DATABASE OF LONG TERMINAL REPEATS IN HUMAN GENOME: STRUCTURE AND SYNCHRONIZATION WITH MAIN GENOME ARCHIVES

Alexeevski A.V.¹, Lukina E.N.¹, Salnikov A.N.*², Spirin S.A.¹

¹ Belozersky Institute, Moscow State University, Moscow, Russia; ² Department of Computational Mathematics and Cybernetic, Moscow State University, Moscow, Russia

* Corresponding author: e-mail: salnikov@angel.cs.msu.su

Keywords: human genome, long terminal repeats, synchronization, data bases

Summary

Motivation: The complexity of main nucleotide and complete genome databases (DBs) makes it essential to develop special well-structured, equipped by user-friendly interface, and curated by experts secondary DBs. The problem of synchronization of the information between DBs arises. This problem seems to be especially actual for the human genome because of its continued improving and ordering.

Results: Yu.B. Lebedev group, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, collected and annotated long terminal repeats of class 5 (LTR5). This collection was the starting point for the development of Human Endogenous Repeat database (HEREbase). Currently, HEREbase contains nearly all (~1000) appearances of LTR5 objects in human genome. Query forms allows user to select specified entries. The created program scripts weekly synchronize links to relevant objects in the nucleotide DB (EMBL) and human genome DB (EnsEMBL).

Availability: http://math.genebee.msu.ru

Introduction

The sequences of the complete human genome are contained in well-known publicly opened data bases like the archive database of nucleic acid sequences EMBL (http://www.embl.org) and the database of complete genomes EnsEMBL (http://www.ensembl.org). Such large universal archives are rather complicated and not convenient for purposes of studying specific objects in genome. A user can have difficulties to create adequate queries; searching for results, which can contain thousands of sequences, takes a lot of time due to DB size and multiple simultaneous users requests. In addition, an expert hardly can add new information about an existing object to an archive DB. These inconveniences can be reduced by creating smaller specialized secondary databases that are well-structured and equipped by user-friendly interface. Due to the continued reordering of the human genome the problem of synchronization of the information in small specialized DB with the information in archive DBs becomes very important and somewhat nontrivial.

One of intensively investigated objects in human genome are so-called Long Terminal Repeats (LTR), which are traces of ancient retroviral invasions. The potential possibility of LTRs to be involved into gene expression regulation makes them interesting for scientists who study human evolution, inherited diseases, cancerogenesis etc. (Khodosevich et al., 2002). A curated DB for LTRs and neighboring genes can be useful for such studies.

Methods and Algorithms

HEREbase was created in MySQL under Linux (debian) operating system. The user and curator interfaces are made by means of Apache and php4. Scripts to synchronize the information in HEREbase with the databases EMBL and EnsEMBL, are written in Python language.
Results
We developed a small specialized DB named Human Endogeniouse RetroElement database (HEREbase, http://math.genebee.msu.ru). Currently, the HEREbase is filled by one class of LTRs, namely LTR5. An annotated collection of LTR5 created by E.D. Sverdlov and Yu.B. Lebedev (Shemyakin – Ovchinnikov Institute of Bioorganic chemistry), was in the background of HEREbase. It was extended to all annotated LTR5 objects in EnsEMBL, using information from HERVd (http://herv.img.cas.cz/, see Pačes, Pavlíček et al., 2004). The HEREbase is filled also by neighboring genes of LTRs.

HEREbase relational tables contain information on LTRs themselves: their sequence, location in the genome (chromosome, arm, band), 5' and 3' short direct repeat sequences, class of theLTR (according to the classification of Lavrentieva et al., 1998), and links to the relevant EMBL entry and to the coordinates in the chromosome according to the latest release of EnsEMBL. Additionally, HEREbase contains flanking sequences to facilitate the identification of the LTR, the information about genes in 100k neighborhood and repeats in 2k neighborhood. HEREbase can be curated via an Internet browser. For this purpose, a user (‘guest’) and a curator have different privileges. HEREbase queries suggests a number of possibilities for LTR5 selection. It is possible to make restrictions on classes of LTR5, on involvement/non-involvement of a LTR5 in a provirus, on chromosome localization (e.g., “8p12”). Also LTRs can be selected that are close to annotated genes. User can input a segment of LTR sequence or a sequence of an LTR’s flank and possess the list of LTR5 that have this segment in their sequences (resp. sequences of their flanks). If user does not input any restrictions, then he gets the list of all LTRs in the data base. To synchronize the information, a number of program scripts in Python language were created. Each script weekly automatically checks specific links to EnsEMBL or EMBL and extracts the information specified in HEREbase. In the case of disagreement a script tries to correct discordance. If it is possible, the corrected information is put into HEREbase. If no, the script includes the information about the disagreement into the special table visible (via http) to curators of HEREbase. A curator of the HEREbase is supposed to resolve the found conflict. In the case of EMBL the links the script finds are an EMBL entry by accession number specified in HEREbase and coordinates of the LTR in the entry, which are corrected (if needed) using BLAST. More complicated algorithm is needed to deal with EnsEMBL. EnsEMBL is a mysql operated relational database whose tables are open via ftp protocol. The table cross-references reflects complicated relations between all detectable objects in the genome. For example, LTR coordinates in a chromosome can be derived only from the third level of EnsEMBL table inclusions. An additional problem arises after correction of EnsEMBL structure by the EnsEMBL programmer group. For example, the table named ‘contigs’ in release 19 was transferred into the table named ‘sequence_region’ in the release 20. Such EnsEMBL corrections requires corrections in the HEREbase scripts. Currently HEREbase contains 1032 LTR5 entries. The process of annotation of these entries is continued. The HEREbase structure and software can be used not only for LTR5 in human genome, but also for other classes of repeats in any complete eukariotic genomes.

Acknowledgements
The authors are grateful to Yu.B. Lebedev and I. Mamedov (Shemyakin – Ovchinnikov Institute for Bioorganic Chemistry RAS) for the background data and useful consultations. The work was supported by Ludwig Institute for Cancer Research (CRDF grant RB01277-MO-2).

References