GOAPhAR: An Integrative Discovery Tool for Annotation, Pathway Analysis

Sachin Mathur[#,1], Mahesh Visvanathan*[,#,2,3], Stan Svojanovsky[2,4,5], Byunggil Yoo[2,4,5], Adagarla B. Srinivas[3], Gerald H. Lushington[3] and Peter G. Smith[2,4,5]

[1]*School of Computing and Engineering, University of Missouri-Kansas City, MO 64110, USA*

[2]*Department of Integrative and Molecular Physiology, University of Kansas Medical Center, Kansas City, Kansas 66160 USA*

[3]*Bioinformatics Core Facility, University of Kansas, Lawrence, KS 66047,USA*

[4]*Bioinformatics Core, Kansas IDeA Network for Biomedical Research Excellence (K-INBRE), Mailstop 3051, University of Kansas Medical Center, Kansas City, Kansas 66160 USA*

[5]*RL Smith Mental Retardation Research Center, University of Kansas Medical Center, Kansas City, Kansas 66160 USA*

Abstract: We have developed the web based tool GOAPhAR (Gene Ontology, Annotations and Pathways for Array Research), that integrates information from disparate sources regarding gene annotations, protein annotations, identifiers associated with probe sets, functional pathways, protein interactions, Gene Ontology, publicly available microarray datasets and tools for statistically validating clusters in microarray data. Genes of interest can be input as Affymetrix probe identifiers, Genbank, or Unigene identifiers for human, mouse or rat genomes. Results are provided in a user friendly interface with hyperlinks to the sources of information.

The tool is freely available at http://bioinformatics.kumc.edu/goaphar/

Keywords: GOAPhAR: Gene ontology, Annotations and pathways for array research.

1. INTRODUCTION

Microarrays are useful in profiling entire genomes of organisms under specific conditions [1]. Data generated are used to assess relationships among genes and to obtain a detailed understanding of underlying cellular processes. After the data are generated, probeset signals are filtered from noise and background. Normalization techniques [2] are then applied to minimize technical variation and probe subsets are selected for detailed analysis. Typically, an initial step in analysis is to obtain annotative information for the selected probesets [3]. Annotation can include; structural information such as chromosome location, sequence information, coding regions, or homologs in other genomes; functional information such as biological pathways; and associated publications. Fortunately extensive annotative information is freely available on the internet. However, because of its vast and heterogeneous nature, these resources are scattered among many sites, and it can be a daunting task to locate relevant information from the disparate sources [4].

While some applications are available that integrate annotative information, they are frequently limited with regard to the identifiers they use and the completeness of the information they provide. Many existing tools provide multi-dimensional information as a single instance that lacks logical integration, for example, displaying gene annotations, pathways and Gene Ontology on a single page. This makes it difficult for the user to understand and navigate through the results. Some of the applications require the user to enter one gene identifier at a time without providing comprehensive batch mode of analysis, making for tedious annotation.

The objective of this study was to develop a comprehensive tool to extract a wide range of annotative information from microarray data, and to provide this detailed information in a user friendly environment. Here we present a new web-based application that mines information from various sources, integrates this information, and presents it to the user in a logical and accessible format. The integrated information can be classified as 'Gene Annotations', 'Protein Annotations', 'Gene Ontology', 'Biological Pathways', 'Protein Interaction' and 'Statistical Validity'. All results are hyperlinked to their sources so that users can browse and extract information of their choice. It also provides links to results of existing tools that provide additional information thus giving the user comprehensive information at a single source. Importantly, the tool provides batch mode analysis, so the user can query multiple probe sets simultaneously and the results can be downloaded in the form of a text file.

2. METHODS & IMPLEMENTATION

Definition

GOAPhAR is an acronym for Gene Ontology, Annotation and Pathways for Array Research. These categories pro-

*Address correspondence to this author at the Department of Integrative and Molecular Physiology, University of Kansas Medical Center, Kansas City, Kansas 66160 USA; E-mail: mvisvanathan@ku.edu

#Authors have equally contributed.

1875-0362/09

vide information regarding gene identifiers, gene locations on chromosomes, gene nomenclature, gene symbols, gene ontology, protein identifiers, tertiary protein structures, protein interactions mined from literature, signaling and metabolic pathways and publicly available microarray datasets. It also provides a means of assessing statistical validity of clusters derived from microarray data. The Schematic diagram depicting the work flow in the GOAPhAR is shown in Fig. (**1**).

GOAPhAR Databases

The annotations have been classified as gene and protein annotations and extracted for human, mouse and rat from the NETAFFX annotation file [5] available from Affymetrix. Gene Ontology information is extracted from NETAFFX and Gene Ontology Annotation (GOA). Pathway information data sources are Kyoto Encyclopedia of Genes and Genomes (KEGG) [6], Signaling Pathway Database (SPAD) [7], GenMapp [8] and Panther [10]. Microarray datasets are available from Gene Expression Omnibus in NCBI. Protein interaction information has been extracted from the BIND database, protein structures from PDB database. The tool supports human, mouse and rat genomes with Genbank, Unigene or Affymetrix probe set identifiers. The cluster validation component of GOAPhAR makes use of well known statistical algorithms, namely Dunn's, Davies-Bouldin and silhouette indices [11].

GOAPhAR Web Interface

It was developed using PHP4. Additional scripts for information extraction were written in Perl, C and Java. The curated information is stored in MySQL database**.**

GOAPhAR Usage

GOAPhAR is accessible through the web and a free user account can be obtained after registering on the website. The website has been tested on IE, Mozilla, Firefox and Safari web browsers. The input to the system is a new line delimited text file with the probe identifiers. It consists of the previously mentioned aspects of data analysis, entitled 'Gene Annotations', 'Protein Annotations', 'Gene Ontology', 'Biological Pathways', 'Protein Interactions' and 'Statistical Va-

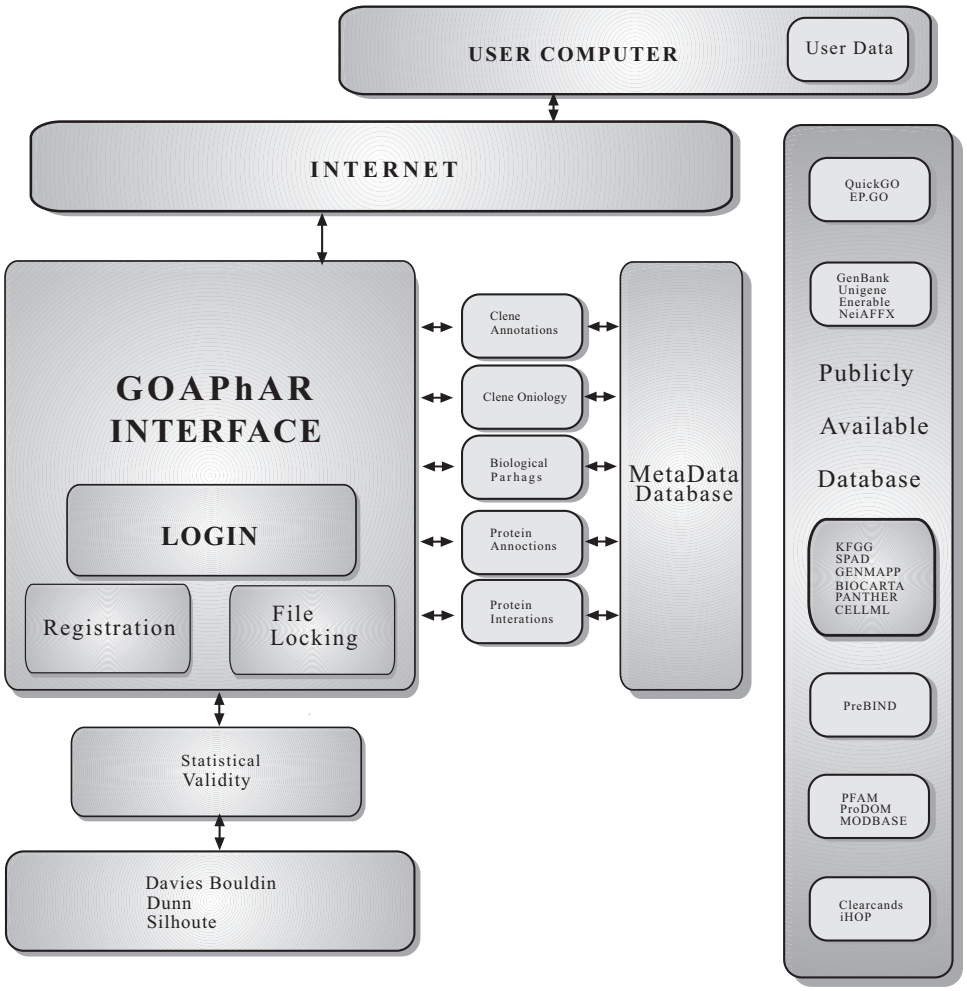

Fig. (1). Schematic diagram depicting the work flow in the GOAPhAR

lidity'. The user is asked to upload a text file that contains the identifiers for the genes, the genome to which it belongs, and the type of identifier. Once a user uploads the file and selects an option the system locks the file and the user can navigate through the entire system without having to upload the file again. The results are displayed in a tabular format and can be downloaded as a text file. Representation of the data analysis aspects of GOAPhAR and their sources from the World Wide Web is shown in Table **1** below.

3. RESULTS

In this section we describe the main analysis modules implemented in GOAPhAR.

Gene Annotations

Gene annotation is a critical feature as it gives the identification, position and functional characteristics of genes in a genome. GOAPhAR extracts information from Genbank [12], Unigene and NETAFFX [9]. In also provides gene titles gene symbols, reference transcript identifiers, associated homologs, enzyme commission numbers and location on chromosome as annotation. The identifiers from all the above mentioned databases are displayed, thus circumventing the problem of multiple identifiers being used for the same gene in different databases. These are hyperlinked to their sources, thus providing more detailed information. Additionally, links are provided to the popular web-based tools Genecards [13] and iHOP [14] which provide additional annotation. This system is capable of retrieving UniGene identifiers as well as other information related to a specific set of probe id's.

Gene Ontology

Gene Ontology provides relevant information on biological processes, molecular functions and the cellular components in which the gene products are involved [15]. This information is useful for determining additional functions of genes, relation to other gene products, and for comparative genomics. Gene Ontology information is obtained from the Gene Ontology Consortium and hyperlinks are provided to QuickGO [16] and Reactome [17] applications. The user can view the hierarchical nature of Gene Ontology in QuickGO and the gene products are ordered by the organisms in Reactome [17].

Biological Pathways

After examining the annotations and expression levels, the investigator typically may select a set of genes for additional statistical analyses such as principal component analysis and clustering. If the genes are co-regulated, it is of interest to determine if they share a common biological pathway. GOAPhAR integrates pathway information from KEGG, Biocarta, Genmapp, Panther, Spad and Cellml databases, thus providing the user with comprehensive information. Some of the pathways are redundant, as many pathways occur in two or more databases, but this is also useful as many pathway schemas are incomplete.

Protein Annotations

A protein domain is a structurally and functionally defined protein region. If a protein contains multiple domains then it may be involved in two or more functions. Protein families are subsets of protein domains with related structure and function. This information is obtained from Structural Classification of Proteins (SCOP) [18] and Protein Families (PFAM) [19] databases and the tertiary structure is obtained from Protein Databank (PDB) [20]. In addition it provides the protein reference identifiers from NCBI and protein identifiers from Swissprot [21].

Protein Interactions

There is abundant public literature providing information concerning interactions between proteins [22]. These interactions are experimentally defined or hypothesized, and can be very helpful in assessing the molecular significance of changes in gene expression. PreBind [23] is a literature mining tool that extracts protein-protein associations from Pubmed and classifies them in various interaction categories based on results of pattern matching. GOAPhAR extracts this information from the Prebind database and displays it to the user. The investigator can then access the relevant citations *via* hyperlink.

Cluster Validity

Clustering is used to identify patterns in the microarray dataset and is often used to find co-regulated genes. There are many algorithms that produce clusters of various granularities. Since a huge number of clusters is possible, the ap-

Table 1. Representation of the Data Analysis Aspects of GOAPhAR and their Sources from the World Wide Web

Feature	Description	Data Source
Gene Annotation	To analyze microarray data in relation to specific biological characteristics	NCBI, NETAFFX, NCBI GEO Datasets
Gene Ontology	Provides information on biological processes, molecular functions and cellular components.	QuickGO, Reactome
Pathways	Provides information regarding the pathways in which gene products interacts.	KEGG, SPAD, BIOCARTA
Protein- Protein Interaction	Provides information on protein to protein interactions mined from public literature	Prebind
Protein Annotations	Extracts information regarding the family of proteins the gene product belongs to and the domains in which it occurs.	Swissprot, PFAM, SCOP, PDB
Statistical Validity	Statistically validates clusters in microarray data.	DB Index, Dunn's Index, Silhouette Index

propriate cluster must be selected for further analysis. Davies-Bouldin, Dunn's and silhouette indices [24] provide good assessments with respect to intra- and inter-point separation. For example, a low Davies-Bouldin, high Dunn's index and silhouette close to 1 are considered to be a good indication of valid clustering. GOAPhAR implements these indices, allowing the investigator to both identify related genes and to vet their function and annotation using the text files.

The usage of the Gopahar is shown in Fig. (**2**) wherein the user can upload the probe set id's and retrieve various information related to it that includes gene annotation, protein annotations as well as pathway annotations.

Comparison with Other Tools

There are many web-based applications and desktop software programs that extract information regarding gene identifiers. Two of the web-based applications are Database Referencing of Gene Array Online (DRAGON) [25] and MicroArray Data Review and Annotation System (MADRAS) [26]. While these applications provide much useful information, DRAGON does not provide Gene Ontology information whereas MADRAS lacks protein annotations. Commercial software like GeneSpring provide pathway information from only a limited number of databases (i.e. KEGG and GenMapp). None of the above applications provides information on publicly available microarray datasets, protein structures or protein-protein interactions. GOAPhAR overcomes all these limitations and provides a structured and detailed analytical framework. GOAPhAR's functionality is currently being expanded to include additional genomes, tools that map expression profiles onto functional pathways, statistical tools for analysis and tools that mine protein interaction literature.

GOAPhAR provides detailed and comprehensive information from microarray data in a user-friendly and structured manner. Investigators can use the information to filter genes or perform detailed analyses on subsets of genes. GOAPhAR can exponentially reduce the time required for analyzing data obtained from gene profiling microarray experiments. We are in the process of adding extra functionality's to the gopahar server that would allow the user to group objects (for example probe set id with their respective proteins and their interactions. We are also in the process of adding extra functionalities to the output wherein the user can split the results.

4. CONCLUSIONS

GOAPhAR provides detailed and comprehensive information from microarray data in a user-friendly and struc-

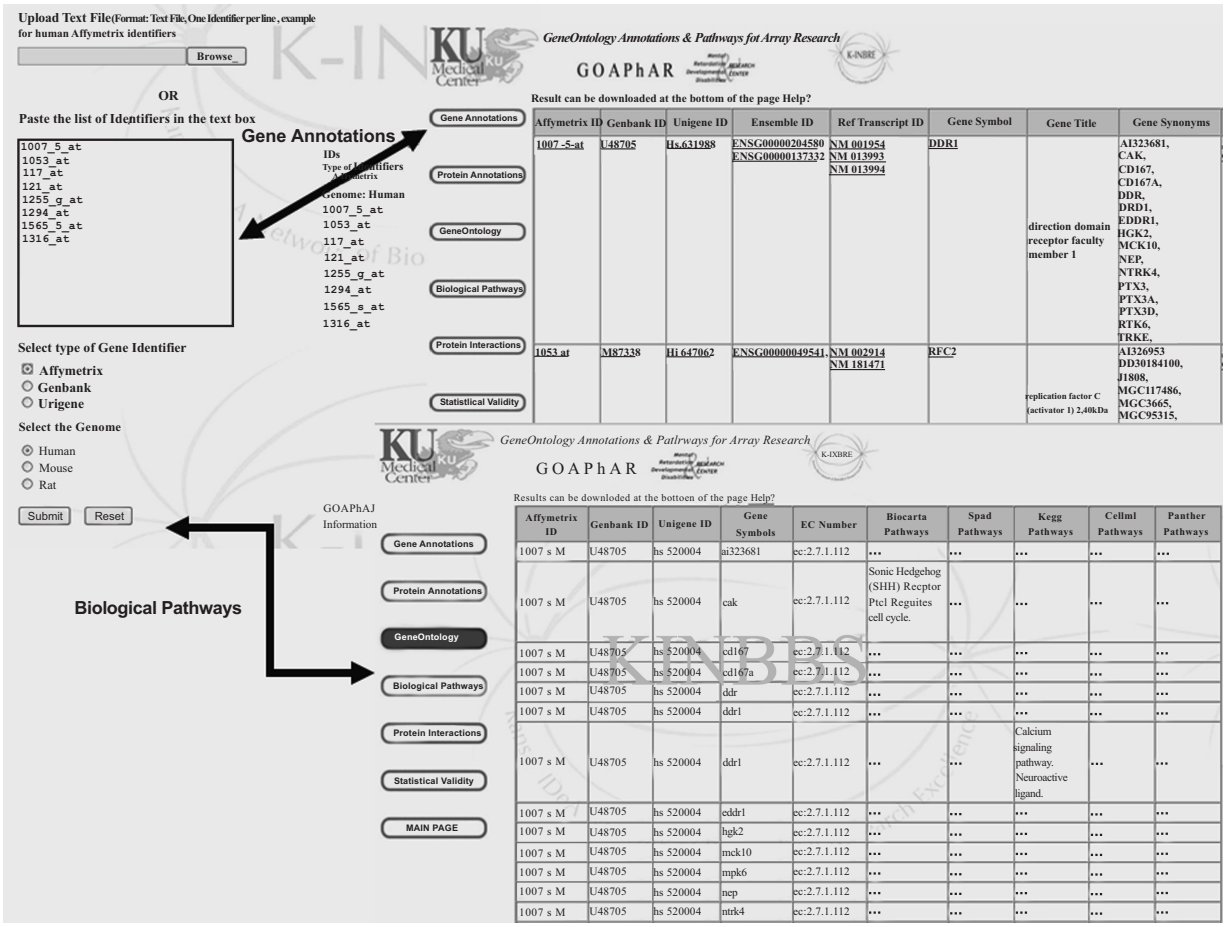

Fig. (2). Simple use age of the GOAPhAR server wherein the user provides the system with a list of probe set id's and retrieve various annotations related to it.

tured manner. Investigators can use the information to filter genes or perform detailed analyses on subsets of genes. GOAPhAR can exponentially reduce the time required for analyzing data obtained from gene profiling microarray experiments. GOAPhAR is useful in preliminary data analysis for finding gene/protein annotations, as well as for detailed analysis including functional pathway and protein interactions. The tool significantly increases efficiency of analysis of microarray data by providing information from many sources on a single interface, thus reducing time and effort. The tool is freely available at http://bioinformatics.kumc.edu/goaphar/

ACKNOWLEDGEMENTS

This work was supported by the K-INBRE, NIH grant number P20 RR016475 and Kansas IDDRC grant number P30 HD002528.

REFERENCES

[1] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol*., vol. 147, pp. 195-197, 1981.

[2] D.J Lockhart, "Expression monitoring by hybridization to high-density oligonucleotide array," *Nat. Biotechnol.*, vol. 14, no. 13, pp. 1675-80,1996.

[3] B.M Bolstad, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185-93, 2003.

[4] A. Riva "Comments on selected fundamental aspects of microarray analysis," *Comput. Biol. Chem.*, vol. 29, no. 5, pp. 319-36, 2005.

[5] M. Navarange, "MiMiR: a comprehensive solution for storage, annotation and exchange of microarray data," *BMC Bioinform.*, vol. 6, p. 268, 2005.

[6] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes", *Nucleic Acids Res*., vol. 28, no. 1, pp. 27-30, 2000.

[7] Signaling Pathway Database: http://www.grt.kyushu-u.ac.jp/spad/

[8] S.W. Doniger, "MAPPFinder: using gene ontology and genMAPP to create a global gene-expression profile from microarray data," *Genome Biol*., vol. 4, no. 1, p. R7, 2003.

[9] NetAFFX Annotation File: http://www.affymetrix.com/analysis/index.affx

[10] H. Mi, "The PANTHER database of protein families, subfamilies, functions and pathways," *Nucleic Acids Res*., vol. 33(Database issue), pp. D284-8, 2005.

[11] N. Bolshakova, F. Azuaje and P. Cunningham "An integrated tool for microarray data clustering and cluster validity assessment," *Bioinformatics*., vol 21, no. 4, pp. 451-5, 2005.

[12] D.A. Benson, "GenBank". *Nucleic Acids Res.*, vol. 34(Database issue), pp. D16-20, 2006.

[13] M. Safran, "Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE," *Nucleic Acids Res*., vol. 31, no. 1, pp. 142-6 , 2003.

[14] iHOP: http://www.ihop-net.org/UniPub/iHOP/

[15] M.A. Harris, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res*., vol. 32(Database issue), pp. D258-61, 2004.

[16] L. Hermida, "MIMAS: an innovative tool for network-based high density oligonucleotide microarray data management and annotation," *BMC Bioinform*., vol 7, p. 190, 2006.

[17] G. Joshi-Tope, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Res*., 33(Database issue), p. D428-32, 2005.

[18] L. Lo Conte, "SCOP: a structural classification of proteins database," *Nucleic Acids Res*., vol. 28, no. 1, p. 257-9, 2000.

[19] R.D, Finn "Pfam: clans, web tools and services," *Nucleic Acids Res*., vol. 34(Database issue), pp. D247-51, 2006.

[20] J.L. Sussman "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules," *Acta Crystallogr. D Biol. Crystallogr*., vol. 54(Pt 6 Pt 1), pp. 1078-84, 1998.

[21] A. Bairoch, and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", *Nucleic Acids Res*., vol. 28, no. 1, pp. 45-8, 2000.

[22] G.D. Bader, D. Betel, and C.W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248-50, 2003.

[23] I. Donaldson "PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine," *BMC Bioinform*., vol. 4, p. 11, 2003.

[24] N. Bolshakova and F. Azuaje, "CVE: cluster validation for gene expression data," *Bioinformatics*, vol. 19, no. 18, pp. 2494-5, 2003.

[25] C.M Bouton, and J. Pevsner, "DRAGON View: information visualization for annotated microarray data," *Bioinformatics*, vol. 18, no. 2, pp. 323-4, 2002.

[26] MADRAS: http://www.madras.uwcm.ac.uk/

Received: February 17, 2009 Revised: April 24, 2009 Accepted: April 24, 2009