PROTMINE: a web based tool for the interpretation of clinical proteomic experiments

Giacomini M[(1)], Petretto A[(2)], Ravaschio S[(1)], De Nadai S[(1)], Melioli G[(2)]

[(1)] Department of Communication, Computer and System Science,University of Genoa
[(2)] Giannina Gaslini Institute; Genoa

Motivation

Humans have been "manually" extracting information from data for centuries, but increasing data volumes in modern times has called for more automatic approaches. As data sets and information extracted from them have grown in size and complexity, direct hands-on data analysis has increasingly been supplemented and augmented with indirect, automatic data processing using more complex and sophisticated tools, methods and models. The proliferation and increasing power of computer technology has helped data collection, processing, management and storage. However, the captured data needs to be converted into information and knowledge to become useful. The development of Internet has allowed to collect the data also in Biological area, especially in Proteomics. Several database are now available for protein description, but they are controlled by different organizations and with different purposes. The main examples of these data bases are UNIPROT (central archive for sequences and functions, merging information from Swiss-Prot, TrEMBL and PIR). Another correlated DB is VEGA (Vertebrate Genome Annotation, central archive for the results of the Human Genomic project and other similar projects). All these archives manage specific keys to search data within them, by this way a complete navigation within such a complex material is very difficult without specific tools. A good starting point was set up by EBI (European Bioinformatics Institute) which maintains a unique identification code: IPI (International Protein Index) that is becoming the unique worldwide identification for each protein. The International Protein Index (IPI) provides a top level guide to the main databases that describe the human, mouse and rat proteomes. IPI is built from the protein sequence data taken from the UniProt Knowledgebase, Ensembl and RefSeq databases, which are combined to create proteome sets for each species that combine a level degree of completeness with a low level of redundancy. Stable identifiers (with incremental versioning) allow the tracking of sequences in IPI between IPI releases, while cross-references are provided between equivalent entries in the source databases.

Methods

With the premises above it is clear that we need to fully get information from these coordinated data bases in order to correctly interpreter experimental results coming from clinical samples. The discriminatory power of sequence technology is continuously increased, so that the number of recognized proteins for each sample is already quite high, but it is foreseeable that in near future these numbers will be so high that no direct human interpretation will be still possible. The proposed structure is a web based tool because with this support it is extremely easy to share data and information among collaborating teams in different locations. The SSL technology is now easily adoptable in order to protect experimental data before publication, so that the use of such web based tools can be recognized as reliable. By this way collaboration among different groups will be encouraged with complete data sharing in order to achieve more stable scientific conclusions. As shown before the organization of the proteomic community is of high level also as regards computer data management, so that many web based tools are available. Specifically, EBI support data integration with their databases in many ways. We decided to use web services maintained by EBI servers. This decision is due to their complete and non redundant content , their speedy answer, completely acceptable within the time requirements of this project. At DIST Server, a MS SQL Server DB has been set up in order to collect experimental data (mainly list of proteins identified in clinical samples) and to correlate them with relevant clinical information of the sample. A web based interface allows people from clinical laboratory to submit the collected lists and to compare results within different experiments. EBI web services are interrogated to collect proteomic data chosen by experimental people in order to support data mining reasoning rules.

Results

At present we are in the testing phase for data collection and comparison. Specifically more than 150 protein lists have been successfully uploaded by people from clinical laboratories of

the Gaslini Institute. In this phase we are comparing these lists mainly according to cellular protein location and protein function in order to verify if significant differences in the profiles of these characteristics can be correlated with specific clinical status in which the samples are collected. Further investigation with modern data mining tools like Artificial Neural Networks will be tried.

Contact : Mauro.Giacomini@dist.unige.it