Poster B-35

FlagelLink – An integrated organellar database dealing with refined pattern recognition of specific motifs/domains targeted to the eukaryotic flagellum

Authors:

Fabiana Freire de Araújo (*NUGEN-UECE*)
Wesley Jefferson Oliveira Alcoforado (*NUGEN-UECE*)
João David de Lira (*NUGEN-UECE*)
Dácio Barros Tavares (*NUGEN-UECE*)
Ana Luiza Bessa de Paula Barros (*LARCES-UECE*)
Diana Magalhães de Oliveira (*NUGEN-UECE*)

Short Abstract: In order to provide a convenient bioinformatics resource covering available information about the flagellum (an organelle of intriguing motile features), an integrated relational database has been created. FlagelLink (http://nugen.lcc.uece.br/lpgate/?p=flagdb) gives this first insight on compiling the modularity of flagellar genes/proteins for allowing further analyses.

Long Abstract:

BACKGROUND. The flagellum is an intriguing, well conserved, propulsive organelle used by many unicellular organisms (and differentiated cells of multicellular organisms) to move through a liquid or semi-fluid medium. By discovering more about the components and the nanostructured systems that maintain the flagellum, it might be possible to unveil new details regarding one of its key features: motility. Flagellated organisms (or cells) – like the pathogenic protozoan Leishmania spp. or a mammalian spermatozoid, for example– rely on a correct motility to achieve their main functions. In order to provide a convenient bioinformatics resource covering all available flagellar information, an integrated database has been created to link data from several different sources and we have named it FlagelLink. Combining different types of data from multiple databases is a key feature, since information about a given biological entity is often scattered across many different databases, which, in turn, often contain redundant or over-lapping information. Thus, database integration allows cross-validation and verification for which a relational database structure is essential. Sequence database searches can also be remarkably useful for finding the function of genes whose sequences have been determined in genome or transcriptome projects. A common reason for performing a database search with a query sequence is to find a related gene or protein in another organism. Alternatively, a query sequence of known function may be used to search through sequences of a particular organism to identify a gene/protein that may have the same function. As bioinformatics databases grow in size, and as biological questions grow in scope, the point-and-click style of database mining becomes less and less practical in such an environment. Instead, both computational and molecular biologists seek discoveries by using programs and complex queries to mine databases. These programs seek new patterns and generalizations by issuing queries to one or more databases to select, combine, and compute over millions of data records. In this sense, specific/organellar databases, such as the flagellar one we have structured, are increasingly necessary to improve the ability to extract valuable information out of the vast amount of unorganized existing data about the flagellum. METHODS. The relational schema employed was version 3.x of the Genomics Unified Schema, GUS (http://gusDB.org) and our database management system is PostgreSQL version 8.1. The Web interface to the database was produced via a Java servlet. A website was designed to include 'one click' access to the most commonly used features of the database. In addition to the relational database, several additional applications are provided such as BLAST and a variety of custom PERL scripts to facilitate data mining and text searches. In order to feed our local flagellar DB, we have started searching for new or unknown motifs/patterns that appear with highest frequencies in flagellar genes/proteins. Initially, we gathered these flagellar gene/protein sequences from public datawarehouses (NCBI, GeneDB, PDB, UniProt and Swiss-Prot). Each entry in our database is identified by FASTA format columns which correspond to the same ones in NCBI, like unique name and accession number. Then, we removed quotation marks and some other characters that might interrupt the processing. From these gathered data, we partitioned the components in accession number, unique name, description and sequence. Numbers were put inside an integer type column, and the remaining elements were considered as strings (character chains). Ater that, we were able to analyze and perform datamining tasks for recognizing the existing patterns (i.e., frequent, repeating or redundant motifs). Then, it was necessary to find out if motifs are already catalogued (or annotated as flagellar-related). Comparison to standard domain annotations at Pfam and SMART is a direct way to do this. Applying the principle that if a new/unknown protein sequence is similar to previously well known protein sequences, then good hypothesis about the function and structure of this new/unknown protein can be quickly established on the basis of the known protein, we have listed several motifs not previously annotated as flagellar-related, but consistently present in known flagellar proteins. Accordingly, we must verify if it is a valid assumption by using proper tools like MEME (http://meme.sdsc.edu/meme/) that is a tool for discovering motifs in a group of related DNA or protein sequences. MEME takes as input a group of DNA or protein sequences (the training set) and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width and description for each motif. RESULTS: Many protein families have already been identified, for instance globins, homeo domains and protein kynases. However, today very little information about relationships between sequences or family membership is contained in any of the available sequence databases. In such a scene, it is extremely relevant, in terms of biological significance, to provide additional information towards families of flagellar proteins. FlagelLink gives this first insight on building a bioinformatics platform for integrating knowledge about flagellar sequences and for further establishing an integrated view of a possible super-family of flagellar proteins. The integration of genomics, proteomic and EST datasets in FlagelLink permits users to construct combined queries for genes on several flagellated species (including Chlamydomonas and Leishmania spp., among others) based on the available evidence of expression, both at the RNA and protein levels. This is particularly important for all trypanosomatid genomes (for which proteomic and EST data are yet lacking or limited) since only 50% -60% of the predicted protein coding genes have been assigned a putative function. FlagelLink database (available at http://nugen.lcc.uece.br/lpgate/?p=flagdb) is designed to compile the modularity of each flagellar gene/protein as accurately as possible, thus allowing the biologist user to discover new patterns and motifs by means of local (and/or external) datamining from the sequences available either at FlagelLink or at other main repositories.