BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 8 2007, pages 1043–1045
doi:10.1093/bioinformatics/btm063

Databases and ontologies

GoSh: a web-based database for goat and sheep EST sequences

Andrea Caprera[1,*], Barbara Lazzari[1], Alessandra Stella[1], Ivan Merelli[2], Alexandre R. Caetano[3] and Paola Mariani[1]

[1]Parco Tecnologico Padano, Via Einstein – Località Cascina Codazza, 26900 Lodi, [2]Istituto Tecnologie Biomediche, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy and [3]Embrapa Recursos Geneticos e Biotecnologia Parque Estacao Biologica, Final Av. W/5 Norte, Brasilia-DF, C.P. 02372 70770-900, Brasil

Received on October 25, 2006; revised on February 5, 2007; accepted on February 16, 2007

Advance Access publication March 24, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The GoSh database is a collection of 58 990 *Capra hircus* and *Ovis aries* expressed sequence tags. A perl pipeline was prepared to process sequences, and data were collected in a MySQL database. A PHP-based web interface allows browsing and querying the database. Putative single nucleotide polymorphism (SNP) detection, as well as search to repeats were performed, and links to external related resources were provided. Sequences were annotated against three different databases and an algorithm was implemented to create statistics of the distribution of retrieved homologous ontologies in the Gene Ontology categories. The GoSh database is a repository of data and links related to goat and sheep expressed genes.

Availability: The GoSh database is available at http://www.itb.cnr.it/gosh/

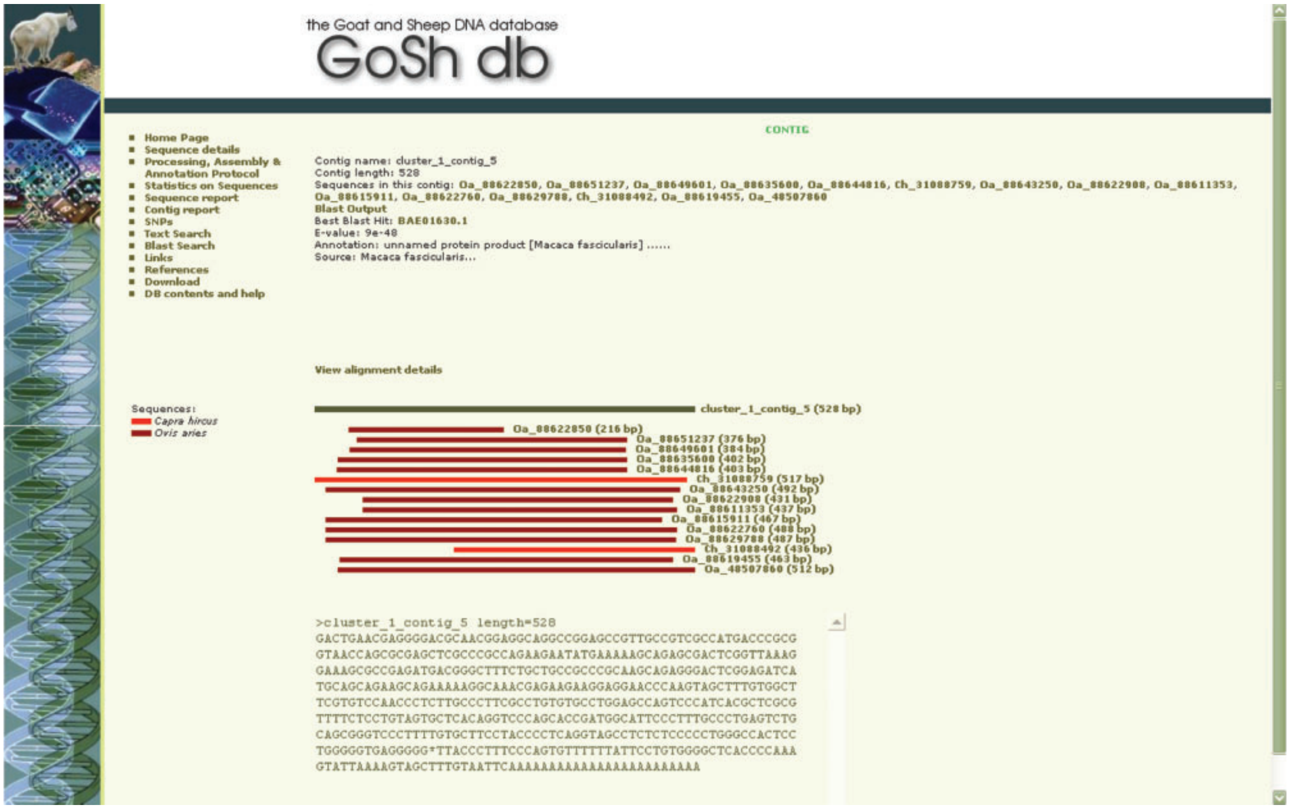
Contact: andrea.caprera@tecnoparco.org

Expressed sequence tags (EST) represent one of the most important sources of information in gene expression studies, however, raw EST sequences must be annotated before they are of value to the research community. The main function of web resources based on organism-specific collections is to provide an overview on the distribution of expressed genes in the various tissues/organs or in different developmental stages. The aim of the present 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.

An analysis pipeline integrating public programs with Perl scripts was created to process 637 goat (*Capra hircus*) and 58 353 sheep (*Ovis aries*) sequences downloaded from GenBank. The AutoSNP program, version 7 (Barker *et al*., 2003), was used to identify putative single nucleotide polymorphisms (SNPs) in assembled sequences. AutoSNP integrates TGICL (Pertea *et al*., 2003) that performs sequence clustering and in turn integrates CAP3 (Huan and Madan, 1999) for contig assembly. Data obtained during the assembly procedure were used to establish a unigene dataset of 20 784 sequences. 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 7437 goat and sheep genomic sequences downloaded from GenBank (named 'genomic blast' in the cited web interface). All the annotations will be periodically updated and the last update is indicated at the bottom of the Home Page of the GoSh web site. Analysis of repeats was performed with Tandem Repeats Finder (Benson, 1999). FrameFinder (http://bioweb.pasteur.fr/docs/man/man/ESTate.1.html) was used to infer the most probable polypeptide sequence from each EST and contig consensus sequence. For this purpose, a word probability file was prepared from 45 753 bovidae mRNA sequences downloaded from the GenBank CoreNucleotide database. The derived polypeptide sequences were compared to the PROSITE database (Falquet *et al*., 2002) with scanPROSITE (Gattiker *et al*., 2002) to retrieve homologous protein patterns and domains. 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 to allow a successive dynamic creation of statistics for the ontology occurrences. Links to the AmiGO web pages (http://www.godatabase.org/) were associated to GO terms in the GO Statistics pages. Full details on programs and parameters used in the GoSh database (db) pipeline are available at the Processing, Assembly and Annotation page of the web site.

The Perl pipeline has a modular structure and fills in tables of a MySQL database. The web interface is based on the PHP

[*]To whom correspondence should be addressed.



Contig name: cluster_1_contig_5
Contig length: 528
Sequences in this contig: Oa_88622850, Oa_88651237, Oa_88649601, Oa_88635600, Oa_88644816, Ch_31088759, Oa_88643250, Oa_88622908, Oa_88611353, Oa_88615911, Oa_88622760, Oa_88629788, Ch_31088492, Oa_88619455, Oa_48507860
Blast Output
Best Blast Hit: BAE01630.1
E-value: 9e-48
Annotation: unnamed protein product [Macaca fascicularis] ......
Source: Macaca fascicularis...

```
>cluster_1_contig_5 length=528
GACTGAACGAGGGGACGCAACGGAGGCAGGCCGGAGCCGTTGCCGTCGCCATGACCCGCG
GTAACCAGCGCGAGCTCGCCCGCCAGAAGAATATGAAAAAGCAGAGCGACTCGGTTAAAG
GAAAGCGCCGAGATGACGGGCTTTCTGCTGCCGCCCGCAAGCAGAGGGACTCGGAGATCA
TGCAGCAGAAGCAGAAAAAGGCAAACGAGAAGAAGGAGGAACCCAAGTAGCTTTGTGGCT
TCGTGTCCAACCCTCTTGCCCTTCGCCTGTGTGCCTGGAGCCAGTCCCATCACGCTCGCG
TTTTCTCCTGTAGTGCTCACAGGTCCCAGCACCGATGGCATTCCCTTTGCCCTGAGTCTG
CAGCGGGTCCCTTTTGTGCTTCCTACCCCTCAGGTAGCCTCTCTCCCCCTGGGCCACTCC
TGGGGGTGAGGGGG*TTACCCTTTCCCAGTGTTTTTTATTCCTGTGGGGCTCACCCCAAA
GTATTAAAAGTAGCTTTGTAATTCAAAAAAAAAAAAAAAAAAAAAAAA
```

Fig. 1. The GoSh database web interface: an example of a single contig page.

language, and manages all the incoming queries as well as all the graphical outputs dynamic creation (Fig. 1).

Three interfaces were set up to query the database and retrieve user-defined sequence subsets: the sequence/contig report tables, the Text Search utility and the Local Blast utility. The sequence/contig report tables report all significant information related to each sequence/contig. Details on features, BLAST outputs or contig alignments can be accessed via the links given in these tables or via the single sequence/contig pages that were assigned to each sequence/contig. The text search utility is structured to perform combined keyword searches on all the fields of the Sequence/Contig report tables. Searches can be restricted to sequence subsets (goat/sheep, unigenes/not unigenes, singlets/contig-related sequences, putative SNP- or repeats-containing sequences). Query outputs can be downloaded both as multifasta sequence files and in tabular format. The local blast interface allows users to blast their own sequences against either the nucleic GoSh db dataset or the derivative putative protein database.

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 data retrieval.

The modular nature of the Perl pipeline and of the PHP-based web interface components, allowed using the same structures for different projects, adding or removing modules to perform sequence analyses according to the needs of different datasets (Lazzari *et al*., 2005 and unpublished data). Other online databases that were produced by the authors are based on similar sequence analysis pipelines, but some of the features that are presented in this article were created on purpose for the GoSh database. This is the case, for example, of the package of scripts that were prepared to infer statistics on Gene Ontologies from GO annotations. If applicable, new modules are transferred to the other databases, giving the authors the opportunity to maintain all the databases upgraded and to improve information related to the different datasets they represent.

ACKNOWLEDGEMENTS

This work was supported by Fondazione CARIPLO (N° 2003.0721/11.8094) and European Specific Support Action Bioinfogrid (FP6-026808). We acknowledge Dr Dave Edwards and Dr Gary Barker for providing the AutoSNP program, Dr Luciano Milanesi, CNR-ITB and L.I.T.B.I.O. for informatics support and Prof. Francesco Salamini for critical reading of the manuscript. Funding to pay the Open Access publication charges was provided by Parco Tecnologico Padano.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Barker,G. *et al*. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acid Res*., **27**, 573–580.

Falquet,L. *et al*. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res*., **30**, 235–238.

Gattiker,A. *et al*. (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 Res*., **9**, 868–877.

Lazzari,B. *et al*. (2005) ESTree db: a tool for peach functional genomics. *BMC Bioinformatics*, **6** (Suppl. 4:S16), PMID: 16351742.

Pertea,G. *et al*. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet*., **25**, 25–29.