Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools

Daisuke Kihara[1,2,3,4], Yifeng David Yang[1] & Troy Hawkins[1]

[1]Department of Biological Sciences; [2]Department of Computer Science; [3] Markey Center for Structural Biology; [4] The Bindley Bioscience Center, College of Science, Purdue University, West Lafayette, IN, 47907, USA

Abstract: The immensely popular fields of cancer research and bioinformatics overlap in many different areas, e.g. large data repositories that allow for users to analyze data from many experiments (data handling, databases), pattern mining, microarray data analysis, and interpretation of proteomics data. There are many newly available resources in these areas that may be unfamiliar to most cancer researchers wanting to incorporate bioinformatics tools and analyses into their work, and also to bioinformaticians looking for real data to develop and test algorithms. This review reveals the interdependence of cancer research and bioinformatics, and highlight the most appropriate and useful resources available to cancer researchers. These include not only public databases, but general and specific bioinformatics tools which can be useful to the cancer researcher. The primary foci are function and structure prediction tools of protein genes. The result is a useful reference to cancer researchers and bioinformaticians studying cancer alike.

Keywords: microarray data management software, microarray data repository, gene function prediction, protein structure prediction, bioinformatics tools, protein-protein interaction database

Introduction

Since its birth in the 1980s, bioinformatics has been rapidly growing, keeping pace with the expansion of genome sequence data. Recent technological development of large-scale gene expression analysis using DNA microarrays and proteomics experiments has further boosted the importance of bioinformatics methods. The integration of wet experiments and the use of bioinformatics analyses have become an indispensable part of the biological and clinical research of this century.

The area of cancer research is not an exception. A typical scenario of cancer research using bioinformatics tools is analysis of global profiles of gene expression in cancer (Hedenfalk et al 2002;Dressman et al 2003; Subramanian et al 2004;Glanzer and Eberwine 2004). Gene expression patterns of cancer cells are compared with those of normal cells or those of other subtypes of the cancer, and genes over/under-expressed in the cancer tissue are identified and clustered (identifying cancer signatures). Additional clinical questions include identifying signatures of metastasis (Weigelt et al 2005;Jones et al 2005) and prediction of clinical outcome (Chen et al 2005;Eschrich et al 2005). Then biological function of the genes of such signatures is also of biological and clinical interest, because they represent selected candidate genes for further biochemical investigation and for the development of targeted therapies, such as siRNA interference. Comparisons of findings across studies are very important.

Our review is organized to provide a sampling of the studies conducted to date, and to review the potential biological and clinical significance of the genes found in such signatures, hopefully to promote further follow-up development of novel routes to prevention and treatment. This review is organized as follows. First, we briefly list software for organizing microarray data and retrieving annotation information for genes from public databases. Next, we highlight several microarray data repositories. Then we review tools for function prediction of genes. In the subsequent section, protein structure prediction methods are reviewed. This is because a predicted tertiary and secondary structure can often give useful information for the design of biochemical experiments on a protein. Sometimes function of genes can be inferred from the predicted structure, too. Next, we review databases of protein-protein interaction. Information about interacting partners of a given gene can

Correspondence: Daisuke Kihara: e-mail dkihara@purdue.edu

provide direct insight of the biochemical mechanism of a particular function of a cell and can also be a clue to guess about the function of that gene. This review is not intended to be a comprehensive survey of the field, but rather give a quick practical guide for recent developments of bioinformatics tools and databases useful for cancer research. Therefore in the choice of the introduced resources, preference is given to those that are non-commercial and well maintained. The bioinformatics tools and databases including those introduced in this article are available from our web site, http://dragon.bio.purdue.edu/bioinfolinks.

Previous Reviews

Rhodes *et al* proposed a statistical model for performing meta-analysis of gene expression data across independent studies, and applied it to expression profiles of prostate cancer (Rhodes et al 2002). They identified the function of significantly differentially expressed genes by PubMed literature searches (Wheeler et al 2002) and a KEGG pathway query (Kanehisa et al 2004). In the study of expression profile analysis of colorectal cancer by Yeh *et al,* functional characterization of up- and down-regulated genes was done using software to visualize expression patterns and function information of a set of genes was retrieved from public databases (Yeh et al 2005). Bono and Okazaki reviewed methods of function characterization of differently expressed genes using KEGG pathway mapping tools (Bono and Okazaki 2005). Statistical analysis of characteristic patterns of gene expression are practically very powerful in distinguishing cancer from normal tissue and distinguishing between subtypes of the cancer (Sorlie et al 2003). However, functional characterization of differently expressed genes can certainly give biological insight to the mechanism of the cancer. A recent excellent review by Rhodes and Chinnaiyan discusses the use of external functional information for interpreting and summarizing large cancer signatures (Rhodes and Chinnaiyan 2005). In their analysis, called the functional enrichment analysis, it is examined whether the difference of the fraction of genes which fall into a functional category from different samples is statistically significant or not.

In a functional analysis of a set of genes, it is desired that the employed method can assign accurate function to as large a number of genes as possible in the dataset. However, conventional homology search algorithms, such as BLAST (Altschul et al 1990) or FASTA (Pearson and Lipman 1988), can typically cover only 50% or less of the genes in a genome. Therefore it happens frequently that almost no functional clues are given to genes in a cluster of interest, which makes it extremely difficult to speculate about biological explanations to why the observed difference of gene expression profiles occurs. Note here that these homology search algorithms are also employed as a major computational procedure in public databases, such as KEGG and UniProt (Bairoch et al 2005), so that refereeing these databases does not necessarily solve the problem. One of the primary foci of this manuscript is to introduce and review bioinformatics tools for gene function and structure prediction, which aim to supplement functional assignment by the conventional homology search methods. Another focus is to introduce recent advanced protein structure prediction methods that will be useful for designing biochemical experiments of selected genes.

Microarray Data Management and Analysis Software

Microarray studies of gene expression usually analyze hundreds to tens of thousands of genes. Typical questions to be asked involve the statistical significance of an observed differential expression pattern between samples, or the function of a set of genes with a different expression pattern. GoMiner, listed at the top of Table 1, is software designed to facilitate function analysis of a set of genes in microarray studies (Zeeberg et al 2003). Functions of a set of input genes are mapped onto the Gene Ontology (GO) tree, which is a hierarchically controlled vocabulary of gene function (Harris et al 2004). Function is assigned to genes by referring to public databases, such as UniProt, species specific databases at The Institute for Genome Research (TIGR) (Lee et al 2005), and Mouse Genome

Table 1. Microarray Analysis Software Focusing on Function Clustering

Software	Location	Website
GoMiner	NCI, NIH	http://discover.nci.nih.gov/gominer/
GoSurfer	Harvard Univ.	http://www.biostat.harvard.edu/complab/gosurfer/
GenMAPP	UC San Francisco	http://www.genmapp.org
ArrayTrack	US Food and Drug Administration (FDA)	http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm

Table 2. Microarray Analysis Software Using Statistical Methodologies

Software/Link	Location	Website
caGEDA	Univ. of Pittsburgh	http://bioinformatics.upmc.edu/GE2/GEDA.html
SAM	Stanford Univ.	http://www-stat.stanford.edu/~tibs/SAM/
NUDGE	Univ. of Washington	http://www.bioconductor.org/
Microarray Software Comparison	The Chinese Univ. of Hong Kong	http://ihome.cuhk.edu.hk/~b400559/arraysoft_rpackages.html

Informatics (MGI) (Eppig et al 2005). Up-regulated and down-regulated genes are flagged on the GO tree, and the relative enrichment of up-/down-regulated genes in a GO category is statistically tested. There are also links to other public databases including LocusLink (Pruitt and Maglott 2001), BioCarta (www.biocarta.com) and PDB (Berman et al 2000). Its recent upgraded version, named High-Throughput GoMiner, handles multiple microarray data, a feature which is useful for a time-course study of gene expression (Zeeberg et al 2005). GoSurfer has similar functionality to GoMiner, including visualization of gene function on the GO tree and statistical tests to search for the GO terms that are enriched in the annotations of a subset of input genes (Zhong et al 2004).

GenMAPP is designed to view and analyze microarray data on biological pathways (Dahlquist et al 2002;Doniger et al 2003). Input genes can be mapped onto a biological pathway, which can be one of the standard pathways imported from KEGG or a user-customized pathway. Up-regulated and down-regulated genes in an experiment can be shown in a different color on the pathway. From each box of genes in a pathway, a user can view function annotation in public databases including UniProt, MGI, and GO. The numerical values of the expression level can be also retrieved. MAPPFinder, an associated program to GenMAPP, can also employ the function enrichment analysis on the GO tree.

ArrayTrack is comprehensive microarray data management and analysis software (Tong et al 2004). Multiple microarray data can be stored in an organized fashion and standard statistical tests can be employed in order to detect genes with a significantly different expression pattern among samples. Data normalization methods available in this software facilitate cross-chip comparison. It also provides a collection of functional information about genes, proteins and pathways imported from public databases. The functional enrichment test on the GO tree can be performed, and also several data plotting and visualization tools are available.

We limited the list in Table 1 to include only software easily downloadable to a local machine and free for academic users. There is also free web-based software, including DAVID (Dennis, Jr. et al 2003) and Onto-Express (Draghici et al 2003).

The above software is mainly aimed to cluster genes based on function and for mapping pathways. Table 2 lists software for gene clustering using statistical methodologies. caGEDA provides many

Table 3. Microarray Data Repositories

Repository	Location	Web Site
GEO	NCBI, NIH	http://www.ncbi.nlm.nih.gov/projects/geo/
ArrayExpress	EMBL-EBI	http://www.ebi.ac.uk/arrayexpress/
CIBEX	Nat. Inst. Genetics, Japan	http://cibex.nig.ac.jp
Standard Microarray Database (SMD)	Stanford Univ.	http://smd.stanford.edu/
The Gene Expression Database (GXD)	The Jackson Lab.	http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml
Oncomine	Univ. of Michigan	http://www.oncomine.org

alternative statistical tools for each step in microarray data analysis (preprocessing, feature selection, and patient prediction model development) (Patel and Lyons-Weiler 2004). Users can easily perform comparative evaluation of different methods on their data sets.

Significance Analysis of Microarrays (SAM) (Tusher et al 2001) and NUDGE (Dean and Raftery 2005) use R, which is a language and environment for statistical computing and graphics (http://www.r-project.org/). The last website contains abundant links to statistical tools for gene expression analysis using R. A good summary of statistical testing for gene expression was given by Dudoit et al. (Dudoit et al 2003).

Microarray Data Repositories

In this section, we briefly review public microarray repositories (Table 3). These repositories are very useful to retrieve data to perform cross-sample studies, identifying robust gene expression patterns across different conditions or different (sub)types of cancer (Rhodes and Chinnaiyan 2005). Data in the databases can also be analyzed using associated online tools. The Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) holds the largest number of high-throughput gene expression data entries, which exceeds 54,000 at the time of writing of this manuscript (Barrett et al 2005). Data from non-array-based high-throughput experiments are also stored, including comparative genomic hybridization, serial analysis of gene expression (SAGE) and mass spectrometry peptide profiling. Individual “Sample” data are also organized into “Series”, which bring related Samples together with summary tables of the Series. Data mining and visualization tools, such as clustering methods, are available for most of the stored data. ArrayExpress is another public repository for microarray data hosted by the European Bioinformatics Institute (EBI) (Parkinson et al 2005). This is useful not only for retrieving data; expression patterns can be visualized by a collection of tools called Expression Profiler (Kapushesky et al 2004). This web-based tool kit includes tools for data preprocessing, clustering, visualization and comparison between multiple samples. CIBEX is another public database, together with GEO and ArrayExpress, recommended by the Microarray Gene Expression Data (MGED) society for storing expression data related to publications (Ikeo et al 2003). In addition to the three repositories, three additional large databases are listed in Table 3. SMD also provides database software developed originally for the authors’ own use. GXD is specific for the expression profiles of transcripts and proteins in different mouse strains and mutants (Hill et al 2004). Oncomine is specific for gene expression in cancer (Rhodes et al 2004).

Protein Function Prediction Tools

Probably some of the most frequently used bioinformatics tools in cancer research are gene function prediction methods. As we have seen above, most of the microarray data management software import gene function from public databases, which typically hold function information

Table 4. Protein Function Prediction Tools

Software	Type	Location	Web Site
BLAST	Homology search	NCBI, NIH	http://www.ncbi.nlm.nih.gov/BLAST select protein-protein BLAST
FASTA	Homology search	Virginia Univ. Kyoto Univ.	http://fasta.bioch.virginia.edu http://fasta.genome.jp/
PSI-BLAST	Homology search	NCBI, NIH	http://www.ncbi.nlm.nih.gov/BLAST select "PSI- and PHI-BLAST"
Pfam	Protein family identification	Washington Univ	http://pfam.wustl.edu
SMART	Conserved Motif search	EMBL	http://smart.embl-heidelberg.de
PROSITE	Functional Motif search	Swiss Inst. Bioinformatics	http://us.expasy.org/prosite http://motif.genome.ad.jp
ELM	Functional motif search in eukaryotes	The ELM Consortium	http://elm.eu.org
STRING	Function prediction by comparative genomics	EMBL	http://string.embl.de
PSORT	Subcellular localization prediction	Human Genome Center, Tokyo Univ.	http://www.psort.org
PFP	Function prediction by mining PSI-BLAST result	Purdue Univ.	http://dragon.bio.purdue.edu/pfp

of only up to half of the genes in a genome. In order to perform the functional enrichment analysis on microarray data, it is crucial that genes in a cluster of interest have annotated function. Here we introduce several interesting gene function prediction methods developed in recent years. These tools are aimed to give functional clue to genes beyond a conventional BLAST search. Function can be predicted from gene (amino acid) sequence, the tertiary structure, interacting partners, or of course, expression patterns of genes (Watson et al 2005). The focus of this section is sequence-based methods, because sequence information is usually available for all of the genes in a microarray analysis.

In Table 4, first, three homology search methods are listed. Although less distributed, FASTA performs better or at least comparable to BLAST (Brenner et al 1998). The site at Virginia University will provide also the local copy of the program. The database search results of course depend on the sequence database to be searched. If a recent version of the sequence database is not available at the Virginia site, it would be better to try the KEGG site at Kyoto University. PSI-BLAST (Altschul et al 1997) is a variant of BLAST. It performs an iterative search of a database using information of retrieved sequences from former rounds; hence generally it has better sensitivity than BLAST and FASTA. But at the same time, caution should be used in examining PSI-BLAST search output, because spurious hits can easily contaminate the results. We reemphasize here that function annotation in public databases is mainly derived by these homology search methods, thus running these methods in a standard fashion may not yield additional useful annotation. Therefore, these analyses may be performed when users want to try a different parameter set for a more aggressive search or a different database to be searched.

Pfam (Bateman et al 2002) is a database of protein families described by Hidden Markov models (HMM), which are statistical representations of multiple sequence alignments (Eddy 1996). Since a query sequence is searched against HMMs that have more information than single sequences, an increased sensitivity in the search is expected. From the Pfam website, a database search can be performed. Also the database itself and software for searching and creating a HMM database can be downloaded.

The next three resources, SMART (Letunic et al 2004), PROSITE (Hulo et al 2004) and ELM (Puntervoll et al 2003) are sequence motif databases with different features. SMART stores conserved regions in multiple sequence alignments of protein families, which can be used as signatures of each gene family. On the other hand, sequence motifs in PROSITE are primarily biologically significant sites described in literature, which include functional sites and sites which are subject to chemical modifications. ELM is a database for functional sites of eukaryotes.

STRING is an interactive database of known and predicted functional associations between genes (von Mering et al 2003). The interesting feature of STRING is that the function of a query sequence is predicted by comparative genomics methods, which are made possible by the growing number of complete genomes available. For example, if a query gene locates next to a gene of known function in several genomes of moderate evolutionary distance from each other, it would indicate that the query gene is involved in the same pathway or function as the adjacent gene. Genes that have the same phylogenetic profile (*i.e.* tree) and genes with the same pattern of co-occurrence and co-absence in genomes may also indicate that they are functionally linked. STRING also uses co-expression patterns in microarray analyses, and previous knowledge mined from PubMed literature abstracts. Users can perform function prediction on the web site, and also the functional association data in STRING are freely available.

PSORT is a server for predicting subcellular localization of genes (Nakai and Horton 1999). Basically, sequence features (signal sequences etc.) in a query sequence are detected and classified to known localization using a machine learning technique. The series of PSORT server families and links to the other servers of the same sort listed in the web site would be also useful.

The PFP (Protein Function Prediction) server was recently developed by our group (Hawkins and Kihara 2005a;Hawkins and Kihara 2005b). Unlike the conventional way to use PSI-BLAST, PFP mines more functional information from sequence hits with generally-thought insignificant hits by applying function association rules learned from genes of known function in public databases. PFP performed the best at the automatic function prediction competition held at the 13th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) in June, 2005 (http://ffas.burnham.org/AFP).

Among the servers listed here, BLAST, FASTA and Pfam are the most reliable but may not provide additional functional information to annotation already stored in public databases. The other methods often outperform the three methods above and have a higher coverage, but should be used carefully because they also have a relatively high rate of spurious hits. A reasonable way to reduce false positives is to use different methods and compare the results to see if the prediction is consistent among the used methods.

Protein Structure Prediction Tools

When candidates of genes are selected for experimental work-up by a microarray analysis, bioinformatics protein structure prediction tools are often very useful for designing biochemical experiments. For example, predicted secondary structure of a gene is a good clue to guess the domain structure of a gene, which is important to design limited proteolysis experiments in order to identify the functional region of the gene. The prediction accuracy of current secondary structure prediction algorithms is about 75% (Rost 2001;Kihara 2005), which would be high enough for routine use. Five secondary structure prediction tools are listed in Table 5. All of them use a machine learning technique to recognize known sequence patterns for α-helices and β-strands. PSI-PRED (Jones 1999), PORTER (Pollastri and McLysaght 2005), SABLE (Adamczak et al 2005) and PredictProtein (Rost and Sander 1994) use artificial neural networks, and SAM-T02 (Karplus et al 2003) uses the HMM. SABLE and PORTER claim the best accuracy in this field to date (78.4% and 79%, respectively). A local copy of the program is available for PSIPRED and SAM-T02. Although the accuracy of PredictProtein is relatively lower among those listed here, the server predicts not only the secondary structure but also other structural

Table 5. Protein Structure Prediction Tools

Software	Type	Location	Web Site
PSIPRED	2ndary structure	Univ. College London	http://bioinf.cs.ucl.ac.uk/psipred/
PORTER	2ndary structure	Univ. College Dublin	http://distill.ucd.ie/porter/
SAM-T02	2ndary structure	UC Santa Cruz	http://www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html
SABLE	2ndary str., solvent accesibility	Cincinnati Children's Hospital Med. Center	http://sable.cchmc.org/
PredictProtein	2ndary structure and others	Columbia Univ.	http://cubic.bioc.columbia.edu/predictprotein/
COILS	Coiled-coil region	EMBnet, Switzerland	http://www.ch.embnet.org/software/COILS_form.html
GlobPlot	Disordered region	EMBL	http://globplot.embl.de/
PONDR	Disordered region	Indiana Univ.	http://www.pondr.com/
TMHMM	Transmembrane domain	Technical Univ. of Denmark	http://www.cbs.dtu.dk/services/TMHMM-2.0/
HMMTOP	Transmembrane domain	Hungarian Academy of Sciences	http://www.enzim.hu/hmmtop/
SWISS-MODEL	3D structure; (Homology modeling)	Swiss Inst. of Bioinformatics	http://swissmodel.expasy.org
HHPred	3D str.; (Homology modeling)	Max-Planck Inst.	http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred
MODELLER	3D str.; (Homology modeling)	UC San Francisco	http://salilab.org/modeller/
FUGUE	3D str., threading	Univ. of Cambridge	http://www-cryst.bioc.cam.ac.uk/~fugue/
Phyre	3D str., threading	Imperial College London	http://www.sbg.bio.ic.ac.uk/~phyre/
SPARKS	3D str., threading	SUNY Buffalo	http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sparks2.html
Robetta	3D str; ab initio	Univ. Washington	http://robetta.bakerlab.org/

information, including disordered regions, coiled-coil regions, per residue solvent accessibility, and motifs in a query sequence. Thus it can be used as a convenient one-stop server for analyzing a protein sequence.

COILS predicts coiled-coil regions of a protein by recognizing unique patterns of periodic occurrence of hydrophobic residues in a sequence (Lupas 1996). Coiled-coil regions have been drawing attention recently because these regions are often binding sites to other proteins. GlobPlot (Linding et al 2003) and PONDR (Romero et al 2001) are prediction tools for intrinsic disordered regions of proteins, which do not have stable secondary structures in their native conformation. Importance of disordered regions has also been recognized recently because many functionally important sites, *e.g.* those responsible for binding to other proteins or ligand molecules, are outside of the stable globular domains and thus intrinsically disordered. Programs for local use are available for all of three tools.

TMHMM (Sonnhammer et al 1998) and HMMTOP (Tusnady and Simon 2001) are transmembrane (TM) domain prediction tools which use HMM. TM domain prediction is one of the most successful structure predictions in bioinformatics (Kihara et al 1998). HMMTOP reports that 98% of the domains and 85% of topology of TM proteins in their benchmark set are correctly predicted. Both tools are web-based servers, and HMMTOP also provides a local copy of the program.

The bottom half of Table 5 lists protein tertiary structure prediction tools. Methodology of protein tertiary structure prediction has made dramatic improvements in the past decade, and the accuracy of some methods has reached a practical level. A recent review concisely describes the current status of this field (Schueler-Furman et al 2005). Structural prediction methods are roughly classified into three categories, namely homology modeling, threading (fold recognition), and "*ab initio*" or "*de novo*" folding (Jones 2000;Baker and Sali 2001;Forster 2002). Homology methods use an experimentally determined tertiary structure of a highly homologous protein to a query protein sequence as a template for modeling. Therefore, when an appropriate template structure is available in PDB, a very accurate model in an atomic detailed level can be built. SWISS-MODEL (Schwede et al 2003) and HHPred (Soding et al 2005) are web-based servers for homology modeling. The HHPred software is also available for download. MODELLER (Sali and Blundell 1993) is the most widely distributed and one of the earliest examples of this type of software. Both MODELLER and SWISS-MODEL have a database of homology models generated by the software.

The next three tools, FUGUE (Shi et al 2001), Phyre (Bates et al 2001) and SPARKS (Zhou and Zhou 2004) fall into the category of threading (Skolnick and Kihara 2001;Skolnick et al 2004). Threading algorithms seek a template protein in a database that structurally fits well to a query sequence. Unlike homology modeling, an apparent sequence similarity between a query sequence and a template protein is not a necessary condition. Threading methods have improved significantly in the past years, and can detect remotely related protein structures very well from a database, if any exist. A statistical score, the Z-score, shows the significance of the match between a query sequence and a template structure. Users should pay attention to the Z-score of retrieved models, and should only use models with a significant Z-score, as recommended by the server. When the Z-score is low, it may simply mean that there are no structures that fit well to a query, or the alignment between the query and the template is not very reliable.

The last server, Robetta (Kim et al 2004), is an *ab initio* method, which assembles a model from pieces of structural fragments retrieved from a database. Although algorithms of this category have also made a dramatic improvement (Kihara et al 2001;Skolnick et al 2003), it is still early to use *ab initio* methods routinely. When using *ab initio* methods, generated models should be checked carefully to see if they are reasonable in the biological sense based on background knowledge of the protein.

Protein Protein Interaction Databases

The last group of resources we describe here are databases of protein-protein interactions (PPI) in model organisms (Table 6). In the past five years, an increasing number of large-scale experiments for revealing PPI in various organisms have been conducted, and most of the data are available at databases on the internet (Auerbach et al 2002). PPI of a gene is very important information to speculate the context of the gene's role; for example, the pathway or subcellular localization of a gene. BIND (Alfarano et al 2005) is currently the largest PPI data repository, and contains over 200,000 interactions from more than 1,500 unique organisms. It also provides tools for visualization and data retrieval. DIP (Salwinski et al 2004) is one of the earliest databases of this kind and stores over 18,000 interactions. MIPS stores mammalian PPI data collected from literature with *Mus musculus* as the reference organism (Pagel et al 2005). HPRD is a unique database of information of human proteins in health and disease, including PPIs, posttranslational modifications, disease associations, tissue expression etc., extracted manually from literature (Peri et al 2003). GRID stores PPI data of the fruit fry, yeast, and worm. Note that data is downloadable from all the databases above.

IntAct (Hermjakob et al 2004) and Ospray (Breitkreutz et al 2003) are an open source database and toolkit for storage, visualization and analysis of PPI data. These packages would be useful to integrate in a microarray data management system to link to PPI data.

Table 6. Protein Protein Interaction Databases and Database Tools

DB/Software	Type	Location	Web Site
BIND	PPI, pathway	Mt. Sinai Hospital, Canada	http://bind.ca/
DIP	PPI	UC Los Angeles	http://dip.doe-mbi.ucla.edu/
MIPS	Mammalian PPIs	Munich Information Center for Protein Sequences	http://mips.gsf.de/proj/ppi/
HPRD	Human protein references	Johns Hopkins Univ.	http://www.hprd.org/
GRID	genetic and physical interactions of yeast, fry, worm	Mt. Sinai Hospital, Canada	http://biodata.mshri.on.ca/grid/
IntAct	open source db systems & tools for PPI data	EBI	http://www.ebi.ac.uk/intact/
Ospray	PPI visualization tool	Mt. Sinai Hospital, Canada	http://biodata.mshri.on.ca/osprey/

Summary

In the last decade, many new techniques have appeared in experimental biology that have had a tremendous impact on directions and styles of cancer research. And the same thing is true for bioinformatics databases and tools; indeed development and improvement of bioinformatics resources might be even more rapid than experimental techniques. A key to effectively handling large-scale experimental data is to use appropriate and reliable bioinformatics tools to organize and analyze that data.

The bioinformatics tools reviewed here were chosen with a scenario that gene-expression patterns of a certain type of cancer are investigated, functional enrichment analyses are performed to identify the signature of the cancer type, and further biochemical experiments are designed for a handful of selected genes with help of protein structure prediction methods (Fig. 1). If the function of genes cannot be retrieved from public databases, homology search methods are the first choice for prediction. If there are still no significant hits in the search, the other sequence based methods, including STRING, PFP, and PSORT can be used. At the same time, motif searches may also be able to provide functional clues for the genes. PPI data will provide the context of the genes' function, and can be used to cluster genes in terms of their interaction patterns. To design biochemical experiments to determine functional/interaction domains of a given gene, it is helpful to predict the secondary structure of the gene. Motif search and homology search methods can also provide conserved functional regions of the gene. Predicted tertiary structure is useful for designing site-directed mutagenesis experiments.

Other types of bioinformatics tools not included in this article but useful for cancer research would be transcription binding site prediction tools (or DNA motif finding algorithms). For DNA motif finding tools, please refer to recent studies on the benchmarking of several programs (Tompa et al 2005;Hu et al 2005). All of the introduced resources can be used on-line from their websites, but some are also downloadable for use on local machines. The resources for which local copies are available are explicitly mentioned in the text because they can be integrated into a microarray data management system to make the system more comprehensive. It is no doubt that bioinformatics are going to play a more important role in cancer research in this new century, and this article is intended to be an aid for selecting useful tools for researchers in this field.

Acknowledgement

D.K. acknowledges the support from the National Institute of General Medical Sciences of the National Institutes of Health (grant number R01 GM-075004). The authors are grateful to Stan Luban for proofreading the manuscript.

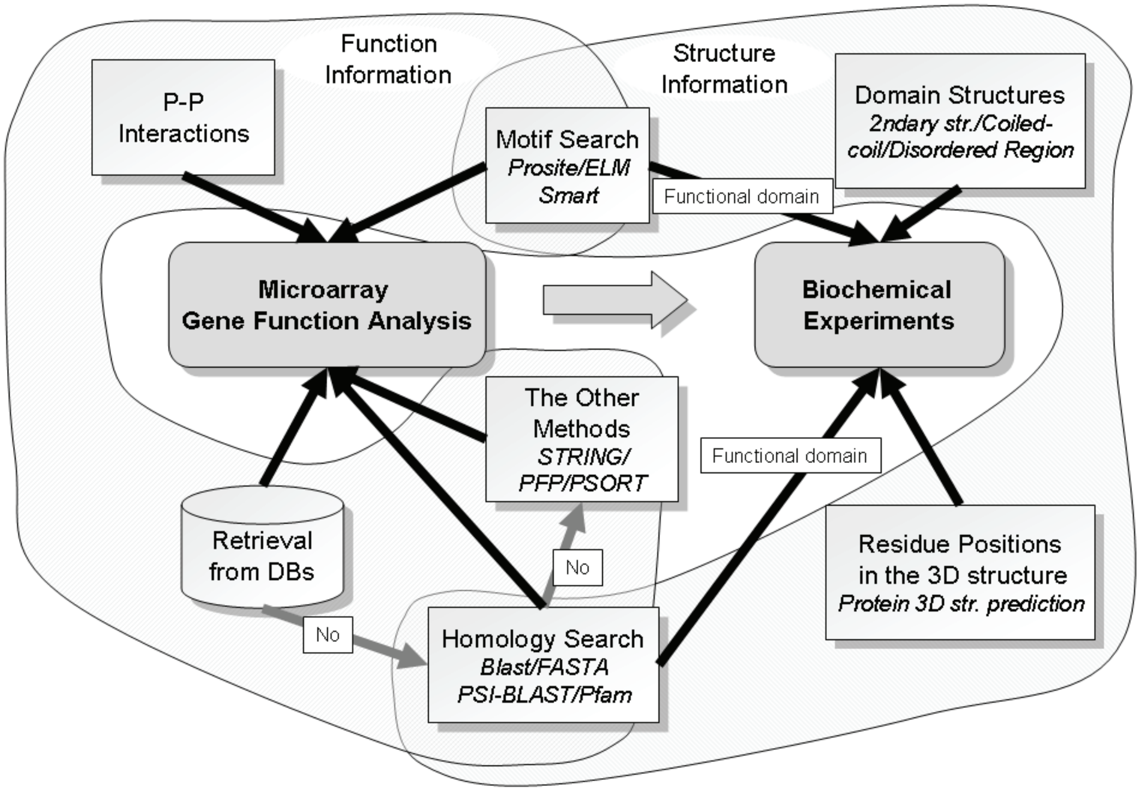

Figure 1. Bioinformatics tools for protein function and structure prediction. Gene function assignment is used in a microarray analysis. Structure information such as the domain structure or the tertiary structure of a certain gene is useful for designing biochemical experiments of the gene. See the text for the details.

References

Adamczak R, Porollo A, and Meller J 2005. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins, 59:467-475.

Alfarano C et al. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res, 33:D418-D424.

Altschul SF et al. 1990. Basic local alignment search tool. J Mol Biol, 215:403-10.

Altschul SF et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25:3389-402.

Auerbach D et al. 2002. The post-genomic era of interactive proteomics: facts and perspectives. Proteomics., 2:611-623.

Bairoch A et al. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res, 33:D154-D159.

Baker D and Sali A 2001. Protein structure prediction and structural genomics. Science, 294:93-96.

Barrett T et al. 2005. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res, 33:D562-D566.

Bateman A et al. 2002. The Pfam protein families database. Nucleic Acids Res, 30:276-80.

Bates PA et al. 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins, 45 Suppl 5:39-46.

Berman HM et al. 2000. The Protein Data Bank. Nucleic Acids Res, 28:235-42.

Bono H and Okazaki Y 2005. The study of metabolic pathways in tumors based on the transcriptome. Semin.Cancer Biol., 15:290-299.

Breitkreutz BJ, Stark C, and Tyers M 2003. Osprey: a network visualization system. Genome Biol., 4:R22.

Brenner SE, Chothia C, and Hubbard TJ 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc.Natl.Acad.Sci.U.S.A, 95:6073-6078.

Chen CN et al. 2005. Gene expression profile predicts patient survival of gastric cancer after surgical resection. J.Clin.Oncol., 23:7286-7295.

Dahlquist KD et al. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat.Genet., 31:19-20.

Dean N and Raftery AE 2005. Normal uniform mixture differential gene expression detection for cDNA microarrays. BMC.Bioinformatics, 6:173.

Dennis G, Jr. et al. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol., 4:P3.

Doniger SW et al. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol., 4:R7.

Draghici S et al. 2003. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Res, 31:3775-3781.

Dressman MA et al. 2003. Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. Cancer Res, 63:2194-2199.

Dudoit S, Shaffer J, and Boldrick J 2003. Multiple hypothesis testing in microarray experiments. Stat.Sci., 18:71-103.

Eddy SR 1996. Hidden Markov models. Curr Opin Struct Biol, 6:361-5.

Eppig JT et al. 2005. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res, 33:D471-D475.

Eschrich S et al. 2005. Molecular staging for survival prediction of colorectal cancer patients. J.Clin.Oncol., 23:3526-3535.

Forster MJ 2002. Molecular modelling in structural biology. Micron., 33:365-384.

Glanzer JG and Eberwine JH 2004. Expression profiling of small cellular samples in cancer: less is more. Br.J.Cancer, 90:1111-1114.

Harris MA et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, 32:D258-D261.

Hawkins T and Kihara D 2005a. PFP:Automatic annotation of protein function by relative GO association in multiple functional contexts. The 13th Annual International Conference on Intelligent Systems for Molecular Biology, 117.

Hawkins T and Kihara D 2005b. The use of context-based functional association in automated protein function prediction methods. The 13th Annual International Conference on Intelligent Systems for Molecular Biology, Automatic Function Prediction - Special Interest Group, 16-17.

Hedenfalk IA et al. 2002. Gene expression in inherited breast cancer. Adv.Cancer Res, 84:1-34.

Hermjakob H et al. 2004. IntAct: an open source molecular interaction database. Nucleic Acids Res, 32:D452-D455.

Hill DP et al. 2004. The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Res, 32:D568-D571.

Hu J, Li B, and Kihara D 2005. Limitations and potentials of current motif discovery algorithms. Nucleic Acid Res., 33:4899-4913.

Hulo N et al. 2004. Recent improvements to the PROSITE database. Nucleic Acids Res, 32:D134-D137.

Ikeo K et al. 2003. CIBEX: center for information biology gene expression database. C.R.Biol., 326:1079-1082.

Jones DT 2000. Protein structure prediction in the postgenomic era. Curr Opin Struct Biol, 10:371-9.

Jones DT 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 292:195-202.

Jones J et al. 2005. Gene signatures of progression and metastasis in renal cell cancer. Clin.Cancer Res, 11:5730-5739.

Kanehisa M et al. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res., 32 Database issue:D277-D280.

Kapushesky M et al. 2004. Expression Profiler: next generation--an online platform for analysis of microarray data. Nucleic Acids Res, 32:W465-W470.

Karplus K et al. 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins, 53 Suppl 6:491-496.

Kihara D 2005. The effect of long-range interactions on the secondary structure formation of proteins. Protein Sci., 14:1955-1963.

Kihara D et al. 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A, 98:10125-30.

Kihara D, Shimizu T, and Kanehisa M 1998. Prediction of membrane proteins based on classification of transmembrane segments. Protein Eng, 11:961-70.

Kim DE, Chivian D, and Baker D 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res, 32:W526-W531.

Lee Y et al. 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res, 33:D71-D74.

Letunic I et al. 2004. SMART 4.0: towards genomic data integration. Nucleic Acids Res, 32:D142-D144.

Linding R et al. 2003. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res, 31:3701-3708.

Lupas A 1996. Coiled coils: new structures and new functions. Trends Biochem Sci, 21:375-82.

Nakai K and Horton P 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem.Sci., 24:34-36.

Pagel P et al. 2005. The MIPS mammalian protein-protein interaction database. Bioinformatics, 21:832-834.

Parkinson H et al. 2005. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res, 33:D553-D555.

Patel S and Lyons-Weiler J 2004. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. Appl.Bioinformatics, 3:49-62.

Pearson WR and Lipman DJ 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A, 85:2444-8.

Peri S et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res, 13:2363-2371.

Pollastri G and McLysaght A 2005. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics, 21:1719-1720.

Pruitt KD and Maglott DR 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res, 29:137-140.

Puntervoll P et al. 2003. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res, 31:3625-3630.

Rhodes DR et al. 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res, 62:4427-4433.

Rhodes DR and Chinnaiyan AM 2005. Integrative analysis of the cancer transcriptome. Nat.Genet., 37 Suppl:S31-S37.

Rhodes DR et al. 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia., 6:1-6.

Romero P et al. 2001. Sequence complexity of disordered protein. Proteins, 42:38-48.

Rost B 2001. Review: protein secondary structure prediction continues to rise. J Struct Biol, 134:204-18.

Rost B and Sander C 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins, 19:55-72.

Sali A and Blundell TL 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol, 234:779-815.

Salwinski L et al. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res, 32:D449-D451.

Schueler-Furman O et al. 2005. Progress in modeling of protein structures and interactions. Science, 310:638-642.

Schwede T et al. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res, 31:3381-3385.

Shi J, Blundell TL, and Mizuguchi K 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J.Mol Biol., 310:243-257.

Skolnick J and Kihara D 2001. Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. Proteins, 42:319-31.

Skolnick J, Kihara D, and Zhang Y 2004. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins, 56:502-518.

Skolnick J et al. 2003. TOUCHSTONE: A unified approach to protein structure prediction. Proteins, 53:469-479.

Soding J, Biegert A, and Lupas AN 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res, 33:W244-W248.

Sonnhammer EL, von HG, and Krogh A 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc.Int.Conf.Intell.Syst.Mol.Biol., 6:175-182.

Sorlie T et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc.Natl.Acad.Sci.U.S.A, 100:8418-8423.

Subramanian S et al. 2004. Gastrointestinal stromal tumors (GISTs) with KIT and PDGFRA mutations have distinct gene expression profiles. Oncogene, 23:7780-7790.

Tompa M et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat.Biotechnol., 23:137-144.

Tong W et al. 2004. Development of public toxicogenomics software for microarray data management and analysis. Mutat.Res, 549:241-253.

Tusher VG, Tibshirani R, and Chu G 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc.Natl.Acad.Sci.U.S.A, 98:5116-5121.

Tusnady GE and Simon I 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics, 17:849-850.

von Mering C et al. 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res., 31:258-261.

Watson JD, Laskowski RA, and Thornton JM 2005. Predicting protein function from sequence and structural data. Curr.Opin.Struct.Biol., 15:275-284.

Weigelt B et al. 2005. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res, 65:9155-9158.

Wheeler DL et al. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res, 30:13-16.

Yeh CS et al. 2005. Fatty acid metabolism pathway play an important role in carcinogenesis of human colorectal cancers by Microarray-Bioinformatics analysis. Cancer Lett., EPub.

Zeeberg BR et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol., 4:R28.

Zeeberg BR et al. 2005. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC.Bioinformatics, 6:168-

Zhong S et al. 2004. GoSurfer : A Graphical Interactive Tool For Comparative Analysis of Large Gene Sets in Gene Ontologytrade mark Space. Appl.Bioinformatics, 3:261-264.

Zhou H and Zhou Y 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins, 55:1005-1013.