Pacific Symposium on Biocomputing 12:127-132(2007)

BIOINFORMATICS DATA PROFILING TOOLS: A PRELUDE TO METABOLIC PROFILING

NATARAJAN GANESAN, BALA KALYANASUNDARAM, MAHE VELAUTHAPLLAI

Department of Computer Science, Georgetown University, 3900 Reservoir Rd NW Washington DC 20057 USA

The term *metabolic profiling* is often used to denote the systematic characterization of the unique biochemical trails or fingerprints left behind by cellular processes. Advances in computational biosciences are often invaluable in dealing with the huge amount of raw data generated from the countless biochemical intermediates that flood the cell at any given time. As a prelude to metabolic profiling, it is essential to completely profile and compile all related information about the genetic and proteomic data. *Profiling tools* in bioinformatics refer to all those software (web based and downloadable) that compile all related information in single user-interfaces. Generally, these interfaces take a query such as a DNA, RNA, or protein sequence or *keyword*; and search one or more databases for information related to that sequence. Summaries and aggregate results are provided in a single standardized format that would otherwise have required visits to many smaller sites or direct literature searches to compile. In other words they are software portals or gateways that simplify the process of finding information about a query in the large and growing number of bioinformatics databases.

1.

1.1. *Contents*

1.2. *Introduction and usage*

The "post-genomics" era has given rise to a range of web-based tools and software to compile, organize, and deliver large amounts of primary sequence information, as well as protein structures, gene annotations, sequence alignments, and other common bioinformatics tasks. In general, there exist three types of databases and service providers. The first one includes the popular public-domain or open-access databases supported by funding and grants such as

NCBI, ExPASy, Ensembl, and *PDB*. The second one includes smaller or more specific databases organized and compiled by individual research groups Examples include the *Yeast Genome Database*, *RNA database*. The third and final one includes private corporate or institutional databases that require payment or institutional affiliation to access.

Figure 1 A typical example of a *keyword* profiling tool - *Entrez*

Typical scenarios of a profiling approach become relevant, particularly, in the cases of the first two groups, where researchers commonly wish to combine information derived from several sources about a single query or target sequence. For example, users might use the sequence alignment and search tool *BLAST* to identify homologs of their gene of interest in other species, and then use these results to locate a solved protein structure for one of the homologs. Similarly, they might also want to know the likely secondary structure of the mRNA encoding the gene of interest, or whether a company sells a DNA construct containing the gene. Sequence profiling tools serve to automate and integrate the process of seeking such disparate information by rendering the process of searching several different external databases transparent to the user. The importance of data profiling assumes significance in the burgeoning field of metabonomics which is soon likely to become a humungous warehouse of selectively processed data. *Seamless integration* of all relevant data should then become the buzzword to define the future of metabonomics.

Many public databases are already extensively interlinked so that complementary information in another database is easily accessible; for example, Genbank and the PDB are closely intertwined. However, specialized tools organized and hosted by specific research groups can be difficult to integrate into this linkage effort because they are narrowly focused, are frequently modified, or use custom versions of common file formats. Advantages of sequence profiling tools include 1) the ability to use multiple of these specialized tools in a single query and present the output with a common interface, 2) the ability to direct the output of one set of tools or database searches into the input of another, and 3) the capacity to disseminate hosting and compilation obligations to a network of research groups and institutions rather than a single centralized repository

1.3. *Keyword based profilers*

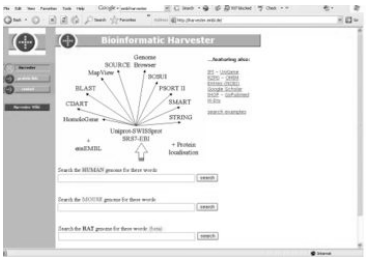

Figure 2 The network of Bioinformatic Harvester

Profiling tools based on *keyword searches* are essentially 'search engines' that are highly specialized for bioinformatics work, thereby eliminating a clutter of irrelevant or non-scholarly hits that might occur with a traditional search engine like Google. Most keyword-based profiling tools allow flexible types of keyword input, accession numbers from indexed databases as well as traditional keyword descriptors. For example, the NCBI search engine *Entrez*[2]segregates its hits by category, so that users looking for protein structure information can screen out sequences with no corresponding structure, while users interested in perusing the literature on a subject can view abstracts of papers published in scholarly journals without distraction from gene or sequence results. The *Pubmed* biosciences literature database is a popular tool for literature searches but is now a competitor with the more general *Google Scholar*[3].

Keyword-based data aggregation services like the *Bioinformatic Harvester* [4,5]performs provide reports from a variety of third-party servers in an *as-is* format so that users need not visit the website or install the software for each individual component service. This is particularly invaluable given the rapid emergence of various sites providing different sequence analysis and manipulation tools. Another aggregative web portal, the *Human Protein Reference Database* (Hprd), contains manually annotated and curated entries for human proteins. The information provided is thus both selective and comprehensive, and the query format is flexible and intuitive. The pros of developing manually curated databases include presentation of proofread material and the concept of 'molecule authorities' to undertake the responsibility of specific proteins. However, the cons are that they are typically slower to update and may not contain very new or disputed data.

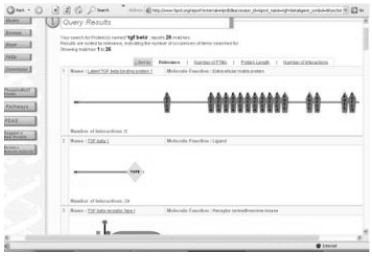

Figure 3 The HPRD is manually curated

1.4. *Sequence data based profilers*

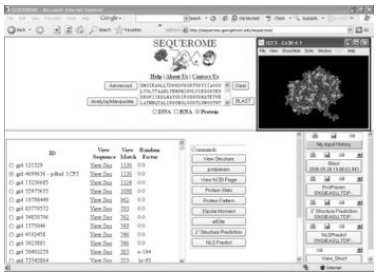

Figure 4 Display of sequence profiling features on a SEQUEROME browser

A typical *sequence profiling tool* carries this further by using an actual DNA, RNA, or protein sequence as an input and allows the user to visit different web-based analysis tools to obtain the information desired. Such tools are also commonly supplied with commercial laboratory equipment like gene sequencers or sometimes sold as software applications for molecular biology. In another public-database example, the BLAST sequence search report from NCBI provides a link from its alignment report to other relevant information in its own databases, if such specific information exists. For example, a retrieved record that contains a human sequence will carry a separate link that connects to its location on a human genome map; a record that contains a sequence for which a 3-D structure has been solved would carry a link that connects it to its structure database. *SEQUEROME*, a public service tool, links the entire BLAST report to many third party servers/sites that provide highly specific services in sequence manipulations such as restriction enzyme maps, open reading frame analyses for nucleotide sequences, and secondary structure prediction. The tool provides added advantage of tabbed browsing interface to track user operations and thus carry a project to its completion within one browser interface. The consequent evolution of such profilers would thus include ability to customize and automate processing of sets of sequence data. Though the presence of sequence based profilers are far and few in the present scenario, their key role will become evident when huge amounts of sequence data need to be cross processed across portals and domains.

1.5. *Microarray data profiling*

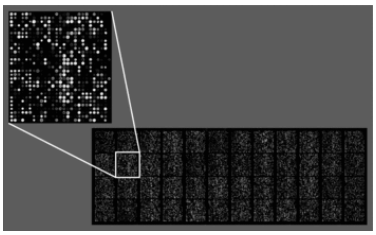

Figure 5 Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail.

Specialized software tools for statistical analysis to determine the extent of over- or under-expression of a gene in a microarray experiment relative to a reference state have also been developed to aid in identifying genes or gene sets

associated with particular phenotypes. One such method of analysis, known as Gene Set Enrichment Analysis (GSEA), uses a *Kolmogorov-Smirnov-style* statistic to identify groups of genes that are regulated together [1]. This third-party statistics package offers the user information on the genes or gene sets of interest, including links to entries in databases such as NCBI's GenBank and curated databases such as *Biocarta* and *Gene Ontology*.

1.6. *Future growth and directions*

The proliferation of diverse bioinformatics tools for genomic and proteomic genetic analysis has led to great advances in helping researchers identify and categorize genes of interest. However, this proliferation can also complicate the user interfaces for advanced users, while confusing and frustrating new users. This is so at time when metabonomics is beginning to spread its wings and branch out into a new area. The importance of such profiling tools is thus likely to expand very rapidly into other areas like metabolite profiling, and modeling dynamic living systems. This will occur as researchers and investigators begin to their directly upload raw information into better interfaces for more comprehensive profiles.

Future tools are, thus, likely to evolve into interfaces that *seamlessly integrate* information from different user defined portals/services e.g. mass spectrometric/NMR databases. They are also likely to allow inputs from images (even sounds) of experimental data. For example, a researcher might want upload the 3D image of a molecule and look for possible targets and genes. The possibilities are endless and the *bio-maze* is just beginning to get interconnected. Tools like SEQUEROME[6] are a step in this direction. As the information pyramid continues to grow and re-assemble itself, new generations of single-interface systems like those described above are bound to spawn a range of similar tools that would address the specific needs of different research groups.

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545-50.

2. *Biomedical language processing: what's beyond PubMed?* Mol Cell. 2006 Mar 3;21(5):589-94.

3. *Google versus PubMed*, Ann R Coll Surg Engl. 2005 Nov;87(6):491-2.

4. *'Harvester': a fast meta search engine of human protein resources.* , Bioinformatics. 2004 Aug 12;20(12):1962-3.

5. *'Harvester': a fast meta search engine of human protein resources.* , Bioinformatics. 2004 Aug 12;20(12):1962-3.

6. *Web-based interface facilitating sequence-to-structure analysis of BLAST alignment reports*, Biotechniques. 2005 Aug;39(2):186