Poster B-33

STING Database Quality Assessment

ISMB 2006
august 6-10, 2006
Fortaleza, Brazil

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

Short Abstract: Inaccurate or inconsistent data can hinder our ability to interpret the research results. An effective data quality strategy can help the research to better handle its scientific objective, reducing costly operational inefficiencies. Here we present the results of our effort to access the quality of data in STING Database.

Long Abstract:

The goal of data management is to provide the infrastructure to transform raw data into consistent, accurate and reliable information. Such information can provide the foundation for strategic decisions and support to research projects. However, most projects have dozens, if not hundreds, of data sources – each with different data formats, conventions and rules. And each data source has a variety of problems – null values, missing fields, inconsistent entries – that lead to an overall lack of data quality. An effective data quality strategy can help research groups better understand their research objectives, reducing costly operational inefficiencies in strategic projects.

In this work, we introduce the Sting Database Quality Assessment (Sting_DB QA), a new module of the Sting Suite (now in its STAR version), developed to meet quality requirements in research databases. In particular, STING_DB is a database composed of structural, sequence, function and stability parameters/descriptor 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, relative entropy, cross-link order, space clash and others). STING_DB is one of the best known databases of structural parameters reported in per-residue fashion with over 300 of them compiled at a single site.

Figure 1. The Main Page of the Module Sting Database Quality Assessment. An example of missing and empty files related to Cross Presence.

Considering its relevance for researchers interested in protein analysis, the module Sting_DB QA was designed to measure the quality of the data deposited in the Sting_DB. 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. In Figure 1, it is shown an example of missing and empty files related to Cross Presence.

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 empty PDB files containing the structure related parameters. However, the situation where the data are missing is not all that frequent since we are trying to improve the quality of our DB by keeping the percentage of missing and empty structure parameters below 3% for almost all PDB files. In addition, we are working diligently to reduce that percentage to 1% or less.

This effort makes the Sting_DB unique in terms of quality assessment when compared with other counterparts in the Bioinformatics domain. To the best of our knowledge, Sting is the only software for protein analysis capable of providing its users with a data quality indicator.