or other databases; 'D' (Definitive) means that the transcript is described in the scientific literature.

When a full length cDNA sequence is not available, a protocol based on Blast, Sim4 (Florea *et al.*, 1998) and GeneScan (Burge and Karlin, 1997) is used to predict the coding and protein sequences. Thus, most entries have an associated protein sequence either experimentally confirmed or putative. Protein sequences are then used to perform BlastP comparison against *S. cerevisiae*, *C. elegans* and *D. melanogaster* sequences. Furthermore, protein domains are searched against Pfam (Bateman *et al.*, 2002) using HMMER. These modules gather different types of information that is stored in MySQL tables and can be used to perform complex queries.

The information retrieval described here is available by clicking the 'features' button on the home page (http://muscle.cribi.unipd.it). The TRAIT query form is simple, yet it is possible to make complex queries. Most fields are entered with the help of menus that only allow the definition of existing instances. The query form is divided into four main sections: transcripts, genes, proteins and orthologs (Fig. 1). Currently, 22 fields are available in the form and can be selected in any number and combination, although only seven are shown in the simplified diagram of Figure 1. The selection is automatically translated into SQL language, using OR operators to merge the searching criteria defined for each field. The resulting query is used to select suitable records from a TRAIT 'view', where all the original tables are merged into a single one, thus overcoming the difficulty of making 'joins' using MySQL.

The results are processed by a PHP script that sorts the hits on the basis of the number of satisfied criteria. For instance, if six criteria were defined in the query form and no entries satisfies them all, then instead of returning 'no hits found' the PHP script will show all the entries satisfying at least five criteria, indicating which of the criteria were satisfied (Fig. 1).

TRAIT records contain several links to gene descriptions, graphic display of human genome context, putative transcript variants, genetic diseases associated to the cytogenetic map location and to GENATLAS (Frezal, 1998). The database contains 4560 transcripts derived from the annotation of 28 893 ESTs obtained from human skeletal muscle (HSPD), 35% of which are still in class 'A'. Recently new ESTs from leukemia (LKPD), blood (BLPD) and heart (CMPD) have been added, bringing the total number of independent entries to 7516.

The authors acknowledge the financial support of Fondazione Telethon ONLUS, grant B57.

REFERENCES

Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Frezal,J. (1998) Genatlas database, genes and development defects. *C. R. Acad. Sci. III*, **321**, 805–817.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Lanfranchi,G., Muraro,T., Caldara,F., Pacchioni,B., Pallavicini,A., Pandolfo,D., Toppo,S., Trevisan,S., Scarso,S. and Valle,G. (1996) Identification of 4370 expressed sequence tags from a 3′-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. *Genome Res.*, **6**, 35–42.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.