BIOINFORMATICS TOOLS FOR GENE LIST ANALYSIS

Imer Muhović*, Larisa Bešić, Adna Ašić, Serkan Dogan, Osman Doluca

International Burch University, Department of Genetics and Bioengineering
*Corresponding author: imer91@gmail.com

ABSTRACT

The advent of the era of high-throughput sequencing has brought a wealth of biological data to researchers, but the vastness of the available data has created a demand for tools that could be used to analyze it. One such type of tools are gene set analysis tools, that take a list of genes that were found to be up or down regulated during an experiment. For the sake of simplicity this review focuses solely on freely available web based tools that have been published or have undergone significant updates in the last 5 years. This review is meant to assist tool developers to better understand the needs of the end-users, and in it we look at the currently available gene list analysis tools, their strengths and weaknesses, and offer suggestions for their improvement.

Key words: microarray, gene set, systems biology, enrichment, gene ontology

INTRODUCTION

Many modern molecular biology experiments result in the production of a list of important molecules. These molecules may be up/down regulated genes obtained from microarray or RNA-seq experiments, or a list of SNP – containing genes. The issue that is created by such lists is in the length of them. Your average co-expression experiment results in a list of hundreds or thousands of „interesting“ genes, and determining the biological significance of such a list is very difficult, as it requires either significant knowledge about the metabolic process being investigated, or it requires the researcher to conduct an extensive literature search to answer questions such as „What does this gene do? Where is it expressed? Does it interact with other genes? Is it linked to a particular disorder?“ Manually performing such a task would be time consuming and tedious, costing the researcher precious time and resources.

To save the time and sanity of researchers undertaking such experiments various tools for annotation enrichment (also known as pathway analysis) have been developed. These tools map genes and proteins to their associated biological annotations (gene Ontology terms, or pathway membership) and then compare the frequency of such terms in the given gene list, with a background list to identify the over expressed, or under expressed terms in the list, following the assumption that such terms are important to the metabolic process that is being studied. As an example, imagine that in a list obtained by a microarray experiment, 20% of the genes are tumor suppressor genes, while in a „normal“ tissue only 5% are. By using standard statistical method we can determine that tumor suppressor genes are enriched in this list, and therefore play an important role in the biological process we are investigating.

Most review articles in this field divide tools according to the statistical method that they use. There are three most common ones: Singular Enrichment Analysis (SEA), Gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA).(Huang, Sherman, & Lempicki, 2009)

SEA – compares annotation terms one by one with a list of interesting genes for enrichment. A p-value for enrichment is obtained by comparing the frequency of an annotation term with the frequency of that term appearing by chance. All terms that are beyond the cut-off value are said to be enriched. The drawback of this approach is that it ignores the hierarchical relationship between GO terms, and results in large lists of enriched terms due to the fact that it treats similar terms as though they were unique.

GSEA – these methods take as an input not only the list of interesting (up or down regulated genes) but all of the genes obtained by an experiment. It functions best in experiments in which two tissue types are compared, because it requires a quantitative value (change in differential expression) for each gene in order to rank them by significant enrichment. A so called maximum enrichment score (MES) is calculated from the ranked list of genes in an annotation category, and enrichment p-values are determined by comparing the MES of the term to a randomly generated MES distribution. To put it in simpler terms, GSEA determines if genes that share a biological annotation (for example belong to the same pathway) are randomly distributed in the gene list (and therefore not significantly attributing to a change in phenotype), or if they are overrepresented in a part of the list (top or bottom, according to fold change, or differential expression), which would indicate that they play a role in the pathway that is being studied.(Subramanian et al., 2005)

MEA – seeks to use the relationships between different annotation terms to remove the redundancy, or underrepresentation of important terms that may be caused by SEA and GSEA methods. They seek to improve sensitivity and specificity by using composite annotation terms. The issue with them may be found if they use only a single information source, usually GO.

Most current tools seem to have switched to using MEA as opposed to SEA, as the link between different levels of annotation has become clearer, and the integration of different databases has become easier.

Molecular interaction network present the easiest, most intuitive way of representing such large and complex datasets, and several curated databases already exist that link all known binary protein interactions, as well as enrichment data, whether extracted from literature of HT experiments.

DATABASES

To better study and keep track of all known pathway data several databases have been constructed. A key difference between databases lays in their data acquisition methods. We can separate curated databases from automatic ones; by the way the data are added in the database, either by trained experts or via automatic methods.

Each has its own advantages, shallow curated databases have larger network coverage, while curated ones have higher quality of data, but still data capture errors such as false positives in the data still can't be excluded.

Another difference is the data source, as some databases take their data from peer reviewed literature, while secondary databases look to integrate primary databases and thus become a one-stop shop for all your protein interaction needs.

One we have a list of PPI interactions we need methods to visualize this data and extract the useful data from them. Due to the large number of PPIs in a possible network the results usually look like a giant ball of yarn that is difficult to interpret so visualization techniques offered by the tool play an important role.

H-InvDB (http://www.h-invitational.jp/) is a human gene database first published in 2004.It contains 244,709 human DNA sequences, and provides the user with a broad variety of tools for genome analysis. According to the authors analysis 19,309 annotated genes were found to be specific to H-InvDB and not to be found in RefSeq or Ensembl.(Takeda et al., 2012)

PINA (The Protein Interaction Network Analysis) is an integrative resource that combines data from six manually curated public databases, and offers a set of tools for network construction, filtering, analysis and visualization. It offers protein-protein interaction (PPI) network construction, by clustering approaches from an interactome constructed for six available model organisms. All identified terms are annotated using GO terms, KEGG pathways, Pfam domains and MsigDB data.(Cowley et al., 2011)

STRING is a database that seeks to provide biologists with a global perspective on as many interactions from as many organisms as possible. It scores both known and predicted interactions, and offers the users tools for statistical analysis and enrichment analysis of queried terms.(Franceschini et al., 2012)

GeneSigDB (http://www.genesigdb.org or http://compbio.dfci.harvard.edu/genesigdb/) is a database of gene signatures collected manually from published literature, focusing on cancer studies, as well as immune cells, stem cells and lung disease. It is an excellent tool for prognostic analysis of cancer and related diseases, or use as an gene set enrichment tool. The visualization of enriched terms is performed via heatmap that provides us with publication-quality images, and GeneSigDB allows us to download data in .gmt file format that can be later used for additional gene set enrichment analysis.(Culhane et al., 2011)

IntAct is an open-source, molecular interaction database that contains data manually curated from literature or raw depositions. It has two levels of curation, and contains around 275 000 interactions, collected from over 5000 publications. A recent upgrade has brought it a visual display of data, which are downloadable in multiple formats.(Kerrien et al., 2011)

The MetaCyc database (http://metacyc.org/) is a freely accessible resource that contains data from metabolic pathways and enzymes from all domains of life. MetaCyc pathway data is obtained experimental and small-molecule metabolic pathways and are curated from the primary scientific literature. Currently there are more than 1800 pathways derived from over 30 000 publications, making MetaCyc the largest curated collection of metabolic pathways. (Caspi et al., 2011)

IPAVS (Integrated Pathway Resources, Analysis and Visualization System) is a manually curated database of known protein pathways. It combines several publicly available pathway databases, and provides the tools to filter search and analyze biological pathways. It is freely available, interactive and integrated pathway database which is designed to address the needs of bench biologists, computational biologists and physicians. It offers biologists a single point of access to several manually curated pathway resources, in addition to its own expert-curated pathways that are in standard format. (Sreenivasaiah, Rani, Cayetano, Arul, & Kim, 2011)

NETWORK CLUSTERING

Proteins are usually represented as nodes, and interactions as vertices, giving us a ball and stick model of interactions. One of the main aims of pathway analysis strategies is to discover clusters of proteins that perform a similar function. This is mostly done by network topology as highly interconnected nodes form clusters, and the basic assumption is that clusters identify proteins that share a common function. Issues that may arise from analyzing pathways in this fashion is that large networks tend to resemble balls of yarn, due to having hundreds of nodes and vertices, thus making the inference of biological data from them very hard, and confusing.

NETWORK ANNOTATION

Annotation of nodes and edges is usually needed to make some sense of the information found in PPINs. The annotations may include info about the method by which the interaction was detected, some confidence scores and similar parameters. Gene Ontology project is the most widely used source of extra information that can be used in network analysis. It's creates a hierarchical list of terms called Ontologies that covers three independent biological domains: 1 - Cellular Components 2 - Biological Processes 3- Molecular Function.(Ashburner et al., 2000)

This enables us to highlight the proteins that perform the same function, thus allowing a functional representation of a network, usually GO is combined with cluster detection to provide greater interpretation of a network.

GENE LIST ANALYSIS TOOLS

Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool

Enrichr is a web based tool that takes in as an input a list of differentially expressed genes, and produces lists of enriched terms. The authors have solved the issues that arise when using only one source of enrichment data, by using 35 gene set libraries split into six groups, with each containing different data about different enrichment terms. It uses

1) ChEA (The ChIP-x Enrichment Analysis Database), it's own resource of putative transcription factor targets created from publications that report experiments of profiling mammalian DNA binding transcription factors. ;

2) position weight matrices (PWMs) from TRANSFAC and JASPAR ; that were used to scan all promoter regions (-2000 to +500 from the start of transcription) of all human genes, they kept all 100% matches to the consensus sequence between a factor and a target gene.

3) target genes generated from PMWs downloaded from the UCSC genome browser , because it produces different results compared to the ones mentioned above

4) transcription factor targets extracted from the ENCODE project . In addition, the two other gene-set libraries in the transcription category are gene sets associated with:

5) histone modifications extracted from the Roadmap Epigenomics Project ; and

6) microRNAs targets computationally predicted by TargetScan .

It provides three different statistical measures of the results, of which one is the Fischer exact test, the other an in-house variation and last a combination of the two. The authors performed a quality evaluation of these methods in their original paper. Enrichr provides many different options for visualizing the data, one of which is a grid of squares, with the most enriched elements being colored more brightly when compared to the rest. It also allows the visualization in the form of a list of enriched terms, bar graph, network and table.

An advantage of Enrichr over other programs of the same type is it's availability and modern design, it's available as a mobile application for smartphones and tablets, and the web-interface is clear, and intuitive. The authors tested the software by comparing nine cancer cell lines and found an upregulation in the PRC2 polycomb group target genes.(Chen et al., 2013)

Network2Canvas is a network visualization program that makes it easier to visualize large protein-protein interaction networks, and enrichment terms. The most common issue with using ball and stick models of PPI networks is that larger networks tend to end up looking like balls of yarn, making it very difficult to visually analyze the properties of the network. Network2Canvas works around this issue by placing the nodes on a square toroidal canvas, the nodes are then clustered on the canvas via simulated annealing in order to have the maximum number of local connections, and their brightness is set to correspond to the local fitness of the node.

This software takes as input a list of differentially expressed genes, or a list of drugs and outputs a set of enriched terms including drug side effects, common pathways etc. The website is accompanied by a video tutorial on how to use the program, and offers a variety of possible canvases, for example Kinase Enrichment Analysis or KEGG pathways, and more. Overall N2C is a very useful and intuitive tool for molecular data analysis, especially for larger lists of genes or drugs.(Tan, Chen, Dannenfelser, Clark, & Ma'ayan, 2013)

Genes2FAN: Proteins interaction studies are mostly done by analyzing binary protein interactions, but these are not the only ways two genes, or their protein products can interact. The authors of this program used knowledge on the shared properties of genes from diverse sources to create functional association networks (FANs), to allow researchers to identify additional interactions between groups of genes, which are not immediately obvious from PPI networks.

G2F uses a database of 14 FANs, and large scale PPI networks to create subnetworks that can connect lists of human and mouse genes. Lists of genes are taken as an input to produce a subnetwork, using a ranked list of intermediate genes that connect the genes from the queried list. This web application offers a powerful new approach to analyzing gene associations, as it can find the intermediate parts of a pathway, and thus allow us to observe a greater, clearer picture of a molecular process.(Dannenfelser, Clark, & Ma'ayan, 2012)

Sets2Network is a tool created to allow the creation of interaction networks by analyzing the co-occurrence of entities in related sets. It gives us a general method for inferring networks by repeated observation of sets of related terms. It interprets the frequency of the occurrence of the link as the probability that it is present in the real-world network.

This tool has usages outside the realm of biology, as it can create a network from any file given in the GMT (Gene Matrix Transpose) format, for example it can be used to create a network of co-authorship by taking in a GMT file of publications and authors, or predict direct PPI from HT MS data. (Clark, Dannenfelser, Tan, Komosinski, & Ma'ayan, 2012, p. 2)

S2N can output the data in various formats, so subsequent analysis can be performed on the data, using additional visualization tools such as yEd.

DAVID (Database for Annotation, Visualization and Integrated Discovery, available at http://david.abcc.ncifcrf.gov/) is one of the oldest and most well-known web-based bioinformatics resources for the functional interpretation of gene/protein lists. It has been cited over 6000 times since its initial publication. It takes inputs in list form and allows the user to perform gene-term enrichment analysis, visualization of the gene-term relationships, search for related genes, pathway analysis and much more. They have recently published DAVID-WS (Web service) an API (application programming interface) which allows for the programmatic automation of requests to DAVID, and thus the easier automation of tasks, without the need for human interactions.(Jiao et al., 2012)

EnrichNet is a web-based tool created in order to address the current limitations of gene set analysis tools. Most GSA tools use the over-representation-based enrichment analysis method which uses the overrepresentation of a gene list of interest in a reference list via a statistical test (usually Fisher's exact test) as proof of biological significance. The issue with this approach is that it low power of discrimination, and significant variance with changes in overlap size, among others. EnrichNet uses an graph-based statistic approach to analyze gene sets, via exploiting information from molecular network structures of, and offers interactive visualization of network sub-structures. It offers integrated data sources (molecular interaction data, pathway and tissue-specific gene expression data) and uses graph-based statistical analysis and forced – directed layout generation to provide a clearer and more detailed understanding of the gene set interactions. It uses a minimalist interface, with clear output, and we refer the reader to the paper for more information.(Glaab, Baudot, Krasnogor, Schneider, & Valencia, 2012)

GeneCodis is a tool for enrichment analysis, available since 2007, its newest version offers a more concise output and removes some redundancy, via summarizing of significantly enriched terms, they also expanded the original application, by adding new sources of information, such as genetic diseases, gene-drug interactions and PUBMED information. GeneCodis offers a very customizable input, as it integrates data from several organisms, which is rather rare as most of the enrichment analysis tools focus only on humans. Its capable of filtering the output.(Tabas-Madrid, Nogales-Cadenas, & Pascual-Montano, 2012)

GeneMANIA (http://www.genemania.org) is a web-app for gene list analysis. Given a list, it will extend it with functionally similar genes, obtained from genomics and proteomics databases. It's capable of finding genes of similar function, and those that are most likely to interact with the ones in the list. It supports multiple organisms, and integrates hundreds of datasets from GEO, BioGRID, IRefIndex and I2D.(Zuberi et al., 2013)

Graphite Web: A new web-app for pathway analysis and visualization, that takes as input gene lists from microarray or RNA-seq experiments. It combines topological methods with multivariate pathway analyses and provides a clear network visualization tool, for efficient interpretation of expression experiment results. It works with three model organisms, and integrates two pathway databases.(Sales, Calura, Martini, & Romualdi, 2013)

CONCLUSION

While new tools are constantly arriving they individually don't see too much use or recognition, this may be due to low popularity, or just being hard to find. This lack of visibility makes it hard for researchers to test out new tools, as they rarely know that they even exist, and this leads to a lack of feedback for the tool makers, which in turn leads to a lack of improvement in the available tools. The usage of targeted internet marketing to possible users should be considered by future tool makers as a way of reaching out to new users, and obtaining feedback on their work. The current focus of enrichment analysis should probably be turned over to better visualization of datasets, as the ball and stick models are prone to looking like a ball of yarn if the input list is too large. Better visualization of data will allow for much easier analysis, and comprehension of experimental results.

API support is another issue, as most of the tools listed in this review rely on manual input of data, DAVID-WS is a nice exception to the rule. APIs could allow for easier testing and automation of enrichment analysis tools, thus simplifying and speeding up a biologist's workflow.

Standardization is another issue encountered in the use of these tools, as few of them support multiple formats of output files, while exception do exist, they are few and mostly consist of older, more established tools. While some tools do offer advantages over others, there exists no gold standard for enrichment analysis, with labs using whichever tools they prefer, this makes it hard to gauge the effectiveness of an approach, as only by repeat usage do the advantages of a tool become clearly apparent.

While we have covered some integrative tools, none of them offers the full package, a modern enrichment analysis tool should offer a customizable input, output, network visualization, different scoring systems, multiple output formats, and allow for publication quality images. The closest we have come to this are the tools from Maya'an Labs (Enrichr, S2N, N2L etc.), which provide a wide array of functionality, but still aren't seeing much use.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … others. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., … Karp, P. D. (2011). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research, 40(D1), D742–D753. doi:10.1093/nar/gkr1014

Clark, N. R., Dannenfelser, R., Tan, C. M., Komosinski, M. E., & Ma'ayan, A. (2012). Sets2Networks: network inference from repeated observations of sets. BMC Systems Biology, 6(1), 89.

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., … Wu, J. (2011). PINA v2.0: mining interactome modules. Nucleic Acids Research, 40(D1), D862–D865. doi:10.1093/nar/gkr967

Culhane, A. C., Schroder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., … Quackenbush, J. (2011). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Research, 40(D1), D1060–D1066. doi:10.1093/nar/gkr901

Dannenfelser, R., Clark, N. R., & Ma'ayan, A. (2012). Genes2FANs: connecting genes through functional association networks. BMC Bioinformatics, 13(1), 156.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., … Jensen, L. J. (2012). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research, 41(D1), D808–D815. doi:10.1093/nar/gks1094

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. Bioinformatics, 28(18), i451–i457. doi:10.1093/bioinformatics/bts389

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research, 37(1), 1–13. doi:10.1093/nar/gkn923

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics, 28(13), 1805–1806. doi:10.1093/bioinformatics/bts251

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., … Hermjakob, H. (2011). The IntAct molecular interaction database in 2012. Nucleic Acids Research, 40(D1), D841–D846. doi:10.1093/nar/gkr1088

Sales, G., Calura, E., Martini, P., & Romualdi, C. (2013). Graphite Web: web tool for gene set analysis exploiting pathway topology. Nucleic Acids Research, 41(W1), W89–W97. doi:10.1093/nar/gkt386

Sreenivasaiah, P. K., Rani, S., Cayetano, J., Arul, N., & Kim, D. H. (2011). IPAVS: Integrated Pathway Resources, Analysis and Visualization System. Nucleic Acids Research, 40(D1), D803–D808. doi:10.1093/nar/gkr1208

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Lander, E. S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545–15550.

Tabas-Madrid, D., Nogales-Cadenas, R., & Pascual-Montano, A. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Research, 40(W1), W478–W483. doi:10.1093/nar/gks402

Takeda, J. -i., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., … Imanishi, T. (2012). H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. Nucleic Acids Research, 41(D1), D915–D919. doi:10.1093/nar/gks1245

Tan, C. M., Chen, E. Y., Dannenfelser, R., Clark, N. R., & Ma'ayan, A. (2013). Network2Canvas: network visualization on a canvas with enrichment analysis. Bioinformatics, 29(15), 1872–1878. doi:10.1093/bioinformatics/btt319

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C. T., Bader, G. D., & Morris, Q. (2013). GeneMANIA Prediction Server 2013 Update. Nucleic Acids Research, 41(W1), W115–W122. doi:10.1093/nar/gkt533

Imer Muhović is a MSc student at the International Burch University. His main interests lie in bioinformatics and systems biology, and he is currently in the process of constructing a novel bioinformatics tool for sequence analysis, which will form his Master's thesis.