

# BioNotes: A System for Biosequence Annotation

Melissa Lemos, Luiz Fernando Bessa Seibel, Marco Antonio Casanova

{melissa, seibel, casanova}@inf.puc-rio.br

Pontifícia Universidade Católica do Rio de Janeiro  
Departamento de Informática  
Rua Marquês de S. Vicente, 225  
Rio de Janeiro, RJ - Brazil CEP 22453-900

## Abstract

One of the most important tasks of genome projects is the interpretation of experimental data in order to derive biological knowledge from the data. To achieve this goal, researchers typically search external data sources, execute analysis programs on the biosequences, analyze previous annotations and add new annotations to register their interpretation of the data.

This paper first elicits the functional requirements of biosequence annotation systems. Then, it describes BioNotes, a tool that meets these requirements, stressing the advantages it brings to the researchers in this area.

## 1 Introduction

One of the most important tasks of genome projects is the interpretation of experimental data in order to derive biological knowledge from the data.

As a consequence, the challenge of Bioinformatics is to create effective tools to help researchers mine large sets of biosequences, that is, sets of DNA or protein sequences. This involves many facets. The tools must help a researcher: explore and rapidly test his hypothesis; access annotations stored in public data sources to compare them with his results; execute analysis programs and browse through the annotations they automatically created; analyze current annotations, with the help of an appropriate interface, and manually generate new annotations.

Indeed, a researcher now has access to a rich set of data sources, as well as a variety of data analysis programs. After experimentally obtaining a set of biosequences, he may submit them to a public source, such as Genbank [1], or store them in his own repository. The researcher may annotate the biosequences to indicate, for example, the laboratory that sequenced them and details used during the sequencing procedure. In general, two of the most

important types of annotations are those that identify a gene in a genome and those that indicate the function(s) of a gene.

He may also submit the biosequences to data analysis programs to try to detect interesting biological properties of the biosequences. For example, GLIMMER [2] locates ORFs (putative genes) and tRNAScan [3] obtains transfer RNAs.

From the scenario just described, we may then classify annotations as *manual*, directly created by the researcher, *automatic*, generated by analysis programs, or *imported* from public data sources.

The characteristics of the annotations also vary according to the goal of the genome project. Indeed, annotations generated in the context of a project that targets the complete DNA of an organism have different requirements from those created in the context of a project whose goal is to obtain ESTs (expressed sequence tags), which is common for large genomes.

The goals of this paper are to elicit the functional requirements of biosequence annotation systems and to describe BioNotes, a tool that meets these requirements. BioNotes is under development at the Pontifical Catholic University of Rio de Janeiro. The tool is currently used by the Department of Biochemistry and Molecular Biology [4], Oswaldo Cruz Institute – FIOCRUZ, to annotate the genome of *Trypanosoma cruzi*, and by RioGene [5], to annotate the genome of the *Gluconacetobacter diazotrophicus*.

The paper is organized as follows. Section 2 presents the major functional requirements of annotation systems, based on an analysis of the major annotation systems available. Section 3 describes BioNotes. Finally, section 4 contains the conclusions and discusses future work.

## 2 Functional Requirements of Annotation Systems

The goal of annotation systems is to help researchers create, retrieve and analyze annotations that tag biosequences.

The functional requirements of annotation systems listed in this section result from a careful study of the following systems: Artemis[6], DAS[7], CeleraBrowser [8], EDITtoTrEMBL[9], GASP[10], GenDB [11], GeneMine [12], GeneQuiz [13], Apollo [14], Gbrowser [15], Imagen [16], MAGPIE [17], Manatee [18], Pedant [19], VisualGenome [20], PseudoCAP [21], Community Annotation Project [22], Alternative Splicing Annotation Project [23], Genestream [24], Cancer Annotation Project [25], Ensembl Genome Annotation Project[26] and NCBI's Genome Annotation Project [27].

The public data sources are heterogeneous, feature large data volumes, are in constant growth, but do not usually have a detailed documentation of the database schema. Also, the analysis programs do not have good documentation that details the execution parameters and the input and output data formats, which makes it difficult to integrate them into more complex workflow processes.

The major functional requirements of annotation systems are:

- 1) The system must store:
  - a) annotations manually created by the researchers;
  - b) annotations automatically generated by analysis programs;
  - c) annotations imported from external data sources; or
  - d) hyperlinks to annotations stored in external data sources.
- 2) The system must offer tools to:
  - a) search the annotations stored, using text retrieval techniques;
  - b) navigate through hyperlinks to annotations stored in external data sources, if it is the case;
  - c) locate annotations by "drilling down" the data pertaining to a genome (such as chromosomes, contigs, reads, ORFs, among others);
  - d) tabulate annotations assigned by different sources to the same biosequences.
- 3) The system must maintain versions of the annotations.
- 4) The system must offer distributed access to the annotations.
- 5) The system must offer workflow-like facilities to control the execution of analysis programs and to specify which programs generate annotations that must be stored.

Requirement (1) favors a data warehouse strategy to guarantee minimum performance. Requirement (2d) helps users compare the annotations. For example, the user may run BLAST to compare an ORF against a data source to detect similar biosequences and then work with the functional annotations of the biosequences detected to help him annotate the original ORF. Requirement (3) is especially useful to annotate genome data with incomplete sequencing. Requirement (4) makes it almost mandatory to offer a Web interface.

The systems investigated do not meet all the

requirements listed. For example, only one of the systems, Imagen [16], permits a workflow-like composition of analysis programs, but it does not offer distributed input of manual annotations.

These observations indeed motivated the BioNotes project, detailed in the next section.

### 3 The BioNotes System

The BioNotes systems, briefly described in this section, was designed to meet the requirements listed in Section 2. Its implementation uses the infrastructure of the Bio-AXS system [28], a biology data warehouse integrating data from several public sources. BioNotes is written in Java 1.4.0, for portability, and offers a Web interface implemented with JSP technology. The interface is certified for Internet Explorer 5.0 or later. The tool runs on top of a relational database management system. The database directly stores XML documents and supports Xpath queries.

#### 3.1 Functional Description of BioNotes

BioNotes implements the concept of a *user community* to control access to the data. A user community is just a set of users, possibly from different institutions. The system also offers different user profiles to further control which users can execute what commands.

BioNotes supports all three types of annotations.

A user may currently search annotations imported from the following external public sources: GenBank [1], PIR [29], SWISS-PROT[30], PROSITE[31] and Interpro [32].

The user may also run and store the annotations automatically generated by the following analysis programs: Phred [35,36], Phrap [37], GLIMMER [2], tRNAScan [3], RBSFinder [39], transTerm [40], BLAST [33], InterproScan [32] and CAP3 [34].

BioNotes transforms the annotations imported from external data sources, or automatically generated by analysis programs, to a common format – expressed as XML documents – before storing them in the data warehouse (see Section 4.2).

Finally, the user may manually add, delete and update annotations, which then become available to his community. Therefore, the concept of user community facilitates sharing annotations, in a controlled way. Updating a manual annotation actually creates a new version of the annotation. Hence, the user may browse through the various versions of an annotation and track down who created them. Only the user who created (a version of) an annotation may delete it.

BioNotes offers a variety of tools to access the annotations stored in the data warehouse. For example, given a biosequence identifier, the user may retrieve all annotations pertaining to the biosequence (contigs, reads and singletons), such as the keywords and feature table

assigned to the biosequence by Genbank. Likewise, the user may retrieve all annotations automatically assigned to the biosequence by the execution of an analysis program, such as the homologous biosequences obtained by running BLAST. In addition, BioNotes offers graphical schemes to help data visualization. For example, the system graphically indicates, for a contig  $C$ , the reads that compose  $C$  and the ORFs, tRNAs and ribosome sites present in  $C$ . These graphical elements are links to their respective annotations.

The system also has a color coding scheme to indicate the quality of a base (that is, A, T, C or G in DNA sequences) that helps the user analyze the quality of the read.

BioNotes also permits the user to access biosequence annotations stored in other sites, such as NCBI [41]. For example, the result of executing BLAST to compare a sequence  $S$  with the NR database [42] (obtained from NCBI) will contain links to the NCBI site that point to the annotations of each biosequence that is similar (according to BLAST) to  $S$ .

The system also supports the concept of a *private data source*, that is, a local set of biosequences and annotations that is private to a user community. The current examples are TCRUZI, a private data source that stores annotations and biosequences pertaining to the *Trypanosoma cruzi* organism, and GLUCONA, that stores annotations and biosequences pertaining to the *Gluconacetobacter diazotrophicus* organism.

These two private data sources were created using very different strategies. TCRUZI exemplifies a private data source created by importing data from a public data source, Genbank in this case. Indeed, data pertaining to the *Trypanosoma cruzi* organism was first extracted from Genbank, which is not difficult since Genbank lets one search the data by organism name. Then, the data were remapped to an internal format, defined by an XML schema stored in BioNotes (see also Section 4.2), and stored in the data warehouse.

On the other hand, GLUCONA was directly created using BioNotes tools. The chromatograms of the *Gluconacetobacter diazotrophicus* organism, the sequencing lab made available, were first submitted to the Phred program, which generated *reads* and annotations. These data were then remapped to an internal format, again defined by an XML schema stored in BioNotes, and stored in the data warehouse.

In addition, the user may analyze biosequences by executing programs at sites external to BioNotes. For example, the NCBI site permits executing BLAST to compare a sequence with several databases. BioNotes allows the user to compare a biosequence with several databases stored at NCBI by sending the sequence to the NCBI server and invoking the BLAST service at the NCBI site.

The Molecular Biology data sources are prone to errors and inconsistencies. To facilitate data analysis, BioNotes tries to store curated (consistent) data from the various sources, such as SWISS-PROT, and to execute programs, such as BLAST, using non-redundant data sources, such as NR, obtained from NCBI.

BioNotes stores, together with each annotation, its source. However, this tracing is inefficient when the annotation originates from a database that has not been curated. Indeed, such data source guarantees neither data quality nor indicates how the annotation was obtained.

We conclude this section with an open issue. While a genome project is in progress, sequencing of the genome will generate new *reads*. As a consequence, the data sources will constantly be updated and new runs of the analysis programs will further generate new derived data. If the data source is external, to remain up-to-date, BioNotes must refresh its data warehouse from time to time by re-accessing the data sources to retrieve new data. Also, the system must re-run locally the analysis programs to generate new versions of the derived data. Therefore, the system must offer tools to help users compare different versions of the data, as new reads become available, and transfer annotations from older to newer versions. This issue is further discussed in Section 4.

### 3.2 The BioNotes Data Model

BioNotes features a semi-structured data model, implemented on top of relational tables with columns storing semi-structured data as XML documents. The system keeps the description of the XML documents as XML schemas, again stored in a special table.

The system currently features:

- pre-defined functions to query XML data;
- special indexes over the XML data to improve performance;
- pre-defined functions to check the syntax of XML documents against XML schemas.

The entity class *Annotation* (see Figure 1) is specialized into the sub-classes *Automatic*, *Imported* and *Manual*. All annotations are implemented as XML documents stored in table columns (of type XML). The entity class *Schema* contains the XML schemas that define the syntax of the XML documents that represent annotations.

The entity class *Data Source* (see Figure 2) represents the annotations imported from external public data sources.

The entity class *Analysis Program* (see Figure 3) represents the annotations automatically generated by analysis programs.

Finally, the entity classes *User*, *Community* and *Institution* (see Figure 4) models the concept of user community.

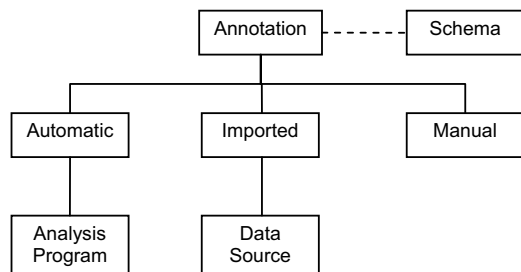


Figure 1. The *Annotation* entity class.

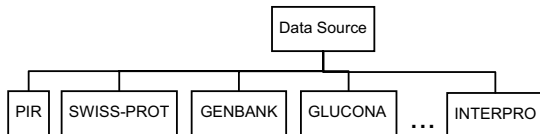


Figure 2. Data Sources.

Certain external data sources, such as PIR, Interpro, SWISS-PROT and Genbank, already distribute data as XML documents, whose structure is also available (either as XML schemas or DTDs). BioNotes then stores data exported from these data sources in their external format. However, other data sources, such as Prosite and Blocks, distribute their data in a proprietary format. In these cases, BioNotes has specific XML schemas to define the local format of the data and the system offers parsers to transform their external format to the BioNotes format.

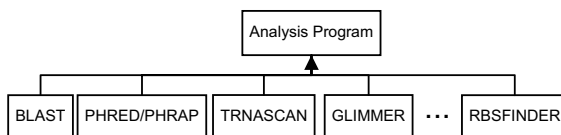


Figure 3. Biosequence analysis programs.

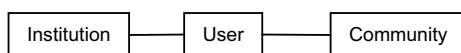


Figure 4. Users, Communities and Institutions.

Likewise, certain analysis programs generate data in XML format according to a well-defined (and published) XML schema, while others generate data in a proprietary format. InterproScan and BLAST are examples of former variety, while Phred, Phrap, GLIMMER, tRNAScan and RBSFinder of the latter variety. BioNotes also offers XML schemas and parsers to transform output data from these programs into XML documents, before storing them into the data warehouse.

Finally, manual annotations are also stored as XML documents that follow a XML schema. When creating an annotation, the user is offered a Web form, where certain fields have controlled vocabulary, while others are free format. The form is first parsed into a XML document,

which is then stored.

## 4 Conclusions and future directions

We briefly described in this paper the functional requirements of biosequence annotation systems. We then outlined the functionality and the data model of BioNotes, a tool that meets these requirements.

BioNotes is under development at the Pontifical Catholic University of Rio de Janeiro and the system is in use at the Department of Biochemistry and Molecular Biology, Oswaldo Cruz Institute, and at RioGene.

Several issues remain to be addressed in the context of the BioNotes project.

The current implementation of Bio-Notes does not yet support workflow-like composition of analysis programs, which now has to be manually programmed. Extending the system to cover this functionality is one of the top priorities of the project.

Adding new analysis programs is also under consideration, especially those that help mining genome data that hint to new biological knowledge, to be verified by further lab experiments.

We conclude with an issue already discussed in Section 4.1. While a genome project is in progress, sequencing of the genome will generate new *reads*, creating new versions of the data. Hence, the system must offer tools to help users transfer manual annotations from older to newer versions of the data, as otherwise it would be very discouraging to annotate data from on-going sequencing projects. However, this is a challenge since it is not simple to detect which genome objects (chromosomes, contigs, reads, ORFs, etc), from different versions of the data, are identical.

We are currently experimenting with different strategies to address this problem. Consider contigs, for example. One strategy is to consider that two contigs are identical if most of their *reads* are the same. Alternatively, two contigs may be considered identical if they are best matches when BLASTing contigs from the different versions of the data.

By developing these strategies, we hope that BioNotes will help detect identical objects from different versions of the data and, thereby, the system will help transfer annotations from one object to the other.

## Acknowledgments

This work was partially supported by CNPq under grant no. 141938/2000-5 for Melissa Lemos.

## References

- [1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank", in *Nucl. Acids. Res.*, vol. 31, pp. 23-27, 2003.
- [2] A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg, "Improved microbial gene identification with GLIMMER", in *Nucleic Acids Research*, vol. 27 (23), pp. 4636-4641, 1999.

- [3] T.M. Lowe, and S.R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence", in *Nucl. Acids. Res.*, vol. 25, pp. 955-964, 1997.
- [4] Department of Biochemistry and Molecular Biology, Oswaldo Cruz Institute, March 2003, <http://www.dbbm.fiocruz.br>.
- [5] Department of Medical Biochemistry, Federal University of Rio de Janeiro, March 2003, <http://www.bioqmed.ufrj.br/>.
- [6] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M-A. Rajandream and B. Barrell. "Artemis: sequence visualisation and annotation", in *Bioinformatics*, vol. 16(10), pp. 944-945, 2000.
- [7] Scott Pearson, "Distributed Annotation System", March 2003, <http://www.biodas.org/>.
- [8] Celera, "CeleraBrowser", March 2003, <http://www.celera.com/genomics/commercial/home.cfm?ppage=literature>.
- [9] S. Moller, U. Leser, W. Fleischmann, and R. Apweiler, "EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation", in *Bioinformatics*, vol. 15, pp.219-227, 1999.
- [10] M.G. Reese, G. Hartzell, N.L. Harris, U. Ohler, J.F. Abril and S.E. Lewis, "Genome Annotation Assessment in Drosophila melanogaster", in *Genome Research*, vol. 10(4), pp.483-501, 2000.
- [11] Bioinformatics Group, Center for Genome Research at Bielefeld University, "GENDB", March 2003, <http://gendb.Genetik.Uni-Bielefeld.DE/>.
- [12] C. Lee, and K. Irizarry, "The GeneMine System for genome/proteome annotation and collaborative data mining", in *IBM Systems Journal*, vol. 40 (2), pp. 592-603, 2001.
- [13] S. Hoersch, C. Leroy, N.P. Brown, M.A. Andrade and C. Sander, "The GeneQuiz Web server: protein functional analysis through the Web", in *Trends in Biochemical Sciences*, vol. 25, pp. 33-35, 2000.
- [14] Berkeley Drosophila Genome Project and The Sanger Institute in Cambridge, UK, "Apollo Genome Annotation and Curation Tool", March 2003, <http://www.fruitfly.org/annot/apollo/>.
- [15] Generic Model Organism Project, "GBROWSER", March 2003, <http://gmod.sourceforge.net/>.
- [16] C. Medigue, F. Rechenmann, A. Danchin, A. Viari, "Imagine: an integrated computer environment for sequence annotation and analysis", in *Bioinformatics*, vol. 15, pp.2-15, 1999.
- [17] T. Gaasterland, C.W. Sensen, "MAGPIE: Automated Genome Interpretation," in *Trends in Genetics*, vol. 12, pp. 76-78, 1996.
- [18] Bioinformatics department at The Institute for Genomic Research, "Manatee", March 2003, <http://manatee.sourceforge.net/>.
- [19] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, HW. Mewes, "Functional and structural genomics using PEDANT", in *Bioinformatics*, vol. 17, pp.44-57, 2001.
- [20] Rational Genomics, "Visual Genome", March 2003, <http://www.rationalgenomics.com/visualgenome.html>.
- [21] University of Washington Genome Center and PathoGenesis Corporation, "Pseudomonas aeruginosa community annotation project", March 2003, <http://www.cmdr.ubc.ca/bobh/PAAP.html>.
- [22] Center for Genome Research, Whitehead Institute, "Community Annotation Project", March 2003, <http://www-genome.wi.mit.edu/annotation/microbes/methanosarcina/sarcinaCA/P/>
- [23] Barmak Modrek, and Christopher Lee, "Alternative Splicing Annotation Project", March 2003, <http://www.bioinformatics.ucla.edu/HASDB/generic.php3>.
- [24] Bioinformatics Unit, Institut de Génétique Humaine, Montpellier France, "GeneStream", March 2003, [http://xylian.igh.cnrs.fr/getseq/genbank\\_sequence\\_finder.html](http://xylian.igh.cnrs.fr/getseq/genbank_sequence_finder.html).
- [25] Bioinformatics Laboratory, Institute of Computing - University of Campinas, "Cancer Annotation Project", March 2003, <http://cancer.lbi.ic.unicamp.br/>.
- [26] M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyraas, J. Gilbert, M. Hammond, T. Hubbard, A. Kasprzyk, D. Keefe1, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik and E. Birney, "Ensembl 2002: accommodating comparative genomics", in *Nucleic Acids Research*, vol. 31 (1), pp. 38-42, 2003.
- [27] R. Agarwala, S. Chetvernin, V. Choi, D. Church, W. Jang, J. Kans, P. Kitts, D. Lipman, D. Maglott, J. Ostell, K. Pruitt, G. Resenchuk, G. Schuler, S. Sherry, T. Tatusova, D. Thierry-Mieg, J. Thierry-Mieg, S. Wheelan, "NCBI's Genome Annotation project - current status", March 2003, <http://hgm2001.hgu.mrc.ac.uk/Abstracts/Publish/Workshops/Workshop09/hgm0074.htm>.
- [28] Luiz Fernando Bessa Seibel, Sérgio Lifschitz, "A Genome Database Framework", DEXA, pp. 319-329, 2001.
- [29] C. H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, R. S. Ledley, K. C. Lewis, HW. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsugita, C. R. Vinayaka, L. S. L. Yeh, J. Zhang, and W. C. Barker, "The Protein Information Resource: an integrated public resource of functional annotation of proteins", in *Nucl. Acids. Res.*, vol. 30, pp. 35-37, 2002.
- [30] B. Boeckmann, A. Bairoch, R. Apweiler, MC. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilboud, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003", in *Nucl. Acids. Res.*, vol. 31, pp. 365-370, 2003.
- [31] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002", in *Nucl. Acids. Res.*, vol. 30, pp. 235-238, 2002.
- [32] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krejstyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan, and E. M. Zdobnov, "The InterPro Database, 2003 brings increased coverage and new features", in *Nucl. Acids. Res.*, vol. 31, pp.315-318, 2003.
- [33] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "A basic local alignment search tool", in *J. of Molecular Biology*, vol. 215, 1990, pp. 403-410.
- [34] X. Huang, A. Madan, "CAP3: A DNA sequence assembly program", in *Genome Research*, vol. 9, pp.868-877, 1999.
- [35] B. Ewing, L. Hillier, M.C. Wendl, and P. Green, "Base-Calling of Automated Sequencer Traces using Phred. I. Accuracy Assessment", in *Genome Research*, vol. 8, pp.175-185, 1998.
- [36] B. Ewing, P. Green, "Base-Calling of Automated Sequencer Traces using Phred. II. Error Probabilities", in *Genome Research*, vol. 8, pp.186-194, 1998.
- [37] P. Green, "Documentation for Phrap", March 2003, <http://bozeman.mbt.washington.edu/phraps.docs/phrap.html>.
- [38] D. Gordon, C. Abajian, and P. Green, "Consed: A Graphical Tool for Sequence Finishing", in *Genome Research*, vol. 8, pp.195-202, 1998.
- [39] The Institute for Genomic Research, "RBSFinder", March 2003, <http://www.tigr.org/software/>.
- [40] The Institute for Genomic Research, "TransTerm", March 2003, <http://www.tigr.org/software/transterm.html>.
- [41] National Center of Biotechnology Information, "NCBI Homepage", May 2003, <http://www.ncbi.nlm.nih.gov/>.
- [42] National Center of Biotechnology Information, "The BLAST Databases", May 2003, <ftp://ftp.ncbi.nih.gov/blast/db/>.