BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2883–2885
doi:10.1093/bioinformatics/bth321

GeneRecords: a relational database for GenBank flat file parsing and data manipulation in personal computers

P. D'Addabbo, L. Lenzi, F. Facchin, R. Casadei, S. Canaider, L. Vitale, F. Frabetti, P. Carinci, M. Zannotti and P. Strippoli*

Center for Research into Molecular Genetics 'Fondazione CARISBO', Institute of Histology and General Embriology, University of Bologna, Via Belmeloro, 8-40126 Bologna, Italy

Received on December 23, 2003; revised on March 24, 2004; accepted on May 2, 2004
Advance Access publication May 14, 2004

ABSTRACT

Summary: Extracting the desired data from a database entry for later analysis is a constant need in the biological sequence analysis community; GeneRecords 1.0 is a solution for GenBank biological flat file parsing, as it implements a structured representation of each feature and feature qualifier in GenBank following import in a common database managing system usable in a personal computer (Macintosh and Windows environments). This collection of related databases enables the local management of GenBank records, allowing indexing, retrieval and analysis of both information and sequences on a personal computer.

Availability: the current release, including the FileMaker Pro runtime application (built for Windows and Macintosh environments), is freely available at http://apollo11.isto.unibo.it/software/

Contact: pierluigi.strippoli@unibo.it

INTRODUCTION

The amount of information in biology has become enormous, even for specialized topics such as genetics. The need to create databases becomes important, to keep a record of the biological data for later references or analysis. The major genetics database, namely GenBank (Benson *et al*., 2002), is in a flat file format, that is a simple text format without formatting adjunctive elements, used by many of the most popular databases for everyday use in genetics research. The flat files are easy to distribute, access and maintain, while the data extraction and elaboration usually require the use of specific scripting language modules. For example, several Perl modules have been developed for parsing flat-file databases (http://www.bioperl.org). The alternative is to perform online search and analysis at Unix based servers. Most of the online analysis software can directly access flat file databases: BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) for sequence similarity search, Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) for keyword-based information retrieval.

The Sequence Retrieval System (SRS) (Etzold *et al*., 1996) also allows the direct download of retrieved sequences, but the search is more similar to a text retrieving operation than to a search in a database, due to the absence of a full relational schema in the indexing of features, qualifiers and descriptions. For example, a biologist could wish searching for the first exon (exon number 1) sequence of the human genes in the nucleotide database. Searching in SRS (release 7.1.1 at http://srs.ebi.ac.uk, library: EMBL, standard query form, option 'get results of type': features) for sequences in Division: 'hum' with feature: 'exon', qualifier: 'number' and feature description '1', translated as: ([embl-Division:hum*] > (([embl-FtKey:exon*] & [embl-FtQualifier:number*]) & [embl-FtDescription:1])), reports 7913 entries. The browsing of the reported features shows that there are many artifactual results: just among the first 100 entries found, there are 5 non-first exons that are listed (i.e. AB008560: exons 2, 3, 4, 5 and 6 of DNASE2 gene, because '1' is present in the term '3.1.22.1' that is contained in the qualifier 'EC_number', and so on). This artifact is due to the absence of adequate splitting of the different feature qualifiers in the SRS schema. Many systems other than SRS have been implemented for restructuring/reorganizing biological data (for a full description, see Lacroix and Critchlow, 2003); these systems often require high-level programming and querying skills and are mainly based on UNIX operating system.

*To whom correspondence should be addressed.

An *ad hoc* relational database, which can run on a personal computer in the Macintosh or Windows environment, could be a useful system for querying, retrieving and manipulating data. Parsing and indexing are the key step to convert data available in a flat file format to the appropriate record fields of a relational database. We developed GeneRecords (http://apollo11.isto.unibo.it/software/), a storage and retrieval system of genetic data to parse information from GenBank flat files, based on the highly popular FileMaker Pro engine, included as a runtime (FileMaker, Santa Clara, CA; see Stephens, 2002, http://www.dnjonline.com/articles/tools/iss28_reviews_filemaker.asp). The GeneRecords database can directly import GenBank flat files data sources, preformatted by a text filter, and it contextually generates a relational database with 43 files containing the parsed data, making them available for complete search and analysis. The system implements an independent representation of each feature and its respective qualifiers in GenBank within a common database managing system usable in a personal computer (Macintosh™ and Windows™ environments).

SYSTEMS AND METHODS

First, a detailed description of GenBank flat file format (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt) has been accurately analyzed for (1) identification of characters usable as consistent limits for each data type, and (2) conversion of the flat file format in a multiple related table series allowing the appropriate import for each data type. Our strategy is thus based on, first, treatment of the downloaded, expanded data file by a fast text conversion utility in order to find, change or insert control characters (e.g. insert carriage return and tabs at a desired point, to define the end, respectively, of each record and of different fields in the same record); this step is performed by automated text filtering (we incorporate in the software Guoniu Han’s ‘PowerReplace’ for Mac, while we enclose ‘Search & Replace’ by Funduc Software Inc. or ‘BK ReplaceEm’ by Boolean Dream for Windows), using an appropriate provided replacement filter based on invariant features of the GenBank flat file established format. At this point the file is imported into the appropriate text fields (e.g. ‘NM Entry’, ‘Features’, ‘Seq 1 50100’) of the GeneRecords database, which will automatically parse data using calculated fields, that extract each single data type from the text fields. The whole import process is driven by FMP scripts launched by ‘Import’ button. The FMP 6 template which includes these calculation fields is available at http://apollo11.isto.unibo.it/software/. It is provided as a standalone tool including the FMP 6 runtime; the file is Mac and Windows compatible. The detailed instructions and technical specifications are distributed with the software, along with a Tutorial. The free included FMP runtime allows full records management and browsing, while the creation of new fields for elaboration or further relationships definition require the installation of the FMP application.

DISCUSSION

It has been stated that ‘Although the [GenBank] database was never meant to be read from computers, an army of computer-happy biologists have nevertheless parsed, converted, and extracted these records by means of entire suites of programs’ (Ouellette, 1998). Indeed, we find that to date no single, simple tool exists for convenient parsing and analysis of GenBank data into a local database for personal computer users. This category largely comprises the biologists, who usually have little confidence in programming languages and Unix workstations. This fact creates a barrier between the molecular biology researchers and the straightforward processing of original sequence data and associated information, the latter often being restricted to specialized bioinformaticians. Interestingly, in a recent special issue on genome analysis in the Nature Genetics journal (Wolfsberg *et al*., 2002), the authors of the editorials complain of the unexpected scarce use of valuable sequence data source from the research community (see also Buckingham, 2003).

We present a new tool to facilitate the analysis of sequence datasets which does not require high-level computer skills, and converts large data files in GenBank format into a standard database immediately usable on a personal computer. Each feature subdatabase is composed of several fields representing all possible qualifiers attributed to a given feature, so searching for ‘first exon’ can be made in ‘Exon’ subdatabase by searching ‘1’ in ‘number’ field.

Moreover, the very simple menu commands of this software readily allows the performance of calculations on the imported data without the knowledge of any programming language. For example, a ‘Mean’ summary field could be added to readily extract the mean value of the sequence length of a particular records subset. Finally, the availability of complete GenBank datasets in a relational database format allows the easy integration with other biological databases available in the same or similar format; for example, Unigene (collection of ESTs clusters) and LocusLink (a table of correlation among different types of biological data) can be easily imported in an FMP template. Our release includes the integration with LocusLink (http://research.nhgri.nih.gov/microarray/downloadable_cdna.html), with the respective GeneOntology tags.

REFERENCES

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.

Buckingham,S. (2003) Bioinformatics: programmed for success. *Nature*, **425**, 209–215.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Lacroix,Z. and Critchlow,T. (Ed.) (2003) *Bioinformatics—Managing Scientific Data*. Morgan Kaufmann Publishers, San Francisco, CA.

Ouellette,B.F.F. (1998) The GenBank sequence database. In Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics—A Practical Guide to the Analysis of Genes and Proteins*. John Wiley & Sons, New York.

Stephens,P. (2002) Alternative access. Developer Network Journal, **28**, 30–32.

Wolfsberg,T.G., Wetterstrand,K.A., Guyer,M.S., Collins,F.S. and Baxevanis,A.D. (2002) A user's guide to the human genome. *Nat. Genet.*, **32**(Suppl.), 1–79.