Poster B-44

An Integrated database of structural parameters for protein analysis

Authors:
Stanley R. M. Oliveira (*Embrapa Agricultural Informatics*)
Gustavo V. Almeida (*Embrapa Agricultural Informatics*)
Kassyus R. R. Souza (*Embrapa Agricultural Informatics*)
Diego N. Rodrigues (*Embrapa Agricultural Informatics*)
Goran Neshich (*Embrapa Agricultural Informatics*)
Paula R. Kuser-Falcão (*Embrapa Agricultural Informatics*)
Michel E. B.Yamagishi (*Embrapa Agricultural Informatics*)
Edgard H. Santos (*Embrapa Agricultural Informatics*)
Fábio D. Vieira (*Embrapa Agricultural Informatics*)
José G. Jardine (*Embrapa Agricultural Informatics*)

Short Abstract: STING_DB is a relational database composed of structural parameters for protein analysis operating with a collection of both publicly available data (PDB, HSSP, Prosite) and its own data (contacts, interface contacts, surface accessibility). STING_DB has over 300 parameters compiled at the single site and was implemented using free software.

Long Abstract:
The relational database of STING is composed of structural parameters for protein analysis. This database operates with a collection of both publicly available data (PDB, HSSP, Prosite) and its own data (contacts, interface contacts, surface accessibility). This database was designed to support applications for interesting biological cases. For this reason, STING_DB is one of the best known databases of structural parameters with over 300 of them compiled at the single site.

Considering its relevance for biologists, researchers and others interested in protein analysis, we decided to proceed with the STING_DB migration from flat files to a relational database in order to provide more complete and modern environment for structure analysis. The information for analyzing structure relationships, the quality of the structure, nature and volume of atomic contacts of intra and inter chain type, relative conservation of amino acids at the specific sequence position based on multiple sequence alignment, indications of folding essential residue (FER) based on the relationship of the residue conservation to the intra-chain contacts and Ca-Ca and Cb-Cb distance geometry is now going to be more accessible and readily addressable for data grouping and mining.

The main features of the relational database of STING can be summarized as follows:

• It is based on indices, which speeds up the search for information and, consequently, improves the response time of the protein analysis process.

• It is available for different platforms. Currently, it is implemented and available in the database MySQL. However, it could be ported to other platforms, such as ORACLE or even

Postgres.

• It greatly reduces the redundancy of information.

• It allows users to compare different protein structures, at the same time, which was not possible with the previous version (flat files).

• Its update is much simpler, since it was built on relational database facilities.

Apart from the features mentioned above, this database provides its users with a data quality indicator, i.e., a new module developed to meet quality requirements in research databases. On a weekly basis, a checklist procedure is performed to identify the parameters/files that are both missing and/or empty for the new PDB files added to the database. The main goal of such a procedure is to guarantee that the quality of the data will not be degraded as the updates take place. When the checklist procedure identifies a group of parameters that are missing and/or empty, a report is automatically sent to the Sting_DB administrator who will run a set of scripts to update the parameters concerning the new PDB files, and subsequently, perform the checklist procedure to evaluate the quality of the updated data.

When a Sting user selects a PDB file for analysis, if one or more parameters of that PDB are not available at the STING_DB, the user can search for such a PDB name in the Sting_DB QA to verify the existence of those parameters. For each parameter, there is a list of missing and empties PDB containing the PDB's parameters. However, this situation is unusual since we are keeping the percentage of missing and empty PDB's parameters below 3% in almost all PDB files. In addition, we are working diligently to reduce that percentage to 1% or less.