GoSh: a goat and sheep ESTs database

A. Caprera[2], B. Lazzari[2], A. Stella[2], I. Merelli[3], A. R. Caetano[4], P. Mariani[1]

[1] Livestock Genomics Unit. Parco Tecnologico Padano, Lodi, Italy
[2] Statistical Genetics and Bioinformatic Unit. Parco Tecnologico Padano, Lodi, Italy
[3] Istituto di Tecnologie Biomediche. Consiglio Nazionale delle Ricerche, Segrate, Italy
[4] Embrapa Recursos Geneticos e Biotecnologia. Brasilia, Brasil

Corresponding author: Andrea Caprera. Parco Tecnologico Padano. Via Einstein, Polo Universitario, Località Cascina Codazza, 26900 Lodi, Italy - Tel. +39 037 14662693 - Fax: +39 037 14662349 - Email: andrea.caprera@tecnoparco.org

ABSTRACT: The GoSh database (http://www.itb.cnr.it/gosh/) is an online resource including expressed sequence tags (ESTs) from *Ovis aries* and *Capra hircus*. A total of 58,990 sheep and goat sequences were downloaded from GenBank and processed by a semi-automated pipeline, integrating public programs and Perl scripts. Data were collected in a MySQL database, which can be queried via a PHP-based web interface. Sequences were assembled and a unigene dataset was defined. Three annotation procedures were carried out on all the EST sequences and all the contig consensus sequences. A procedure was also implemented to infer statistical classification among Gene Ontology (GO) categories from the ontology occurrences related to the sequences included in the database. A number of programs were used to extract features and give significance to rough sequences. Among these, AutoSNP was used to perform putative SNP detection. Further analyses were performed on the GoSh db dataset, including tandem repeats search and protein patterns identification. The web interface allows users to retrieve significant data and correspondent external links and to download selected sequences and accessory information in different formats. The resulting web site is a resource of data and links related to goat and sheep expressed genes.

Key words: Database, EST, Goat, Sheep.

INTRODUCTION - Expressed sequence tags (EST) represent one of the most important sources of information in gene expression studies. Nevertheless, it is not always simple to extract information from huge amounts of sequences without the help of dedicated bioinformatic resources. The main function of web databases based on specific EST sequence collections is to give an overview on gene expression features in the various tissues/organs and/or to allow comparisons of similarity among related organismsThe main function of web databases based on specific EST sequence collections is to give an overview on gene expression levels in the various tissues/organs and/or to allow comparisons among related organisms. Furthermore, the possibility to perform searches on sequence annotations is a powerful tool to deeply investigate expressed genes distribution. Aim of this work was to create a comprehensive database dedicated to goat and sheep, encompassing data derived from sequence analysis as well as links to related online resources. Particular care was given to the preparation of the web interface, that allows performing complex queries on the database in a simple and user-friendly way.

MATERIAL AND METHODS - An analysis pipeline integrating public programs with in-house developed Perl scripts was created to process 637 *Capra hircus* and 58,353 *Ovis aries* sequences downloaded from GenBank. Input data, as well as data produced during the pipeline steps, are automatically stored in a MySQL database. The AutoSNP algorithm (Barker *et al.*, 2003), version 7, was used for putative SNP detection. As TGICL (Pertea *et al.*, 2003) and CAP3 (Huan 1999) are integrated in the AutoSNP procedure, sequence clustering and assembly were performed by AutoSNP, as well, and a number of accessory scripts was prepared to extract assembly information from the SNP detection procedure intermediate steps. The CAP3 parameters were set to –p 95 –o 100 and 5,462 contigs were generated from the assembly procedure. A unigene dataset of 20,784 sequences, encompassing all the

singlets and the longest sequence of each contig, was defined. All the EST sequences, as well as all the contig consensus sequences, were annotated against three different databases. Two of these annotation procedures were carried out by BLASTx (Altschul *et al.*, 1990), the former versus the GenBank nr database (referred to as 'NCBI blast' in the cited web interface), and the latter versus the UniProtKB database (http://www.ebi.ac.uk/uniprot/) (referred to as 'GO blast' in the cited web interface). A third supplementary annotation was performed by BLASTn against an in-house prepared database encompassing 5,474 goat and sheep genomic sequences downloaded from GenBank (named 'genomic blast' in the cited web interface).
The most probable polypeptide sequence was inferred from each EST and from each contig consensus sequence with FrameFinder (http://bioweb.pasteur.fr/docs/man/man/ESTate.1.html) and the resulting polypeptides were compared to the PROSITE database (Falquet *et al.*, 2002) with scanPROSITE (Gattiker *et al.*, 2002) for homologous protein patterns and motifs identification. Tandem Repeats Finder (Benson, 1999) was used to identify repeats in the sequences contained in the database. Significantly homologous NCBI blast hits were scanned for the presence of EC numbers. When present, these were used to retrieve links to the ExPASy NiceZyme pages (http://au.expasy.org/) and to the KEGG pathways database (http://www.genome.jp/kegg/pathway.html). Based on data contained in the database of associations among proteins and GO elements (www.ebi.ac.uk/GOA), a Perl script was developed to relate GoSh GO blast best blast hits to Gene Ontology categories (The Gene Ontology Consortium, 2000). Matching ontologies were stored into the database, and statistics for the ontologies occurrences are dynamically created upon user's request. A PHP-based web interface was set up to manage the incoming queries, as well as the preparation of graphical outputs (Figure 1).

Figure 1. the GoSh web interface: a page dedicated to the presentation of a single contig.

The text search utility and the local blast interface, that allows users to blast their own sequences against either the nucleic GoSh db dataset or the derivative putative protein database, represent the two main accession points to data stored in the database.

Table 1. The GoSh database status.

Total nr of sequences	58,990
Average Base Count	547.79
Number of singletons	15,322
Number of contigs	5,462
Number of putative unigenes	20,784
Percentage of 'NCBI blast' annotated sequences	79.55
Percentage of 'GO blast' annotated sequences	76.79
Percentage of 'genomic blast' annotated sequences	12.44
Percentage of SNP-containing contigs	14.94
Percentage of repeats-containing sequences	7.25

Tabular presentations of sequences and their features are also given in the Sequence/Contig report pages, where links are provided to related information internal or external to the GoSh database. A page describing the processing, assembly and annotation procedures that were used to produce data stored in the GoSh database was included in the web site, as well as a detailed help page. Statistics describing the status of the GoSh database are given in Table 1.

RESULTS AND CONCLUSIONS – The GoSh database represents, at the moment, the most complete repository of EST sequences dedicated to goat and sheep. Aim of the Authors was to set up a flexible bioinformatic procedure that could easily fit different purposes. The GoSh pipeline has a modular structure and accessory programs can be integrated or removed with little manual work. The GoSh MySQL database stores all the data produced by the pipeline steps and can be considered a powerful and dynamic structure where new sequences and related information can be added. Access to data stored in the database is granted by the web interface. Queries can be performed on the whole dataset or on logical sequence subsets (i.e. unigenes). Particular care was taken to allow data retrieval according to user's specific needs, and all query outputs can be downloaded both in multifasta and tabular format. All sequences, contig consensus sequences and putative SNP reports, together with complete outputs from the assembly procedure, are also downloadable in different formats from the download page of the web interface.
The GoSh db proved to be very useful for the selection of sequences to be used for specific purposes (i.e. SNP-mapping) and it is currently implemented by the Authors in a SNP validation and biodiversity study. In particular, the possibility to use the dataset performing subset-specific text searches offers an efficient tool for data mining and retrieval.

The Authors want to thank Dr. Dave Edwards and Dr. Gary Barker for providing the AutoSNP program and Dr. Luciano Milanesi, CNR-ITB and L.I.T.B.I.O. for informatics support.

This work was supported by Fondazione CARIPLO (n° 2003.0721/11.8094).

REFERENCES - **Altschul**, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410. **Barker**, G., Batley, J., O' Sullivan, H., Edwards, K.J. and Edwards, D., 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 12:19(3):421-422. **Benson**, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acid Res. 27:573-580. **Falquet**, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A., 2002. The PROSITE database, its status in 2002. Nucleic Acids Res. 30:235-238. **Gattiker**, A., Gasteiger, E., and Bairoch, A., 2002. ScanPROSITE: a reference implementation of a PROSITE scanning tool. Appl. Bioinformatics 1:107-108. **Huan**, X. and Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Research 9:868-877. **Pertea**, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19(5):651-652. **The Gene Ontology Consortium** (2000) Gene Ontology: tool for the unification of biology. Nature Genet. 25:25-29.