Temporal Analysis of Nucleotide Sequence Annotation

ID - 164

Dalla Emiliano[1], Braidotti Michele[1,2], Buttol Andrea[1,3], Zantoni Marco[3], Montanari Angelo[3], Schneider Claudio[1]

[1]Laboratorio Nazionale CIB - Area Science Park, 34012 Trieste, Italy

[2]Arizona Genomics Institute - University of Arizona, Tucson, AZ 85721, USA

[3]Dipartimento di Matematica e Informatica - Università di Udine, 33030 Udine, Italy

Motivation

The availability of new releases of existing biological databases represents a source of relevant changes in the information associated with a nucleotide sequence. As a consequence, in order to gain maximum knowledge from clone/sequence annotations, it becomes crucial to record the history of every clone, that is, to trace possible variations in genomic location, matches with available ESTs, matches with known or predicted genes or transcripts, and so on. Few of the existing bioinformatics tools for biological sequence characterization (analysis) are provided with a database component where the results of the analysis can be recorded and easily accessed for further elaboration. Moreover, complex queries are limited within traditional database applications: the level of interaction that the few existing databases allow is fairly limited, enabling only a narrow range of analysis. Finally, and most importantly, a critical limit of all public and local databases is the lack of a systematic treatment of temporal information. This is the case, for instance, with some new generation platforms for genome-scale analysis, whose local database is mainly used as an easy-to-access repository for user histories. From here on, we will focus our attention on the management of temporal information, by introducing the possibility to label relevant information, e.g., clone annotation, with the time interval during which it is current in the database.

Methods

The system we propose records BLAST results over time and makes it possible to use them in temporal analysis. Temporal information associated with sequence annotations is used to assign sequences a score that measures the reliability/conservation of their annotation. The score is computed on the basis of permutations, number of nulls, insertions and deletions that take place from one BLAST results set to the next one. The multiple entry point analysis pipeline is designed to handle sequence chromatograms and to perform their functional characterization. It consists of a series of Perl scripts that perform two different quality control steps, a vector trimming phase and the clustering and functional annotation steps of the annotation phase. They interact with the Temporal Sequence Data Base (TSDB), based on a temporally-extended data model of the Enhanced Entity-Relationship, where sequence, quality, and mapping data are stored, as well as with internal and external analysis tools.

TSDB is one of the fundamental components of the system. Its flexibility allows one to deal with incomplete information and the presence of a temporal dimension makes it possible to keep track of data history. Finally, it supports information sharing and exchange. Additional data, such as new information about genomic location, ontologies, or orthologous comparisons, possibly generated by other experiments carried out in the laboratory (e.g., cDNA microarrays gene expression experiments), can be easily integrated into TSDB. In a similar way, relevant data contained in TSDB can be easily exported to other databases.

Temporal information associated with sequence annotations is used to assign a score to sequences which measures the reliability of their annotation. The score is computed on the basis of permutations, number of nulls, insertions and deletions that take place from one annotation set to the next one. It penalizes sequences with changes in their annotation due to recent BLAST executions much more than sequences with annotation sets which are permutations of previous ones or sequences that only present changes due to old BLAST executions.

As it is apparent, the score is not an index of sequence or public database quality, but a way to detect if and when a sequence annotation changes, in which way and for which reasons, penalizing sequences with recent meaningful variations.

Results

The proposed computational system was used to support the production and the (temporal) analysis of human cDNA libraries obtained with the CAP-Trapper method. The entire system, however, is applicable to any other biological protocol. The changes that occurred in the results of the different BLAST executions we performed over time were recorded in TSDB and then used for sequence analysis. The scoring function was exploited to measure the stability, and thus the reliability, of cDNA sequence annotations. The main reason for supporting a temporal management of sequence annotation is to compare BLAST results obtained at different time.

New BLAST executions, indeed, do not necessarily return updated versions of previous results, that is, query sequences do not necessarily match the same subject sequences over time. Unfortunately, when an update of BLAST databases occur, there is no way to access their previous state anymore. On the contrary, TSDB supports the so-called as-of queries that allow one to recover (BLAST) information that would have been obtained by querying a previous state of (BLAST) databases.
Availability: http://www.bioinfo.lncib.it
Email: dalla@lncib.it