**PROGRAMS FOR DATA INPUT TO THE TRANSCRIPTION REGULATORY REGIONS DATABASE.**

ANANKO E.A., NAUMOCHKIN A.N., FOKIN O.N., FROLOV A.S.

Institute of Cytology and Genetics, (Siberian Branch of the Russian Academy of Sciences), 10 Lavrentieva ave., Novosibirsk, 630090 Russia

Keywords: data input, transcription, regilatory regions, database, gene networks, verifying the syntax, vocabularies, Internet, MS SQL Server

The informational flow on regulation of gene expression is rapidly growing now. A database for accumulation of this information [*Kel’ A. et al., 1997*] has been developed at the Institute of Cytology and Genetics. In this work, we describe two systems for collecting the formalized data from original papers and verifying the syntax of the previously accumulated information. The first system is designed for local computers and does not require the Internet connection. It is used mainly by experts for checking and editing the data. The second system is designed for remote users and operates via the Internet. It allows input the data on regulatory regions of genes, obtained by researchers in different countries. Both programs are supplemented with a system of vocabularies, verifying the correct data input, and context help.

**1. The local version of the program for entry editing of the TRRD database**

This version is designed for editing the TRRD entries in a flat file format and verifying the field structure and compliance with the vocabularies. The program ID\_TREE.exe is realized in Visual FoxPro 5.0 using OLE technology and ActiveX elements.

The entry to be edited is loaded from the text file and converted into a structure providing its tree-view representation, the nodes of which represent the individual blocks of this entry (Fig. 1).

The major blocks of an entry are: (1) block of the fields that are common for the entire gene (ID block); (2) a group of blocks describing the regulation pattern of the gene (RE blocks); and (3) the block describing regulatory regions (RG blocks), which include the subblocks describing hypersensitive sites (HN blocks) and promoters (AP blocks). The latter, in turn, are consisted of subblocks for site description (AN blocks) that contain the subblocks describing transcription factors (TF blocks). The references (AU blocks) are in the end of the entry.

The blocks can be minimized to a single line or expanded. Each block type has its own icon. The entry is verified during the loading. The overall structures of individual fields and blocks are checked. Obligatory fields are introduced into the entry automatically. Individual fields are checked for their compliance with the vocabularies. The uniqueness of block-identifying fields is verified. The lines containing any errors are marked with an interrogation point. The complete list of the errors can be displayed too.

Navigation along the lines of the tree activates the access to the vocabularies corresponding to certain fields. Context-sensitive help (F1 key) is also available. Editing of both an individual line and the entire text of the entry is provided. Query-based insertion from the vocabularies to the corresponding field of the entry is available under both modes. The vocabularies themselves (Fig. 2) are edited and updated, as required. The edited entry is exported again into a text file or compiled into a specialized database.

**2. The system for Internet-based data input to the TRRD database**

The system for Internet-based data input to the TRRD database consists of several tightly interacting components: (1) User-TRRD Client Module; (2) TRRD Server; and (3) MS SQL Server (Fig. 3).

The main task of the User-TRRD Client Module is the reception and transmission of the data between the Internet user and the TRRD Server. This module is a CGI-bin application starting on the server

(WWW) by the user's browser using POST method. The module consists of two interconnected blocks: one, for connection with the Internet user, the other, with TRRD server. Running User-TRRD Client Module (UTCM) receives the data from the Internet user, connects with TRRD server, and in case of a valid connection, transmits the data received from the user to the server. Then it waits for the data to be transferred from the server back (results of the user's data processing) to convey them to the user. After the TRRD server has processed the data received, it transfer the result to User-TRRD Client Module, which conveys it to the user. In case the TRRD server is disconnected for some reason, has not been connected initially, or the waiting time for the data processed has elapsed, User-TRRD Client Module identifies the errors and informs the user.

The next constituent is the TRRD Server, which performs the main processing of the users' data. On connection of User-TRRD Client Module with TRRD Server, the server logs it and receives the input data (the received query), then organizes the stream of the user's task processing (input query), and adds it to the queue. A number of user's tasks may occur simultaneously; their maximal number is determined by the system resource capacity and the settings of the TRRD Server. The flows of user's tasks is controlled by the Client Task Manager, which coordinates the traffic on the whole, synchronizes the steps of performances, distributes the resources, etc.

Each user's task is a complete subsystem for processing of the input query. When the command of the Client Task Manager to run the processing arrives, the input data are processed by Data Conversion Module (of the user's task), which verifies the data and converts them into an internal format required for the subsequent processing. Then the control (within the user's task) is passed to the module Pre-Verification, which recognizes the type of the input data and determines the corresponding processing scenario. The module Data Processing Scenarios performs the selected scenario. The initial task in performance of the scenario is verification of the input data, that is, checking their consistency with the data type and correction of data format.

When the verification has been completed, the next operations are determined depending on the data contents. The queries to SQL Server to extract, input, or additionally process the data are performed through the interface provided for by the module Database Query Manager. In the process of these operations, the output data are determined and transferred to the Data Conversion Module to be converted into HTML format and transmitted to the Internet user. The input errors of the Internet user, incorrect data format as well as the errors connected with the operation of the SQL server and user's task are considered. If necessary, the return to the step preceding the error emergence is performed. The message on the error detected is send to the Data Conversion Module for formatting. When the output data are received, the user's task informs Client's Task Manager that it has been completed; and the output data are transferred to the Module of Connection with User-TRRD Client Module (CGI-bin) to determine the receiver of these data and transmit them to user.

SQL Server is a basic component of the system: it compiles the data tables of the TRRD database, preserves the reference integrity, supports SQL queries, and carry out various procedures. Interaction with SQL Server is performed via the DataBase Query Manager of the user's task.

The system is designed to provide the authorization of the information, that is, the access of a user only to his own data and the data from the official TRRD release.

Separation of the Internet-based data input system into several individual modules allowed to provide the integrity of the data under heavy traffic and the general control of the resources.

Thus, the both systems described above are complementary and provide correctness and validity of the input data and compiled information. After the some reconstruction these systems can be used for data input to other databases of such type.

**3. Acknowledgments**

This work was supported by the State Science and Technology Program "The Human Genome" of the Russian State Committee for Science and Technology, Russian Foundation for Basic Research (grants 96-04-50006, 97-04-49740, 98-04-49479), and Integration Program of the Siberian Branch of the Russian Academy of Sciences.

**4. References:**

1. Kel A.E., N.A. Kolchanov, O.V. Kel, A.G. Romashchenko, E.A. Ananko, E.V. Ignateva, T.I. Merkulova, O.A. Podkolodnaya, I.L. Stepanenko, A.V. Kochetov, F.A. Kolpakov, N.L. Podkolodny, and A.N. Naumochkin “TRRD: database on transcription regulatory regions of eukaryotic genes” Mol. Biol. (Mosk)., 31, 521-530 (1997).

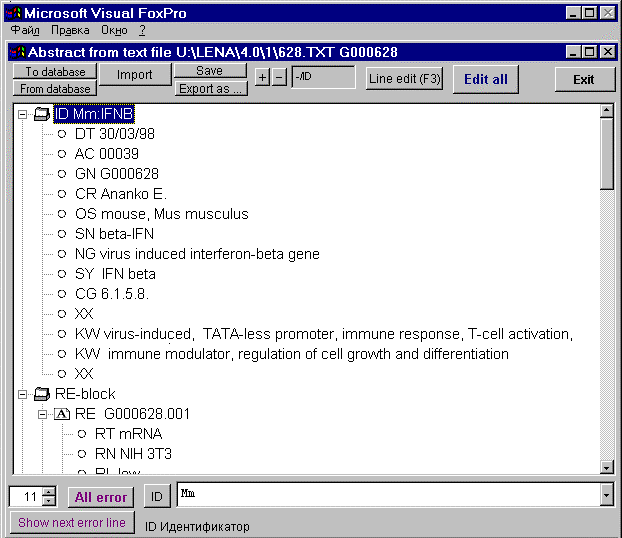


Figure 1. An example of the main window of the program with partially expanded blocks.

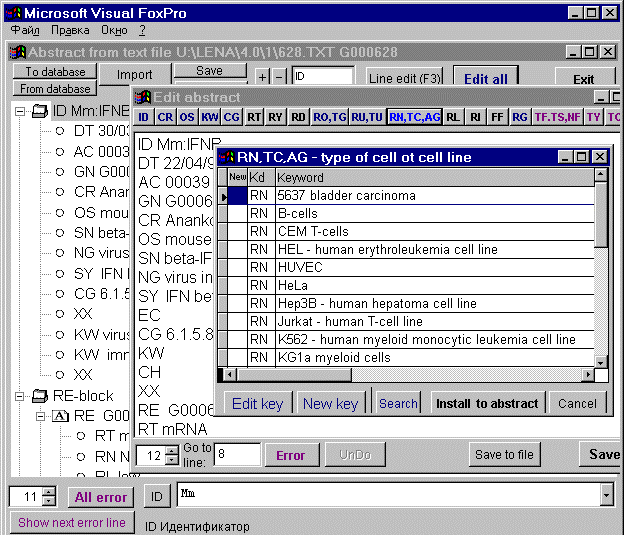


Figure 2. An example of calling the Key Word Vocabulary from the Text Editing Window of the entry.

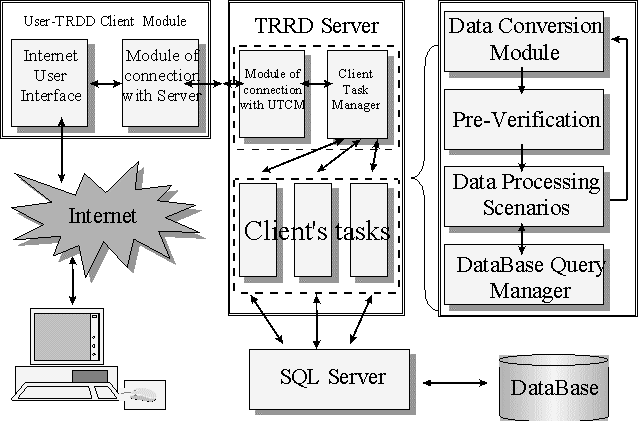


Figure 3. Basic modules of the system of data input to the TRRD database.