Brazilian Journal of Medical and Biological Research (2005) 38: 1571-1574
ISSN 0100-879X Short Communication

The GATO gene annotation tool for research laboratories

A. Fujita[1,2]*, K.B. Massirer[1]*, A.M. Durham[2], C.E. Ferreira[2] and M.C. Sogayar[1]

[1]Instituto de Química, [2]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brasil

Abstract

Large-scale genome projects have generated a rapidly increasing number of DNA sequences. Therefore, development of computational methods to rapidly analyze these sequences is essential for progress in genomic research. Here we present an automatic annotation system for preliminary analysis of DNA sequences. The gene annotation tool (GATO) is a Bioinformatics pipeline designed to facilitate routine functional annotation and easy access to annotated genes. It was designed in view of the frequent need of genomic researchers to access data pertaining to a common set of genes. In the GATO system, annotation is generated by querying some of the Web-accessible resources and the information is stored in a local database, which keeps a record of all previous annotation results. GATO may be accessed from everywhere through the internet or may be run locally if a large number of sequences are going to be annotated. It is implemented in PHP and Perl and may be run on any suitable Web server. Usually, installation and application of annotation systems require experience and are time consuming, but GATO is simple and practical, allowing anyone with basic skills in informatics to access it without any special training. GATO can be downloaded at [http://mariwork.iq.usp.br/gato/]. Minimum computer free space required is 2 MB.

Correspondence
M.C. Sogayar
Instituto de Química, USP
Av. Prof. Lineu Prestes, 748
05508-900 São Paulo, SP
Brasil
Fax: +55-11-3091-3820
E-mail: mcsoga@iq.usp.br

Researh supported by FAPESP, CNPq, FINEP, and PRP-USP.

*These authors contributed equally to this study.

Received July 7, 2004
Accepted June 27, 2005

Key words
- Gene annotation tool
- (GATO)
- Automatic annotation
- Pipeline
- Bioinformatics

Gene annotation is important to assign a set of characteristics to gene sequences and thus obtain relevant information for the interpretation of experimental results (1-3). Automatic annotation is important because it provides complete and homogeneous gene characterization to researchers. However, it is also important that the researcher be allowed to manually complement and correct information in order to update and determine the accuracy of the data set. In addition, visualization of data obtained by project collaborators and by complementary experiments for a given biological model is crucial to accelerate the interpretation of the results and their publication.

The gene annotation tool (GATO) program [http://mariwork.iq.usp.br/gato/] was designed to allow researchers with no computer science skills to annotate gene sequences by accessing information from numerous databases and Bioinformatics tools and to make their results easily available to others. To facilitate easy data viewing, GATO displays Web-based results in HTML format. This system is supported by a MySQL database which permits easy search, in addition to updating and data storage.

The GATO system is composed of three graphical client-server interfaces, namely: data submission, annotation output and stored information search, each of which can be accessed through the Web.

When the user submits data, GATO's interface allows him to select the experimental conditions associated with data generation from a list that can be adapted to each case or laboratory. The program's overall structure is similar to ASAP [http://bioinformatics.fccc.edu/software/OpenSource/ASAP/ASAP.shtml] (2) and, as in Genescript [http://tcag.bioinfo.sickkids.on.ca/genescript/], GATO uses several publicly available analysis programs to annotate and generate the output (4). GATO's editor, in which the annotator may edit the annotation manually, is simpler than Apollo [http:/www.fruitfly.org/annot/apollo/] (5) by providing a text window and by keeping a reliable history of the annotation.

The stored information search interface is designed to be a user friendly page that hides the MySQL database, and allows the annotator to select search fields from selection boxes, such as the downloadable dbERGE II (6), which has the interesting feature of presenting a link to the corresponding PubMed entry, rather than overcrowding the display page.

GATO's input data may be either in the form of chromatograms or FASTA files corresponding to experimentally obtained sequences. Sequences are submitted in a format containing additional information related to experimental conditions and results. Chromatograms are accepted when a minimum of 100 continuous bases display quality levels higher than 20 on Phred [http://www.phrap.org/] (7,8). FASTA files can be submitted in the case of EST assemblies or when a chromatogram is not available. Then Crossmatch is used to screen out vectors in the FASTA file. This last step is essential for reliable similarity searches.

The nucleotide sequences are automatically characterized through searches in publicly available databases. The first step is identification of the sequence using the best BLASTN (9) hit against a database specified in the configuration file. Alignment is then performed against the genome of the organism that was previously chosen on the submission page using BLAT [http://genome.ucsc.edu/cgi-bin/hgBlat?command=start] (10). The main aim of this BLAT search is to find orthologs to the gene of interest.

The sequence is then characterized using LocusLink and PFAM databases. LocusLink returns descriptive information about gene loci, including related diseases, chromosome position and nomenclature (11). PFAM is accessed by a program called HMMPFAM, based on hidden Markov models, which evaluates protein domain profiles. Due to the difficulty in performing automatic functional characterization searches, the annotation system also includes links to other public databases and Bioinformatics tools, including Gene Ontology, Gene Finder and GDB (The Genome Data Base).

For manual annotation, GATO also generates interactive HTML graphical views (Figure 1), which allow the user to browse the results of the analysis and input additional information that can be used to confirm the automatic annotation and allow new deductions. In addition, the system also permits the introduction of corrections and the updating of previous annotation results by the user, without the need to rerun the whole annotation process. This cannot usually be done with annotation tools that are run on-line. This system uses a MySQL database to store the input sequence file and specific information related to each returned result, structured as illustrated in Figure 2. One problem in annotation is that information is updated in public databases as the result of the discovery and characterization of new sequences. This causes the annotation to become outdated (12). To solve this problem, the GATO database was designed to store new versions of the annotation and to include the case history, which is helpful to users, allowing them to determine when and how the annotation was carried out. GATO is important because it decreases the need for repetitive Bioinformatics operations and it provides reliable information. Minor new

information can easily be added by manual annotation, whereas the complete description of a gene is better carried out by re-running the tool.

Our GATO detailed submission page provides specific information on experimental conditions and results which, for example, can be a very useful feature to design microarray experiments. The initial evaluation of sequence quality is important to avoid errors in sequence alignment and thus identify the right gene with confidence. In case several sequences are to be annotated, there is a program called GATOALL, which accepts a text file containing multiple sequences in FASTA format and automatically performs all the annotation. This prevents the annotator from submitting one sequence at a time. The results are stored in the database for further analysis using the Web graphical views. Thus, GATO permits both kinds of annotation, namely: large scale using GATOALL and individual sequence annotation using the GATO Web interface.

We used the GATO annotation system to annotate 697 human genes derived from large EST sequencing projects with theoretically unknown function (genes with prefixes such as LOC, KIAA, MCG, FLJ, and DKFZP). GATO was very useful to update information since ~10% of these genes are already associated with a gene symbol representing their function. One of 697 genes was identified as a mouse gene but this occurred because of the use of an outdated local BLAST database [http://www.ncbi.nlm.nih.gov/blast/]. Therefore, the database version can be very important to obtain a good annotation. For the same set of genes, to which no known function had been associated, we found protein domain profiles, in addition to mapping these genes in the human genome (Figure 3), finding ~25% putative orthologs (sequences with identity higher than 90%) in the mouse genome. A summary of the Gene Ontology data found for this set of genes is presented in Table 1. These data are useful to design further experiments to probe into protein-protein interactions and

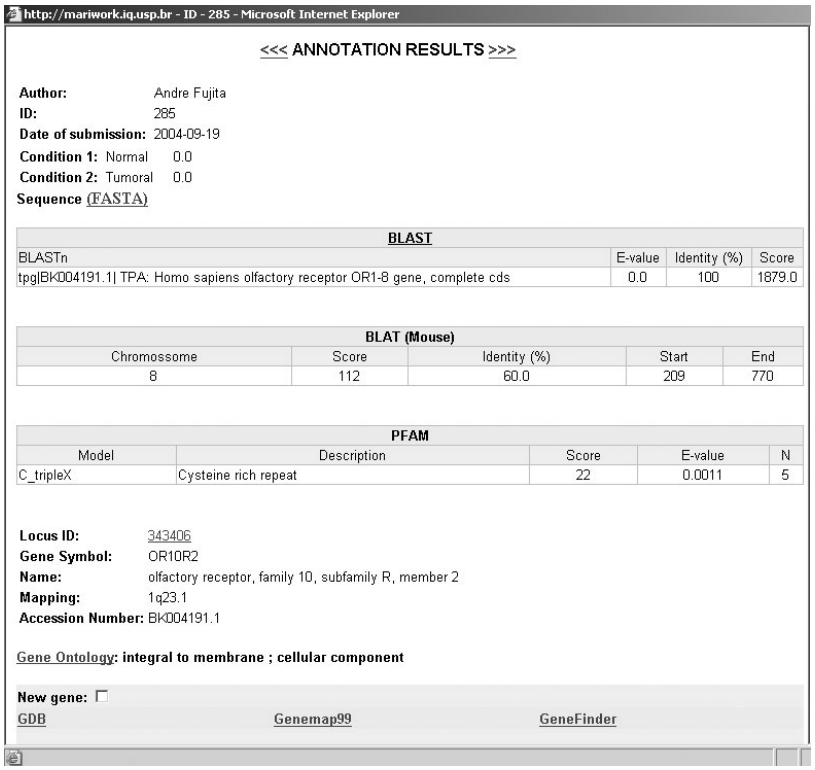

Figure 1. GATO output. A user friendly graphical view which facilitates data interpretation.

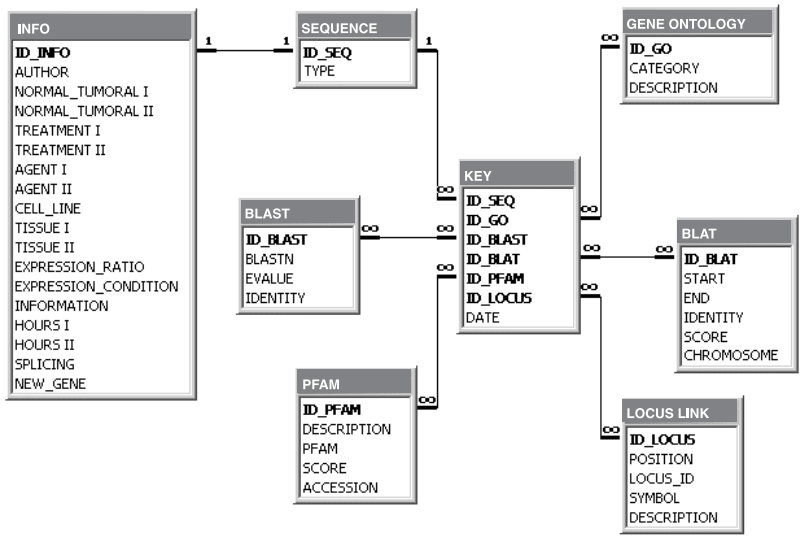

Figure 2. The structure of GATO. The database is composed of a table called KEY, which stores the case history. The INFO Table stores experimental conditions, and SEQUENCE stores the pointer to submitted file and the type, whether it is a chromatogram or a FASTA file. This structure facilitates the insertion of new tools and the maintenance of annotation versions.

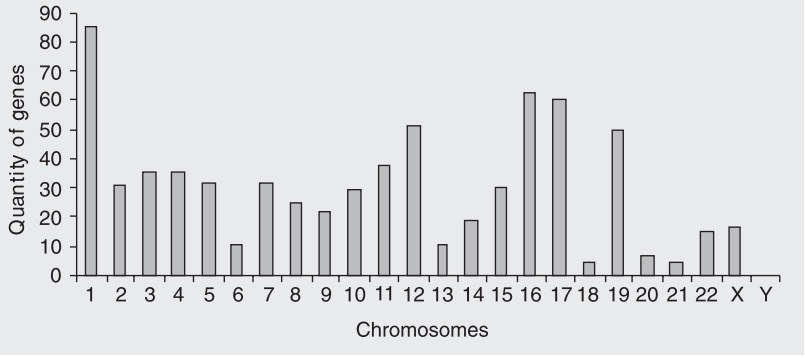

Figure 3. Number of the genes in the human chromosomes.

Table 1. Functional classification of the gene set and the percentage of the genes which belong to each Gene Ontology classification. "No GO" means the proportion of genes which did not have a Gene Ontology classification yet.

	%
Metabolism	79.1
Cellular physiological processes	15.8
Regulation of physiological processes	11.5
Cell communication	8.6
Response to a stimulus	7.2
Regulation of cellular processes	0.7
Regulation of viral protein levels	0.7
Morphogenesis	0.7
Reproduction	0.7
Organismal physiological processes	0.7
Sex determination	0.7
No GO	66.5

metabolic pathways.

The GATO system facilitates the use of numerous publicly available genomic tools and allows members of a research group to display their data in a complete and homogeneous manner. Research laboratories can use GATO as an aid for manual annotation, which is also useful in wide efforts of automated annotation. The program provides graphical overviews of all data required for annotation and integrates several commonly used tools and database searches. The fact that GATO has independent parsers for each specific tool and the modular structure of its database renders this system flexible and also extensible, unique properties which are important to allow other users to implement and improve this system. The basic steps for GATO installation are: to edit a very easy configuration file and to create the database by running a given script. As could be seen above, GATO is accessible to any researcher with basic concepts in informatics.

Acknowledgments

We would like to thank the members of the Cellular and Molecular Biology Laboratory, who contributed with suggestions for improvement of the GATO system.

References

1. Andrade MA, Brown NP, Leroy C et al. (1999). Automated genome sequence analysis and annotation. *Bioinformatics*, 15: 391-412.
2. Kossenkov A, Manion FJ, Korotkov E et al. (2003). ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics*, 19: 675-676.
3. Rutherford K, Parkhill J, Crook J et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16: 944-945 [http://www.sanger.ac.uk/Software/Artemis/].
4. Alexander KH, Joseph C, Andrew PB et al. (2003). Genescript: DNA sequence annotation pipeline. *Bioinformatics*, 19: 1177-1178.
5. Lewis SE, Searle SMJ, Harris N et al. (2002). Apollo: a sequence annotation editor. *Genome Biology*, 3: Research 0082, 1-14.
6. Elnitski L, Giardine B, Shah P et al. (2005). Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Research*, 33: 466-470.
7. Ewing B & Green P (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, 8: 186-194.
8. Ewing B, Hillier L & Wendl MC (1998). Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8: 175-185.
9. Altschul SF, Gish W, Miller W et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215: 403-410.
10. Kent WJ (2002). BLAT - the BLAST-like alignment tool. *Genome Research*, 12: 656-664.
11. Pruitt KD & Maglott DR (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29: 137-140.
12. Gaasterland T & Oprea M (2001). Whole-genome analysis: annotations and updates. *Current Opinion in Structural Biology*, 11: 377-381.