24

Pathway Mapping Tools for Analysis of High Content Data

Sean Ekins, Yuri Nikolsky, Andrej Bugrim, Eugene Kirillov, and Tatiana Nikolskaya

Summary

The complexity of human biology requires a systems approach that uses computational approaches to integrate different data types. Systems biology encompasses the complete biological system of metabolic and signaling pathways, which can be assessed by measuring global gene expression, protein content, metabolic profiles, and individual genetic, clinical, and phenotypic data. High content screening assays can also be used to generate systems biology knowledge. In this review, we will summarize the pathway databases and describe biological network tools used predominantly with this genomics, proteomics, and metabolomics data but which are equally as applicable for high content screening data analysis. We describe in detail the integrated data-mining tools applicable to building biological networks developed by GeneGo, namely, MetaCore™ and MetaDrug.

Key Words: Cervical cancer; database; genomics; high content screening; metabonomics; MetaCore; MetaDrug; networks; ontologies; pathways; proteomics; signature networks.

1. Introduction

The complex nature of human biology ultimately necessitates a systems approach to data analysis, integrating different data types *(1)* using computational methods, which in turn enable automated analysis *(2)*. Systems biology is widely acknowledged as the new paradigm for understanding the complex biological data sets derived from high-throughput technologies and the accumulated knowledge on human protein interactions *(3,4)*. Hence, systems biology can be defined as the integration of genetic, proteomic, transcriptomic, and metabonomic data using computational methods *(1)*. To understand the perturbing effect of a molecule or condition on the complete biological system encompassing metabolic and signaling pathways or networks and the effects on gene or protein expression, the collection of high-throughput data is required. This can include global gene expression, protein content, metabolic profiles for the same samples as well as individual genetic, clinical, and phenotypic data.

The increasing generation of biological data derived from these high-throughput techniques necessitates the use of computational technologies to store, analyze, interpret, and learn from this information *(5)*. Such high-throughput data are important for drug discovery from target identification and validation through to clinical development. These data are generally poorly utilized because of the lack of available methods for interpretation of diseases and biological function. Method development to visualize complex expression data has also recently expanded beyond the widely used clustering methods *(6)*. With the outcome of microarray analysis being dependent on the widely used statistical procedures applied to derive those genes that are significantly differentially expressed *(7)*, newer approaches that do not necessarily require data clustering might be an advantage.

From: *Methods in Molecular Biology, vol. 356:*
High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery
Edited by: D. L. Taylor, J. R. Haskins, and K. Giuliano © Humana Press, Inc., Totowa, NJ

In parallel to high-throughput screening methods we have seen the development of high content screening (HCS). This latter approach represents a method for understanding the roles of genes, proteins, and other small molecules and ions under different cellular conditions. The development of this approach has required advances in the methods for data generation encompassing cell culture, fluorescence microscopy, molecule labeling, image generation, and storage. This approach is high content as specific biological processes can be determined in individual cells as well as across the population of cells. Such a data intensive method also presents challenges for the effective archiving and analysis of images to ultimately generate information that can impact drug discovery *(8–11)*. There are many examples, in which HCS has been used to understand the effect of small molecules like siRNA on whole cells *(12–17)*. HCS data can be linked to unique gene or protein identifiers (e.g., gene-specific siRNAs) and, therefore, can be analyzed using the same methods as used with high-throughput genomics data derived from microarrays. HCS assays are however different from such high-throughput molecular experiments as HCS data represent a phenotypic end point, as opposed to the indirect nature of expression or metabolic data. Therefore, the real power of HCS assays is that, one can directly associate the stimulus with the final outcome in the cell, potentially bypassing the commonly used in vitro biology studies, such as recombinant enzyme or receptor-based assays. The major drawback however is in the complexity and deconvolution of HCS assay data. Multiplexed-HCS assays can be used to generate systems biology knowledge by providing responses for multiple cells to a drug *(18)*.

As we have described previously *(19)* there are at least nine levels of information flow in the cell between the gene sequence and folded active protein (**Fig. 1**), which cannot be directly reconstructed from HCS assays alone, but can be assessed with help of complimentary data and knowledge-based data-mining tools. Therefore, we should consider and analyze HCS data in the context of the underlying biology as much as possible. First, molecular high-throughput or high-content data are snapshots of cellular responses at different levels of information flow in the cell (**Fig. 1**) and the data compliment each other and can be cross referenced. Currently, microarrays are the major type of high-throughput data, because they are based on a readily standardized technique. Second, the HCS data should be considered in the context of functional categories, which we can define as pathways, cellular processes, and biological networks. In this review, we will summarize the tools used predominantly with genomics, proteomics, and metabolomics data, which are equally as applicable for HCS data analysis. In particular, we will focus on the most widely used pathway databases and describe biological networks, which are currently the most relevant, highest resolution research tools for systems biology. Our emphasis will be on the integrated data-mining tools developed by GeneGo (www.genego.com) that includes MetaCore™ and MetaDrug.

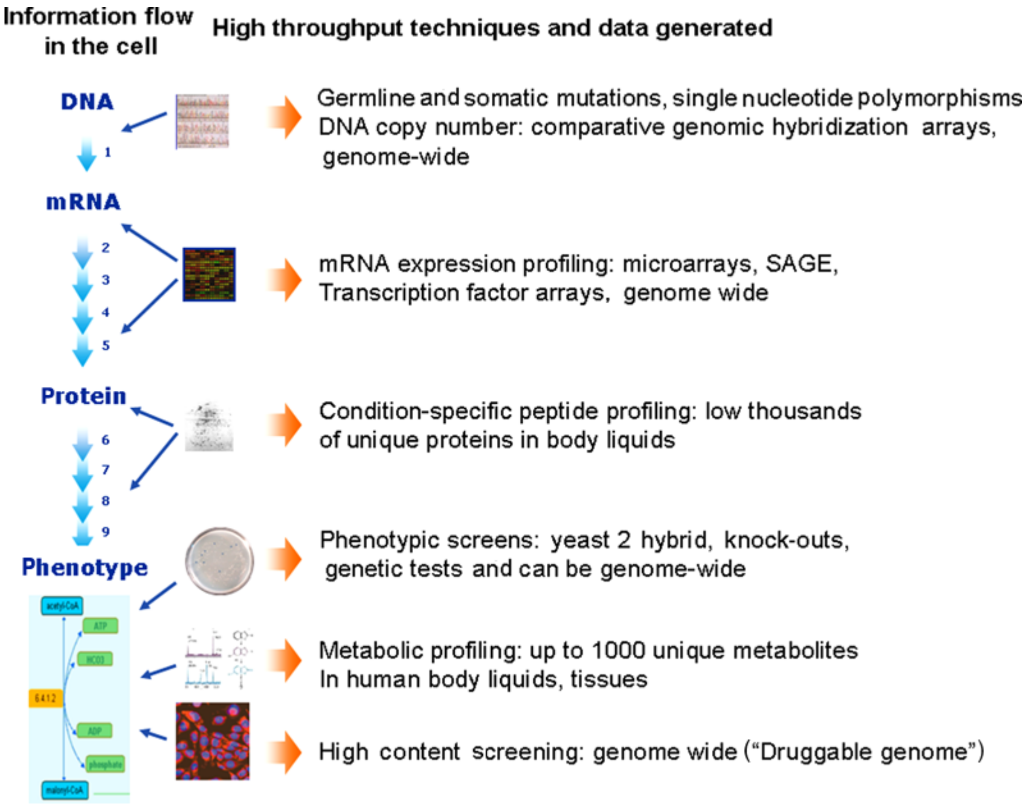
2. Pathways and Networks

2.1. Definition

It is important to define the terms pathways and networks at the outset. We consider the major tools for data analysis as pathways and networks. Pathways are consecutive reaction steps, which are either biochemical transformations or sequences of signaling events, such as signal transduction. Both are static as predefined by previous studies. Networks in contrast are dynamic, as they are built *de novo* out of building blocks from binary interactions and are specific for each data set. The process of data analysis therefore consists of narrowing down the list of potentially many thousands (if not more) data points to something more interpretable. This can be achieved by using statistical analysis p-values, different scoring methods for the intersections between categories, and calculation of the relevance of the result to the data set in question (using the relative saturation of pathways and networks with data).

2.2. Pathway Databases

Metabolism was the first functional level in human biology studied experimentally and therefore was the source for the first databases of biochemical reactions and pathways. These databases



Germline and somatic mutations, single nucleotide polymorphisms
DNA copy number: comparative genomic hybridization arrays, genome-wide

mRNA expression profiling: microarrays, SAGE,
Transcription factor arrays, genome wide

Condition-specific peptide profiling: low thousands of unique proteins in body liquids

Phenotypic screens: yeast 2 hybrid, knock-outs, genetic tests and can be genome-wide

Metabolic profiling: up to 1000 unique metabolites
In human body liquids, tissues

High content screening: genome wide ("Druggable genome")

Fig. 1. This is the general schema for network analysis from different data types. Various high-throughput or high content data can be linked to the tables of human protein interactions. Nine levels of regulation of protein activity in a human cell can be summarized: (1) gene transcription, (2) mRNA processing and editing, (3) mRNA transport from nucleus, (4) mRNA stabilization, (5) protein translation, (6) protein transport, (7) folding and protein stabilization, (8) allosteric modulation, and (9) covalent modification. The types of data generated corresponding to these levels are also shown. (Please *see* the companion CD for the color version of this figure.)

include EMP/MPW, BRENDA, ERGO, and KEGG, which have been described previously ***(19)***. KEGG is probably the database most frequently referred to for metabolic pathways in multiple species. Unlike bacteria with more than 80% of genome-content encoding core metabolic genes, the majority of eukaryotic biology is in signaling: membrane receptors, signal transduction enzymes, transcription factors, and modulators. With the progress of eukaryotic biology in the last decade, signaling pathways have also been collected in multiple databases. Some of examples of these signaling databases are listed later.

Biocarta is a commercial collection of about 350 maps on human biology representing canonical, mostly signaling pathways. Biocarta does not support mapping of experimental data on the pathways.

Gene MicroArray Pathway Profiler is a database of gene ontology (GO) derived diagrams designed for viewing and analyzing gene lists from experimental data ***(5)***. A free tool, Gene MicroArray Pathway Profiler became popular among academics as the first pass functional analysis of microarray expression data.

Protein Lounge is a commercial package with about 600 mixed-species metabolic and signaling pathway maps. These pathways can only be browsed, as there are no tools for mapping experimental data.

MetaCore is a commercial package that contains more than 400 human signaling and metabolic pathway maps available for mapping gene expression, proteomics, metabolic, and HCS data. The data generated can be exported from individual maps and clusters of maps and analyzed further with networks. MetaCore will be described in considerably more detail later in this chapter (**Subheading 3.**).

iPath (Invitrogen; http://escience.invitrogen.com/ipath/) 225 detailed, interactive maps of interconnected biological signaling and metabolic pathways that are freely available. These maps have been created for Invitrogen by GeneGo and are linked to reagents commercially available from Invitrogen.

Databases of high-confidence human protein–protein interactions and associated tools for network generation, analysis, and mapping of experimental high-throughput data have also been developed *(**19,20**)*. There are multiple methods to elucidate protein–protein interactions. One approach is to screen the experimental literature, using text-mining algorithms, for cooccurrence and, therefore, associations between gene/protein symbols and names in the same text. Typically, natural language processing (NLP) and other text-mining algorithms are used for this automated mining of abstracts and titles from PubMed articles *(**21–23**)*. It is important to note that comparative studies have shown that up to one half of NLP associations do not correspond to experimentally verified protein interactions *(**24**)*, although more than 60% of interactions can be elucidated by automated text mining *(**25**)*. The reliability of NLP-derived associations can be enhanced further by the compilation of field-specific synonym dictionaries, using longer word strings for searching and full-text articles to query against, and statistical methods *(**26**)*. In a recent study, the NLP engine MedScan was used to extract hundreds of thousands of functional relations including 20,000 protein interaction facts between human proteins from full text articles with a precision of 91% for 361 randomly extracted protein interactions *(**27**)*.

The protein–protein interactions can also be derived from high-throughput experimentation. For example, the yeast 2-hybrid (Y2H) screen test identifies protein interactions in yeast cells *(**28**)*. A widely used wet lab technique, Y2H was scaled-up for global mapping of protein interactions in yeast *(**29**)*, fly *Drosophila melanogaster* (***30***), and worm *Caenorhabditis elegans* (***31***); however, Y2H-derived interactions are known for a high (more than 50%) level of false positives and false negative interactions *(**32**)*. In one study, over 70,000 interactions for 6231 human proteins were predicted assuming the interactions between these proteins' orthologs in yeast, worm, and fly *(**33**)*. However, the accuracy of predicted interactions remains questionable. Although it was assessed computationally based on the relative correspondence of interacting protein pairs to the GO processes, the interactions were not directly compared with any high-confidence experimental data set available *(**5,34**)*. The interactions can also be deduced from condition-specific cooccurrence of gene expression based on the assumption that interacting proteins must be expressed in parallel *(**35**)*, especially when encoded by the homologous genes *(**34**)*. However, the overall confidence in coexpression-derived interactions in yeast is about 50% (47% anticorrelation for novel interactions) *(**37,38**)*. Another method, coimmunoprecipitation (CoIP) consists of the affinity precipitation of protein complexes in mild conditions using antibodies to one of the complex's subunits, followed by mass-spectrometry or Western blot analysis. CoIP enables direct and quantitative detection of interactions between active proteins, and so it is a true proteomics method. CoIP was used in simultaneously published studies of the yeast interactome *(**39,40**)*. The other, less commonly used experimental and computational methods include protein arrays, fusion proteins, neighboring genes in operons (for prokaryotic proteins), paralogous verification method, colocalization, synthetic lethality screens, and phage display; each method with its own particular pros and cons *(**41,42**)*. The overall confidence in protein interactions, defined as the intersection between interacting pairs obtained using different methods, remains poor. In general, it is believed that only manually curated, physical protein interactions extracted from the original small-scale experimental literature can be used with sufficient confidence in databases (***41,42***).

There are a number of academic or commercially available pathway databases and network building tools available to the user. We have previously reviewed in some detail the major protein interaction databases, pathway maps and process ontologies (19,43). However, there many more databases and they generally enable the visualization of cellular components as networks of signaling, regulatory, and biochemical interactions. These databases include at the time of writing: human protein reference database (HPRD [44]) is the most advanced public source of curated human protein–protein interactions; a joint project between Johns Hopkins University and the Indian Bioinformatics Institute. HPRD currently contains more than 25,000 interactions for 15,000 human proteins, protein domain information, and seven signaling maps. Drawbacks of HPRD include there being no network building tools and this database can only be browsed at present. Second, most of the cited interactions are indirect and only two interaction mechanisms are supported.

Database of interacting proteins is a database of experimentally determined protein–protein interactions, mostly from yeast. About 10% of database of interacting proteins interactions are derived from high-confidence small-scale experiments (41).

Biomolecular interaction network database is a curated database of interactions, derived both from the literature and experimental data sets. Currently more than 8500 interactions are deduced from high-confidence small-scale experiments from multiple species. Biomolecular interaction network database can be used for querying and browsing (45).

Molecular interaction database is a searchable interaction database with more than 40,000 total interactions, mostly from yeast and fly. Seventy percent of these interactions are from lower-confidence Y2H screens. Only 3800 interactions include human proteins (46).

PathArt (Jubilant Biosystems; http://www.jubilantbiosys.com/pd.htm) is a manually curated database of about 7500 protein–protein and protein–compound interactions and pathways.

Pathways knowledgebase (Ingenuity, Inc.; http://www.ingenuity.com) represents mammalian interaction content for which the number of interactions has not been described.

ResNet database (Ariadne Genomics; http://www.anadnegenomics.com) is an automatically extracted and manually validated database of human protein interactions (more than 30,000), with transcriptional regulation (10,000), protein modifications (10,000), and functional regulations (350,000) (27).

Search tool for interacting genes and proteins is a database of known and predicted protein interactions deduced from more than 110 genomes using high-throughput experiments and gene coexpression data (47).

Other categories for the functional classification of proteins are gene, protein, and processes ontologies. Unlike pathways in which proteins are connected through consecutive single-step reactions and direct interactions, in ontologies proteins are assigned to particular categories because of functional or sequence similarity. They might or might not be physically connected in the cell. The best known among these is GO, a publicly available protein classification based on cellular processes developed by the GO Consortium (48) and curated by the European Bioinformatics Institute. Another popular process classification, PANTHER (protein analysis through evolutionary relationships) classification system is currently freely available from Applied Biosystems and classifies proteins according to families (>6683) and subfamilies (31,705), molecular functions and biological processes.

2.3. Network Theory and Tools

In recent years, it has become apparent that biological networks are ubiquitous (26) and therefore network analysis represents a powerful tool for the functional mining of large, inherently noisy experimental data sets. Proteins are represented as the nodes and physical protein–protein and protein–DNA interactions are represented as edges on these networks (4,49). One-step binary interactions between proteins can be extracted from the experimental literature and when combined they form multistep modules and pathways. These inturn are connected into higher level clusters of multistep pathways using all proteins of known function (50). Such protein interactions described for one cell type and condition are also possible in other cells and tissues, resulting in billions of possible multistep network combinations (***19***), although only a fraction might be in use at any one time. A subset of the functioning or activated cellular machinery can be captured by high-throughput experiments, which in turn can be visualized on networks using unique algorithms.

There is an important distinction to be made between networks and other functional analysis tools. Unlike the preset, somewhat arbitrary groupings of objects into categories like pathways and processes, network edges bear the primary experimental information on connectivity between proteins, their subunits, DNA sequences, and compounds. The complete set of interactions therefore defines the potential of the core cellular machinery with potentially billions of physically possible multistep combinations (***19***). Obviously, only a fraction of all possible interactions are activated with any given condition, as only some (10–30%) of the genes are expressed at a given time in a tissue and only a fraction of the cellular protein pool is therefore active. The subset of activated (or repressed) genes and proteins are unique for the experiment and are captured as snapshots by high-throughput data. Because each interaction represents a binary connection between individual proteins, the network gives the highest possible resolution of the resulting data patterns. The use of such networks can, therefore, overcome some of the drawbacks of high-throughput analysis. First, the interaction modules graphically represent biological mechanisms connecting the data. Second, interaction sets are comprehensive and eventually cover the majority of an organism's genes and proteins. Third (and most importantly), networks are dynamic and unique for each data set. Once generated, the networks can be interpreted in terms of these higher level processes, and the mechanism of an effect can be elucidated. This is achieved by linking the network objects to GO (***43***) and other process ontologies, as well as metabolic and signaling maps, and statistical analysis.

Biological networks are currently generated and analyzed by the methods of modern graph theory (***26***). The default random network theory states that pairs of nodes are connected with equal probability and follows a Poisson distribution. This implies that it is very unlikely for any node to have significantly more edges than average (***51***). As the field of network analysis has developed biological networks in yeast and elsewhere have been shown to be nonrandom (***28,29,50***). The distribution of edges is very heterogeneous in these networks, with a few highly connected nodes (hubs) and the majority of nodes possess very few edges. Such a topology is defined as scale free, meaning that the node connectivity obeys the power law: $P\ (k) \sim k^{-g}$, where $P\ (k)$ is the fraction of nodes in the network with exactly k links (***52***). The hubs of such networks are predominantly connected to low-degree nodes, a feature that gives biological networks the property of robustness. Hence, the removal of even a substantial fraction of nodes still leaves the network connected (***53***). The overall possible network topology correlates with the biological properties of the constituent node proteins (***31,54,55***) as highly connected hubs are conserved by evolution (***28,30,31***). These essential proteins also tend to be more closely connected to each other. Furthermore, essential proteins are frequently the more promiscuous transcription factors and target genes that are in turn regulated by fewer transcription factors. Many of these targets are known as house-keeping genes with high-expression levels and demonstrate less fluctuation in expression (***56***). The use of such network visualizations suggests an organized modularity in complex systems (***57***), which has also been applied to interpret the connectivity of small molecules and their interaction with proteins (***58–61***). Combined, these findings might have substantial implications for the practice of drug discovery in terms of target prioritization and identification of multigene/multiprotein biomarkers.

Network organization therefore has a characteristic importance for all levels of information flow in the cell. The networks can be generated at each level directly from available high-throughput data and assembled from protein–protein interaction databases. Although, traditionally the differentiation of all cellular processes into metabolic and signaling components is the norm, in reality, in a living cell both cascades work together and are codependent. In this case many endogenous compounds act as ligands for signaling cascades and almost all transcription factors, ultimately regulating metabolic enzymes and transporters. Some of these networks at different levels of the cell will now be described.

Genetic association networks visualize interconnections between the gene variants associated with a certain phenotype or disease, which are typically associated with mutations, single nucleotide polymorphism (SNPs), and in some cases, chromosomal rearrangements on the order of dozens to hundreds (in the case of cancers) of genes. Disease-associated genes vary greatly by their impact in particular diseases and are interconnected by complex epistatic relationships ***(62,63)***. Although, more than 3000 disease associated genes are described and stored in the National Center for Biotechnology Information's (NCBI) online Mendelian inheritance in man (OMIM) database, disease-specific epistatic clusters (or networks) are poorly studied because of the inherent complexity of disease genetics and to the lack of tools with which researchers can tackle the problem. Recently, some of the first network studies were conducted for Alzheimer's disease ***(64)*** and glaucoma ***(19)***. In both cases, large sets of protein–protein interactions were used to connect approx 60 genes associated with the diseases and provided new insights into potential therapeutic target genes. This approach is likely to be repeated for many other diseases.

Genome-wide transcription data are ubiquitous in disease research, thanks to the relative robustness and reliability of mRNA microarray technology. Based on the assumption that functionally related genes should be cotranscribed at the same time under the same conditions, several computational methods have been applied for the generation of gene coexpression networks in human and model organisms ***(65)***. In one study, more than 3000 individual microarrays from human, fly, worm, and yeast were tested for coexpression of orthologs in multiple organisms (metagenes) and 3400 such orthologous metagenes appeared to be connected through 22,000 interactions ***(66)***. By virtue of the underlying experimental data, coexpression networks mostly describe transcriptional regulation and at a basic level includes transcriptional factors and their downstream targets. Regulatory networks are however topologically complex and multilayered, with basic elements organized in; small one-step connected motifs which repeat frequently in the networks; semi-independent larger modules consisting of several motifs and, finally, the whole network as an interconnected set of modules ***(65)***.

The level of active proteins impacts the information flow in the cell, as proteins are the main building blocks for biological function. In recent years it was realized that most proteins function as physically connected complexes best described as combinations of physical binary interactions ***(55)***. Protein interaction networks show properties similar to other networks, with a few highly connected proteins and domains defining the network topology ***(67)***. These hubs can be separated into two types: "party hubs" of simultaneously connected partners and "date hubs" with different partners at different times and conditions ***(57)***. An important new direction consists of deducing protein interactions based on structural domain information ***(68)***. For instance, proteins are considered to be interacting if their domain sequences are compatible with the X-ray structure of heterodimers ***(69)*** or with domains that have been observed among interacting proteins ***(70)***.

A study on 43 organisms including human showed that similar to other types of biological networks, endogenous metabolic networks follow the power-law distribution with a few highly connected major metabolites, and that any two metabolites can be connected by at most three steps ***(71)***. Metabolic networks allow a semiquantitative evaluation of the balance of major metabolites, known as metabolic flux analysis and constraint-based modeling ***(72)***.

2.4. Other Resources for Network Analysis

The 2005 version of Molecular Biology Database Collection, the benchmark resource compiled at NCBI, contains more than 700 mostly public databases of variable quality and utility. However, several other biology resources are available.

- NCBI includes many useful subsections; Gene, Nucleotide, OMIM, UniGene, SNP, and PubMed. These all provide additional information about genes and proteins. PubMed is especially useful because it is a repository for published papers in the form of abstracts.
- The human genome database contains annotations for the Human Genome.

- The Human Gene Mutation Database lists any mutations for a given gene; it also lists the information in which mutations were first reported.
- European Molecular Biology Laboratory Contains the sequence for a given gene or protein along with PubMed references.
- SwissProt (Swiss Protein) lists additional information about proteins. It contains information such as synonyms, accession numbers, and links to related Medline and PubMed articles.
- The protein information resource supports genomic and proteomic research. The protein information resource maintains the Protein Sequence Database, which is a protein database containing more than 283,000 sequences.
- The Protein Data Bank is a worldwide repository for the processing and distribution of 3-D structure data of large molecules, for example, proteins and nucleic acids.

3. Integrated Network Data-Mining Suites

3.1. MetaCore

MetaCore is a web-based computational platform for multiple applications in systems biology. It is primarily designed for the analysis of high-throughput molecular data (microarray-based and serial analysis of gene expression (SAGE) gene expression, array-comparative genomic-hybridization DNA arrays, proteomics data, metabolic profiles, and so on) in the context of human and mammalian networks, canonical pathways, diseases, and cellular processes. MetaCore is an integrated system, which consists of (1) a curated database of mammalian biology, (2) a suite of tools for querying, visualization, and statistical analysis including pathways maps, network algorithms, and filters, (3) a toolkit (pathway editor) for custom assembly of functional networks, and (4) a set of parsers for uploading and manipulating of different types of high-throughput molecular data ***(19,43)***.

3.1.1. Content

As a foundation, MetaCore has a database of protein–protein, protein–DNA, and protein–compound interactions, metabolic reactions, pathway maps, bioactive compounds (metabolites, drugs, and ligands), and diseases. Human pathways have been manually collected from the experimental literature for more than 5 yr. This represents one of the most comprehensive databases in the field, the core of MetaCore consists of more than 4.5 million individual findings resulting in about 50,000 signaling interactions and 20,000 human metabolic transformations (covering both endogenous and xenobiotic metabolism). The database has interaction information for more than 90% of known human proteins, including 1720 transcription factors and 650 GPCRs. This content is linked to 3200 human diseases and conditions. The bioactive chemistry component includes more than 7000 known drugs with protein targets and 5000 endogenous metabolites. The pathway information is organized in more than 400 signaling and metabolic maps with more than 3000 canonical pathways represented. This information is organized in an Oracle database.

3.1.2. MetaCore Database Architecture

The software currently runs on an Intel-based 32-bit server running RedHat Linux Enterprise 3 AS (RedHat, Raleigh, NC) and the web server runs Apache 1.3.x/mod_perl. Software on the server side is written in Perl, whereas the client side requires HTML/JavaScript and the Macromedia Flash Player Plug-in (Macromedia Inc, San Francisco, CA). The MetaCore database is generated from manual annotation of full text articles as well as disease relevant information from OMIM and EntrezGene. This database has functional processes as the core objects, which can be unique and have different relationships with molecular entities. We have used three major types of functional processes: effects, transformations, and blocks. In addition, we introduce the notion of a component that describes molecular species or functional groups of molecules in their biological context described as follows.

3.1.2.1. Component

This represents major functional objects in the context of a biological system. For example, a component might represent a single gene and its protein product, protein complex, a family of related proteins, small molecules, such as drugs, and metabolites. In the visual representation scheme of MetaCore, component corresponds to objects represented on the network and pathway maps. A component is related to a molecular entity, localizations, cells/tissues, and/or organisms. Thus, network classes represent biological molecules within their biological context. The molecular entity is treated in a broader sense than just being a specific chemical compound. In our current representation a component could also be a group of molecules (e.g., a protein family or class of chemical compounds) or a molecular complex. This is particularly useful for representing cellular processes, or when the exact chemical composition or a particular isoform of a protein participating in a pathway is unknown or ambiguous (e.g., EC numbers). Essentially the component category unites proteins, and compounds (small molecule ligands, endogenous metabolites, xenobiotics), and proteins. For example, p53 and ATP localized in the nucleus would both be components. Similarly, ATP in the cytoplasm will present itself as another component. In the nucleus, ATP is needed for RNA synthesis and in the cytoplasm as an energy source. Activated (phosphorylated) p53 in the nucleus is a potent transcriptional factor and a different component than the inactive p53 in the cytoplasm. At the same time, the family of integrins can be considered as one component in certain conditions.

3.1.2.2. Transformation

This is an entity that is used to store information on biochemical reactions, transport, transcription, and translation, or on any biological process whose primary function is to change the state of a molecule (e.g., a reaction, in a broad sense), which is considered in its particular environment as linked to a subcellular compartment, tissue and organism. Transformation defines the morphing of components into each other. One example of a transformation is a one-step metabolic reaction, such as the synthesis of ATP from ADP and phosphate during oxidative phosporylation in the mitochondria. The transfer of ATP from the mitochondria to the cytoplasm is also a transformation. Another example is of protein phosphorylation in signal-transduction cascades. In all these cases, transformations share the property of modification of one component into another.

3.1.2.3. Effect

This is an entity that represents the influence that molecules exert either on transformations or on each other. Each effect has an agent (a component, which corresponds to the molecule[s] involved), a target (transformation, another component, or entire block [*see* **Subheading 3.1.2.4.**]), type, and a set of numerical values that could be associated (e.g., a kinetic end point). The notion of an effect is convenient for the description of biological activity whether or not the exact mechanism is known, as incomplete information can be stored allowing for the later reconstruction of cellular networks. For example, H-dependent ATP-synthase catalyzes ATP synthesis. This effect is presented as a link between the protein and the corresponding reaction. Another example of an effect is the phosphorylation of p53 by CHLK1 kinase. In turn, p53 is a transcription factor whose effect is modulation of transcription of multiple genes. This effect is the first step in a long chain of chemical and transport transformations induced by activation of the transcription of multiple genes.

3.1.2.4. Block

This is used to describe functional units, be it a particular category of metabolism, or any other functional process. Blocks link together components, effects, and transformations that are themselves functionally related. Blocks are hierarchical as they might contain other blocks as elements. On the other hand every element might be a part of more than one block. Blocks are linked to each other by shared elements. Assembling different entities within functional blocks therefore enables the rapid searching of functional links and the function-centered analysis of expression and other high-throughput molecular data. A sequence of chemical reactions and regulatory protein interactions might have its own biological effect, and therefore can be identified as a block. An example of metabolic block is the oxidation of pyruvate followed by CoA attachment. This process consists of six discrete metabolic steps connected in three cycles and catalyzed by three enzymes. These enzymes form a pyruvate-dehydrogenase complex needed for coregulation kinetics of all reactions in such a way that the products of a downstream reaction are timely and directly assessable to the active site of an upstream enzyme. As a result, the coregulated and spatially unified set of six reactions can be considered as one functional block. At a higher level, The Krebs Cycle can be summarized as a unified block of reactions, which transform pyruvate into carbon dioxide, a process accompanied by reduction of NAD, ubiquinone, and GTP synthesis from GDP. On the one hand, a block can be divided into individual reactions, linked by functional (kinetic or regulatory) connections. On the other hand, it can also be considered as a united and separate entity interacting with other blocks.

This overall organizational structure has been described as a graphic diagram previously *(19,73)*. To summarize, functional processes and components serve as the core information space-holders in our database with many-to-many relationships between them. The corresponding molecular and mechanistic data are then linked to these space-holders as they become available. Functions serve as the "linking portals" for heterogeneous data. Once linked, the heterogeneous types of high-throughput data become a part of a larger system-level picture, in which functional relations among them can be more easily established and elucidated (e.g., all proteins in a pathway and their genes with expression patterns). Every pathway and its elements (interactions, reactions, enzymatic functions) are linked to available molecular data (genes, proteins, compounds, expression data, SNPs, and so on) annotated with relevant information about their involvement and importance in a number of common human diseases. This software allows the superimposition of relevant biological data such as microarray, SAGE expression data, metabolic profiles, and protein interactions on the pathways networks. Currently, by manual annotation of full text articles we have established more than 32,000 links between pathways from our collection and more than 3200 disease states, classified into six major categories:

- Cause—The highest level of verification. Cause means that it was clearly established experimentally that a deviation in a pathway directly causes the disease.
- Manifestations—A clinically confirmed strong correlation between the disease and a deviation in the pathway, but without the direct evidence for cause.
- Hypothesis—A correlation between the disease and the pathway has been demonstrated in some cases, but not in the others.
- Animal models—A correlation is described for one or more model organisms.
- Treatment—The changes in the pathways observed during or after therapeutic intervention.
- No relation—No links have been found between the disease and particular pathway or its elements.

3.1.3. Network Algorithms and Filters

Within MetaCore the networks are generated as a combination of binary single-step interactions (edges) which connect proteins and genes (nodes). The nodes and edges derive from the corresponding interaction tables in the MetaCore database and are visualized as clusters of interconnected nodes with the Macromedia Flash Player Plug-in. The end nodes on the networks have only one edge; the internal nodes might have anywhere from two to several hundred edges depending on connectivity with other nodes. The networks can be built from any input list of genes, proteins, and compounds corresponding to the components (network classes) in the database. The nodes in the input list are therefore considered as root nodes. The input list can be generated in several ways. Gene and protein names can be input and recognized with a built-in synonyms dictionary; gene lists can also be imported as text and Excel files or directly parsed from Affymetrix (www.affymetrix.com), Agilent (www.agilent.com), and other microarray analysis software. MetaCore recognizes most of the commonly used gene and protein identifiers such as LocusLink, SwissProt, RefSeq, and Unigene. The process that ultimately results in the generation of a network can be initiated as described as follows.

First-pass data filtration: Before building networks, the interactions can be preselected based on the level of trust, interaction direction, effects, mechanisms, and tissue specificity (in which only the edges with both nodes belonging to a chosen tissue remain). The nodes from the input list with no connections with other nodes on the list are removed. The edges of networks can be assigned with weights depending on the type of interactions. When a user sets up the list of objects (genes or proteins are translated into components or network "classes"), these are represented as nodes and are visualized connected by the edges if there is an interaction between two nodes, which is present in the interactions table. The edges can then be assembled into clusters by using one of multiple available algorithms described briefly as follows:

3.1.3.1. Direct Interactions

The direct interactions algorithm is the most stringent, in which the only edges allowed are those between two nodes which are root nodes, e.g., objects from the list directly connected to each other.

3.1.3.2. Shortest Paths

The shortest path algorithm is based on Dijkstra's algorithm, which efficiently finds the shortest paths from a given vertex *x* to all *n*-1 other vertices.

3.1.3.3. Analyze Network

The analyze networks algorithm builds on Dijkstra's algorithm and takes a list of root nodes and for each node creates shortest paths networks to the other root nodes in the list and stops the network at a size defined by the user in the advanced options. This process is repeated iteratively until every node from the list is included in at least one network. Each subnetwork is associated with a *G*-score and *p*-value (**Subheadings 3.1.5.** and **3.1.6.**), which rank the subnetworks according to saturation with the objects from the initial gene list.

3.1.3.4. Auto Expand

The Auto expand algorithm starts with a number of root nodes as specified by the user and builds subnetworks around every object from the uploaded set consisting of nearest neighbors. The expansion halts when the subnetworks intersect. The objects that do not contribute to connecting subnetworks are automatically truncated and there is no user control over the size of the network. Each connection represents a direct, experimentally confirmed, physical interaction between the objects.

3.1.3.5. Analyze Transcriptional Regulation

Analyze transcriptional regulation finds transcriptional regulators in the list of nodes or transcriptional regulators that are related to those on the input list and suggests multiple possible networks from this starting point.

3.1.3.6. Self Regulations

The Self Regulations algorithm is similar to the Shortest Paths algorithm except that it tries to build paths that contain transcription factors to identify networks that are held together by regulatory loops. The algorithm takes paths with transcriptional factors that follow the same direction.

3.1.3.7. Expand By One Interaction

The Expand By One Interaction algorithm simply sums up all one-step interactions around each root node and finds islands of objects from the user's list connected by no more than one bridging object.

3.1.4. Mapping Experimental Data on Networks

Every node on the network is associated with genes and proteins through the tables in the general database schema. The novel database architecture enables mapping of the high-throughput experimental data associated with genes and proteins onto the networks (**Fig. 2**). Every experimental data point (e.g., a set of probes on the microarray or a frequency for a certain SAGE tag) represents an attribute of the unique gene or protein identifier. Therefore, the high-throughput data can be linked with the corresponding node in the database and visualized on the networks containing this node. Visually, the altered expression or protein abundance data is presented as a solid circle above the node (red and blue represent increased and decreased abundance, respectively).

The applications of this quite straightforward procedure are ubiquitous in basic research and in drug discovery. For instance, one can directly compare the lists of genes derived from different types of high-throughput or "small-scale experiments" on the same networks. When the same data type and experimental platform is used, the conditional networks can be readily compared for common and different subnetworks and patterns. Such fine grained mapping can also be performed in order to compare the tissue and cell type specific response, different time-points, drug dosage, and different patients from the same cohort, and so on.

3.1.5. Network Statistical Analysis: Scoring and Prioritization According to the Relevance of Input Data

In many cases the high-throughput experimental data sets are very large. For example, a differential gene expression pattern from the whole human genome array in complex diseases might include many thousands of genes. In such cases, the issue of prioritizing networks and modules becomes increasingly important (**Fig. 2**). Prioritization can be based on different parameters, but follows the same procedure, which we will describe next. A data set of interest (e.g., the list of all prefiltered nodes) is divided into two random subsets overlapping in this general case. The size of the intersection between the two sets represents a random variable with a hypergeometric distribution. We apply this fact for numerical scoring and prioritization of the previously discussed node-centered small shortest path networks. Let us consider a general set size of N with R marked objects/events (e.g., the nodes with expression data). The probability of a random subset of size of n, which includes r marked events/objects is described by the distribution:

$$P(r,n,R,N)=\frac{C_R^r\cdot C_{N-R}^{n-r}}{C_N^n}=\frac{C_n^r\cdot C_{N-n}^{R-r}}{C_N^R}=\frac{R!\cdot(N-R)!}{N!}\cdot\frac{n!\cdot(N-n)!}{r!\cdot(R-r)!}\cdot\frac{1}{(n-r)!\cdot(N-R-n+r)!}.$$

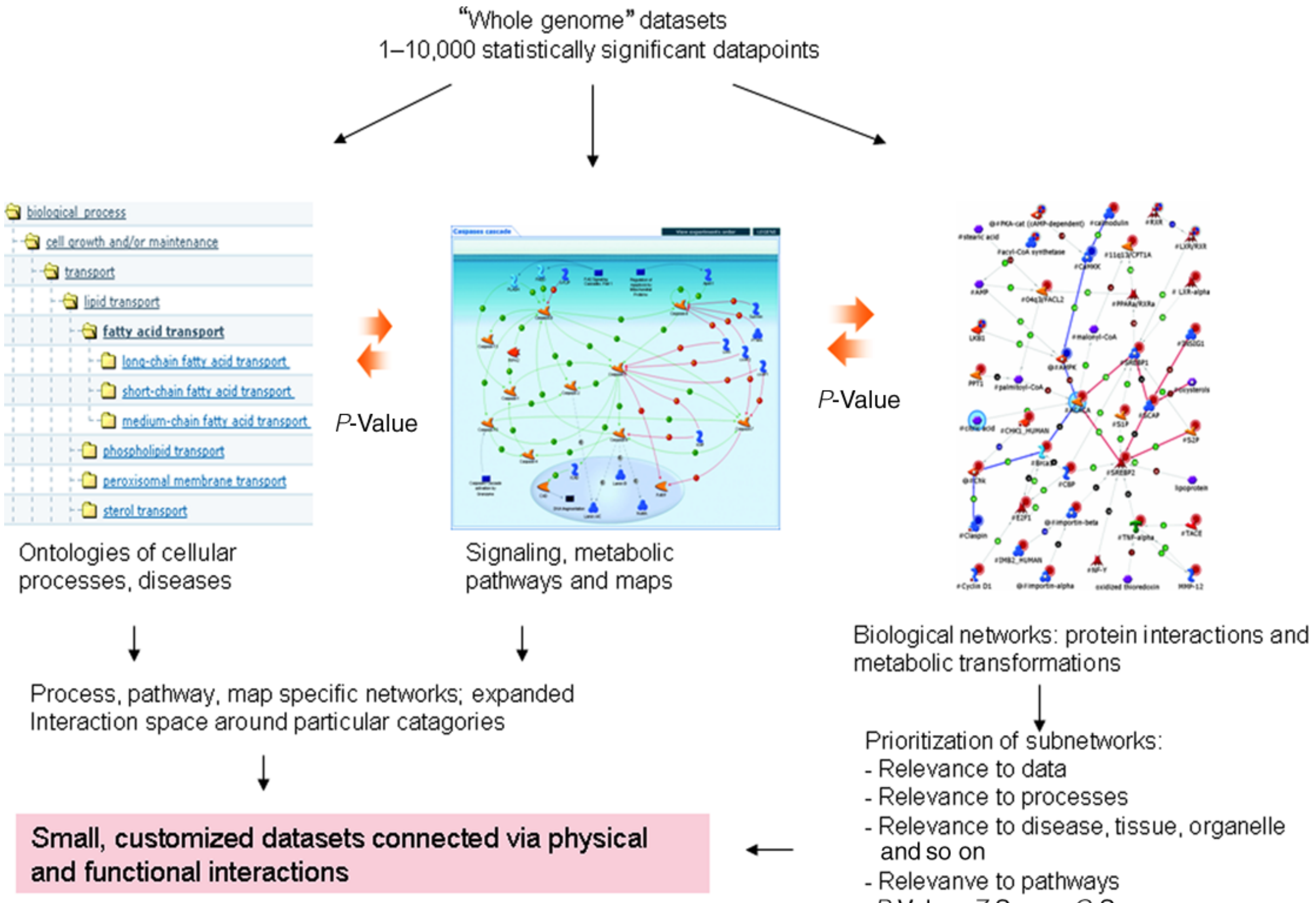
The mean of this distribution is equal to the following:

$$\mu=\sum_{r=0}^{n}r\cdot P(r,n,R,N)=\frac{n\cdot R}{N}=n\cdot q,$$

where, $q = R/N$ defines the ratio of marked objects.

The dispersion of this distribution is described as follows:

$$\sigma^2=\sum_{r=0}^{n}r^2\cdot P(r,n,R,N)-\mu^2=\frac{n\cdot R\cdot(N-n)\cdot(N-R)}{N^2\cdot(N-1)}=n\cdot q\cdot(1-q)\cdot\left(1-\frac{n-1}{N-1}\right).$$



“Whole genome” datasets
1–10,000 statistically significant datapoints

biological_process
cell growth and/or maintenance
transport
lipid transport
fatty acid transport
long-chain fatty acid transport
short-chain fatty acid transport
medium-chain fatty acid transport
phospholipid transport
peroxisomal membrane transport
sterol transport
Caspases cascade

Ontologies of cellular processes, diseases

P-Value

Signaling, metabolic pathways and maps

P-Value

Biological networks: protein interactions and metabolic transformations

Process, pathway, map specific networks; expanded Interaction space around particular catagories

Prioritization of subnetworks:
- Relevance to data
- Relevance to processes
- Relevance to disease, tissue, organelle and so on
- Relevanve to pathways *P*-Value, *Z*-Score, *G*-Score

Small, customized datasets connected via physical and functional interactions

Fig. 2. Functional prioritization and analysis of high-throughput or high content data sets. Networks are scored and prioritized based on the relative relevance of the nodes to functional processes (GO) or static maps of canonical pathways. The software tools can be used to elucidate network modules and novel pathways as hypotheses for biomarker and therapeutic target discovery. (Please *see* the companion CD for the color version of this figure.)

It is essential that these equations are invariant in terms of exchange of n for R. This means that the "subset" and "marked" are the equivalent and symmetrical sets. Importantly, in the cases of $r > n$, $r > R$, or $r < R + n - N$, $P(r, n, R, N) = 0$

We will use the following z-scoring for comparison and prioritization of node-specific shortest path subnetworks:

$$z - \text{score} = \frac{r - n\dfrac{R}{N}}{\sqrt{n\left(\dfrac{R}{N}\right)\left(1 - \dfrac{R}{N}\right)\left(1 - \dfrac{n-1}{N-1}\right)}} = \frac{r - \mu}{\sigma},$$

Where,

N is the total number of nodes after filtration;

R the number of nodes in the input list or the nodes associated with experimental data;

n the number of the nodes in the network;

r the number of the network's nodes associated with experimental data or included in the input list;

μ and σ are, respectively, the mean and dispersion of the hypergeometric distribution described earlier.

We have also devised a variant of this score termed the G-score. The G-score combines the Z-score and the sum of the squares of the interactions to and from each of the nodes not related to the initial list. The value for the K coefficient can be specified by the user:

$$G - \text{score} = z - \text{score} - \frac{4}{K \cdot \sqrt{n - r}} \sum_{i \in \{n \xi r\}} V_i^2$$

where,

n is the total number of nodes in each small network generated from user's list;

r the number of nodes with data in each small network generated from user's list;

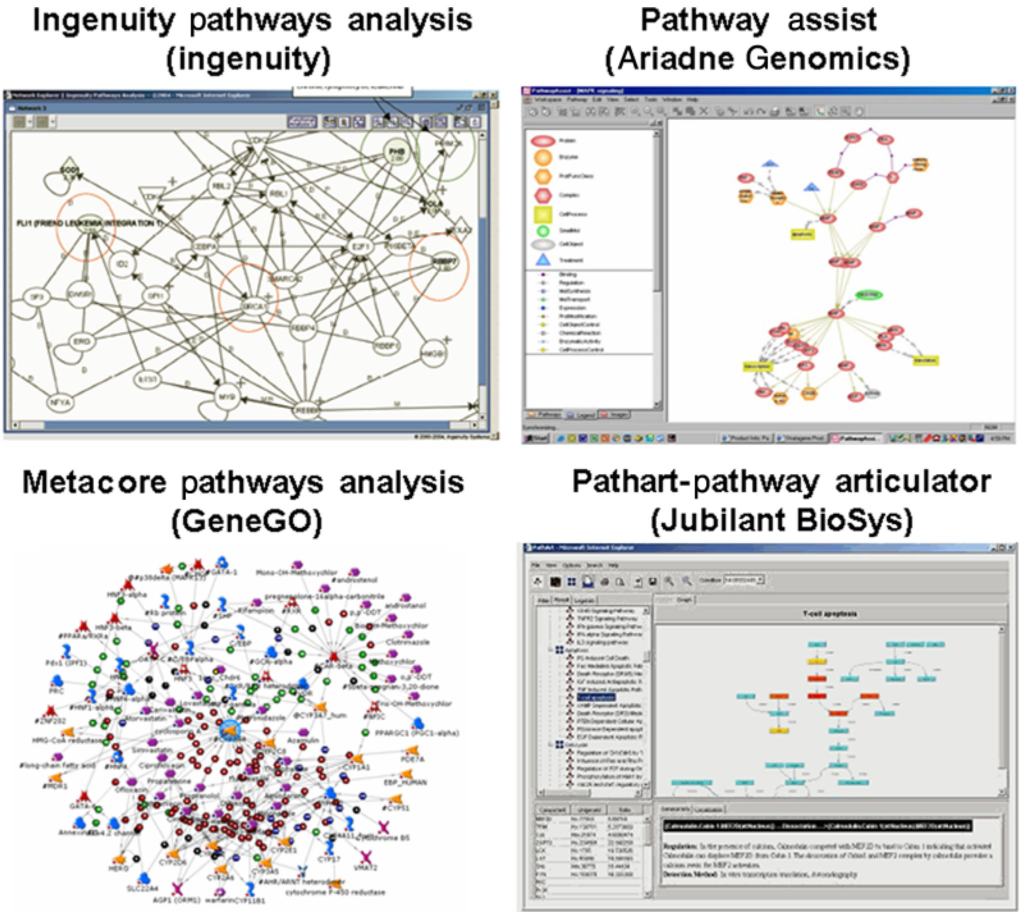
V_i the number of links to/from i-th node;

{n/r} denotes the set of nodes in a small network that are not related to user's list;

K is the user-specified coefficient-used to "demote" networks with high-degree nodes that do not correspond to genes/proteins in user's list.

3.1.6. p-*Value and Evaluation of Statistical Significance of Networks*

For a network of a certain size, we can evaluate its statistical significance based on the probability of its assembly from a random set of nodes of identical or similar size to the input list. We can also evaluate the relevance of the network based on biological processes (defined as a subset of the network nodes associated with the particular process) or any other subset of nodes (**Fig. 2**). For example let us consider a complete set of nodes on the network, divided into two overlapping subsets. These subsets represent the nodes linked to a certain predefined node list (e.g., the list of nodes belonging to GO cellular processes, or a list of genes expressed in a certain tissue) (**Fig. 2**). Generally, these subsets are different but overlapping. Assuming that the intersection between the two subsets is large enough and nonrandom (we do not consider a situation when the intersection is small but nonrandom) the null-hypothesis states that the subsets are independent and, therefore, the size of the intersection satisfies a hypergeometric distribution. The alternative hypothesis states that there is positive correlation between the subsets. Based on these assumptions, we can calculate a p-value as the probability of intersection of the given or a larger size network from two random subsets from the same set.



Ingenuity pathways analysis (ingenuity)

Pathway assist (Ariadne Genomics)

Metacore pathways analysis (GeneGO)

Pathart-pathway articulator (Jubilant BioSys)

Fig. 3. Screenshots of the four major commercial software suites for network and pathway analysis. (Courtesy of Jack Collins, National Cancer Institute.) Note that Pathway assist in now known as Pathway Studio. (Please *see* the companion CD for the color version of this figure.)

$$pVal(r,n,R,N)=\sum_{i=\max(r,R+n-N)}^{\min(n,R)} P(i,n,R,N)=\frac{R!\cdot n!\cdot (N-R)!\cdot (N-n)!}{N!}\sum_{i=\max(r,R+n-N)}^{\min(n,R)}\frac{1}{i!\cdot (R-i)!\cdot (n-i)!\cdot (N-R-n+i)!}.$$

3.2. Other Pathway Suites

Currently there are at least three other commercially available pathway analysis software suites that are widely used by the pharmaceutical industry and major research institutes (Fig. 3). They all have slightly different data and visualization capabilities.

3.2.1. PathwayAnalysis (Ingenuity, Inc.)

An integrated analytical suite based on a manually curated database of literature-derived mammalian protein–protein interactions used for visualization of data on networks and analysis. Networks are connected to GO processes, ~60 KEGG metabolic maps and Cell Signaling Inc.'s signaling maps. Web access and an enterprise solution is available.

3.2.2. PathArt

A curated database of generic protein interactions described earlier, pathways and bioactive molecules supported by high-throughput data parsers and visualization tools. This tool has connectivity with ligand databases and GO categories, while web-based access is available.

3.2.3. Pathway Studio (Formerly Known as Pathway Assist) (Ariadne Genomics)

A desktop software tool for mapping the high throughput data on networks, maps, and pathways. The source of the interaction data is NLP mining of PubMed abstracts. PathwayAssist is bundled with Jubilant and Integrated Genomics pathways content.

3.3. MetaDrug

The parallel development of different high-throughput methods, databases, absorption, distribution, metabolism, and toxicology (ADME/Tox) modeling, and systems modeling is currently ongoing *(74)* and will result in systems-ADME/Tox as a new area for research. We have used MetaCore as a foundation for building a software suite for ADME/Tox, called MetaDrug™. The ultimate goal of this platform is to predict from an input structure the major xenobiotic metabolites in humans and their predicted binding interactions with enzymes and other key ADME/Tox proteins in humans. MetaDrug includes more than 10,000 xenobiotic reactions, more than 1500 enzyme substrates, and 1000 enzyme inhibitors with kinetic data. MetaDrug has been used to derive some of the major metabolic pathways and determine the involvement of particular cytochrome P450s for compounds *(75,76)*. This database has also enabled us to generate more than 85 key metabolic pathways for predicting likely metabolic reactions coded in the software. A molecular structure can be parsed to rapidly create possible metabolites, which are prioritized using a further algorithm. In addition, the molecules can be scored using more than 40 integrated quantitative structure activity relationship (QSAR) models covering a wide range of ADME/Tox properties. Alternatively, the user can generate and use their own QSAR models with the software. Likely reactive metabolites for the input molecule/s are readily highlighted using 89 rules. Ultimately the predicted molecules and their interactions might be visualized as temporary objects with connections on a network diagram derived using one of two network algorithms.

To our knowledge MetaDrug is presently the only commercial product that combines all of the key properties of a human drug metabolism database, QSAR, rule-based methods for metabolism and reactive metabolite formation, and systems-biology approaches. The total effect of combining these different functions represents a significant step toward developing a Systems-ADME/Tox platform approach integrating computational predictions with data from all experiments to provide an understanding of the effect of xenobiotic and endobiotic molecules on ADME/Tox properties in humans *(76)*. The software also has additional valuable roles of providing a means to visualize predicted data in the context of empirical information on complex networks *(74)* and identify gene-signature networks *(77)*.

Future developments for MetaDrug include the integration with pipelining software such as Pipeline Pilot (SciTegic; www.scitegic.com) to allow MetaDrug to be used seamlessly as part of a larger data generation protocol such as, for large virtual library screening. Second, we will produce a version of MetaDrug with rat and mouse metabolism data in the underlying database, to enable predictions for these species and ultimately enable comparisons with the human predictions. Third, we are developing more sophisticated machine learning algorithms for metabolite prioritization to enable increased accuracy of predictions.

3.3.1. Applications of MetaDrug

Previously we have used MetaDrug to generate networks around nuclear hormone receptors (NHR) as well as analyze high-throughput microarray data *(20)*. MetaDrug was applied to analyze NHR, transcriptional factors and their associated interactions with other proteins, and small molecules relevant to drug disposition and toxicology to result in a very complex network using

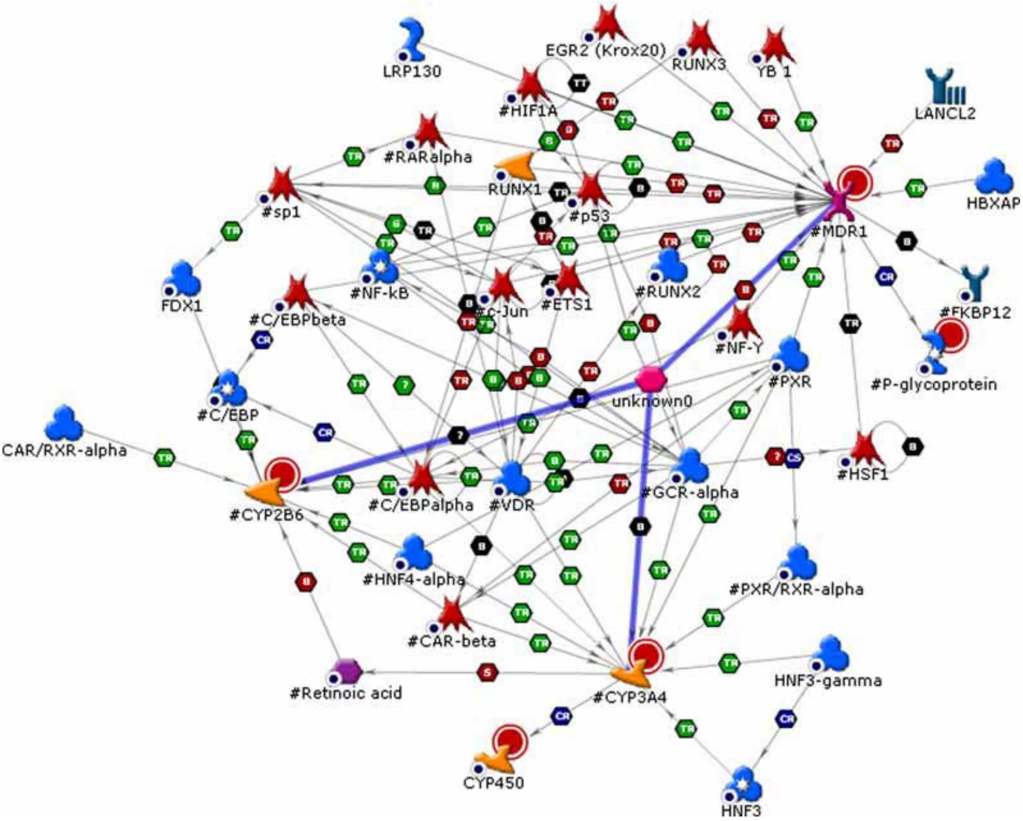


Fig. 4. Network visualization for Artmisinin (pink hexagon) using PCR data from human hepatocytes (red circles) ***(98)***. Small molecules are hidden from this network for clarity. Ligands (purple hexagons) linked to transfactors (red), enzymes (yellow arrows), and transporters (blue) from the MetaDrug database. Highlighted lines show predicted interactions. Small colored hexagons indicate the type of interaction between two nodes, for example, Tr, transcriptional regulation; P, phosphorylation; B, binding. (Please *see* the companion CD for the color version of this figure.)

the autoexpand algorithm ***(20)***. This visualization represented the current literature around NHR in terms of a network focused on the proteins of importance to drug disposition. Microarray data was also transposed on metabolic and signaling networks generated with this database using data from published experiments, in which MCF-7 breast cancer cells had been treated with 4-hydroxytamoxifen (OHT) for 24 h ***(78)***. A network was generated around the enzymes of interest relating to the metabolism of OHT and the microarray data was visualized on this. We have recently provided further test cases using MetaDrug that: enabled the prediction of metabolites for molecules based on their chemical structure; predicted the activity of the original compound, and its metabolites with various ADME/Tox models; incorporated the predictions with human cell signaling, metabolic pathways, and networks; and integrated networks and metabolites with relevant toxicogenomic, or other high-throughput data. We have demonstrated the utility of such an approach using recently published data from in vitro metabolism and microarray studies for Aprepitant, L-742694, Trovofloxacin, and artemisinin (**Fig. 4**), and other artemisinin analogs as well as OHT. This enabled us to show the predicted interactions with cytochrome P450s, pregnane X-receptor and P-glycoprotein, the metabolites and the networks of genes that are affected (Ekins. 2006). These examples represent how MetaDrug could be used as a novel method for analysis of computational predictions and microarray data on networks of interacting genes in order to visualize data in the context of the complete biological system. This provides insights for the up or down regulation of particular genes involved in a phenotypic response and also highlights genes not on the microarray but central to a network.

4. Examples Using MetaCore With Different Data

4.1. Mapping HCS Data on Networks

MetaCore is well suited for mapping phenotypic data such as HCS and siRNA as long as the data points are linked to either genes, or protein, or metabolic IDs. The mapping, visualization, and analytical procedures are virtually the same as mapping of molecular data such as gene expression or protein abundance. It is important that HCS data can be compared and cross-validated with molecular data on the same pathways and networks. One such analysis is currently in progress at the Translation Genomic Research Institute (Jeff Kiefer, personal communication). In this study, 162 genes/proteins were identified as hits from a high-throughput siRNA screen of 5000 genes constituting the "druggable genome" *(79)*. The shortest paths network was built from this data (**Fig. 5A**). When targeted by siRNA these genes were able to increase the sensitivity of the cancer cell line used to the effects of a low dose of a chemotherapeutic compound. The cell proliferation GO process was selected and traced on the same network (**Fig. 5B**) indicating an agreement with the observed data.

A recent study by Cellomics, Inc. (www.cellomics.com) describes the prototype FluoroTox system used for detection and classification of chemical and biological agents using HCS *(14)*. This focuses on one cell type and multiple parameters that were measured including p38 activation, NF-κB activation, NF-κB inhibition, CREB activation, ERK activation, and cytotoxicity. Apart from the latter general assay, the rest relate to four specific proteins, which can be mapped in MetaCore using the Analyze Network algorithm (**Fig. 6A**). A second HCS study studied a collection of 720 natural compounds to find inhibitors of the mitogen-activated protein kinase phosphatase-1 (MKP-1), a dual specificity phosphatase overexpressed in many cancers *(12)*. An alkaloid sanguinarine was found to inhibit MKP-1 and induce phosphorylation of *ERK* and *JNK*. Using MetaCore we can visualize the linkage between MKP-1, *ERK* and *JNK* with the Analyze Network algorithm (**Fig. 6B**). These MetaCore networks could therefore be used to visualize HCS data following treatment with different compounds to visualize the extent of protein deactivation or activation. This data can also be combined with any of the other data types described later.

4.2. Mapping Metabonomics Data

MetaCore and MetaDrug have the capability to upload metabonomics data either as a list of molecule names or molecular formulas at present. We have previously illustrated the frequency distribution of molecules and their molecular formula in MetaDrug *(76)*, which indicated the majority of molecular formulas corresponded to one to two metabolites. At present, we can visualize all the metabolites suggested for each unique formula or name, and after highlighting a molecule of interest further information can be retrieved including the molecule structure, synonyms, and reactions. We can also visualize these metabolites on maps or networks in MetaCore. Using a data set from a recent publication *(80)*, which determined the differences in hydrazine toxicity between rat and mouse by collecting urine and analysis using ^{1}H NMR, we are able to demonstrate this utility. We have used the endogenous urinary metabolites observed in rat and parsed them with our software to visualize the seven of the 17 metabolites alongside proteins on networks, after using the Analyze Network algorithm (G-score = 46.31, p = 9.30 e^{-18}, **Fig. 7**). These networks might be useful for indicating the type of toxicity that could be observed following compound treatment from metabolite data alone and lead to the generation of signature metabolite networks.

4.3. Mapping Genomics Data

4.3.1. Tat-Upregulated Genes at the G_1/S Phase

MetaCore has been primarily used to date for the analysis of microarray data *(19,73,77,81–87)*. To further illustrate the utility for analysis of genomics data we have taken a recently published

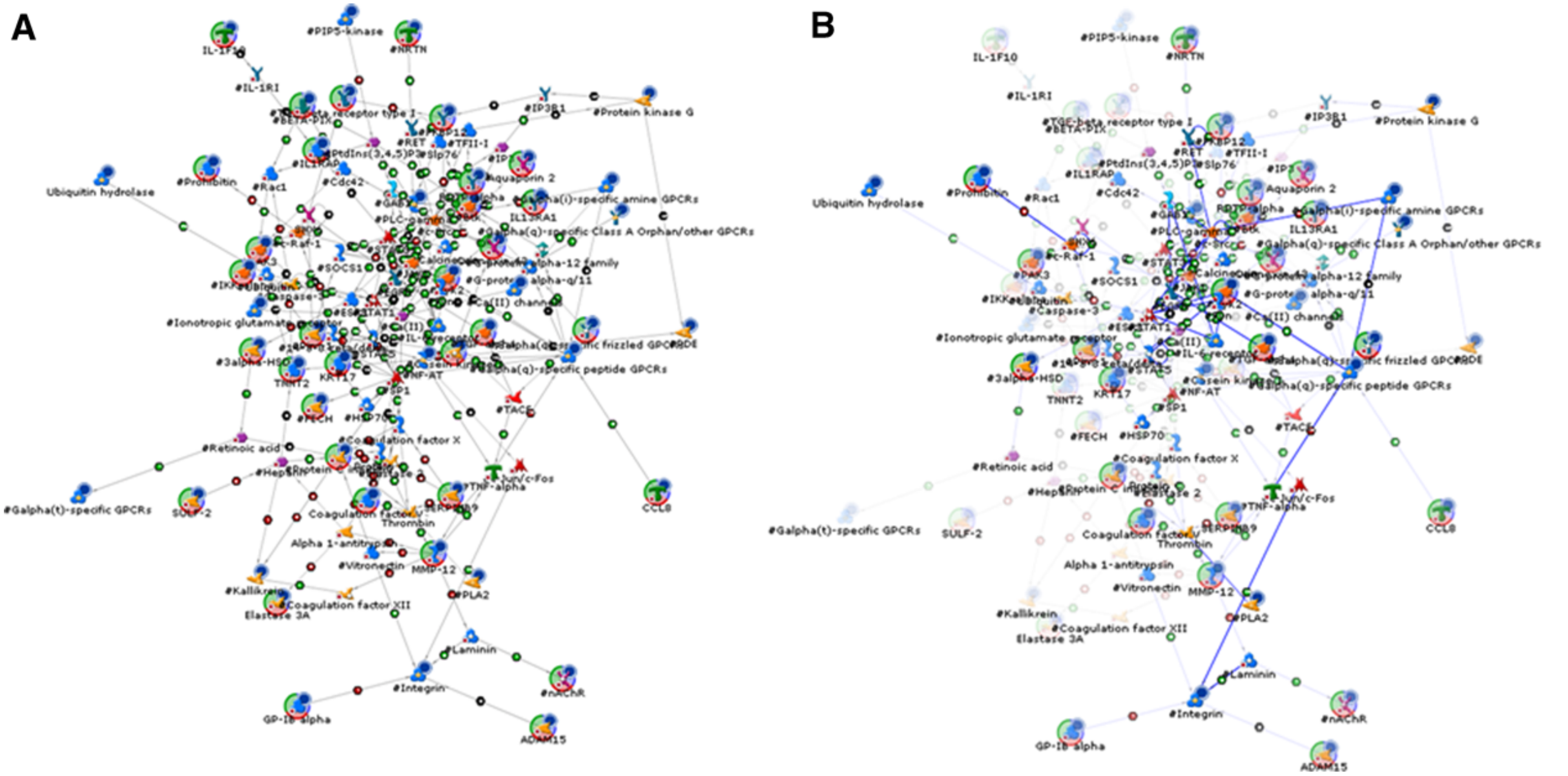


Fig. 5. Mapping of siRNA data on networks. High-throughput siRNA cell assays were conducted with 5000 genes and networks were built in MetaCore. **(A)** The network of 162 initial nodes was built run using Shortest paths algorithm. Tricolored circles mark the genes from input list; solid blue circles mark the nodes with numerical data. **(B)** GO process for cell proliferation is traced on the same network as a blue line. (Courtesy of Jeff Keifer, Translation Genomic Research Institute.) (Please *see* the companion CD for the color version of this figure.)

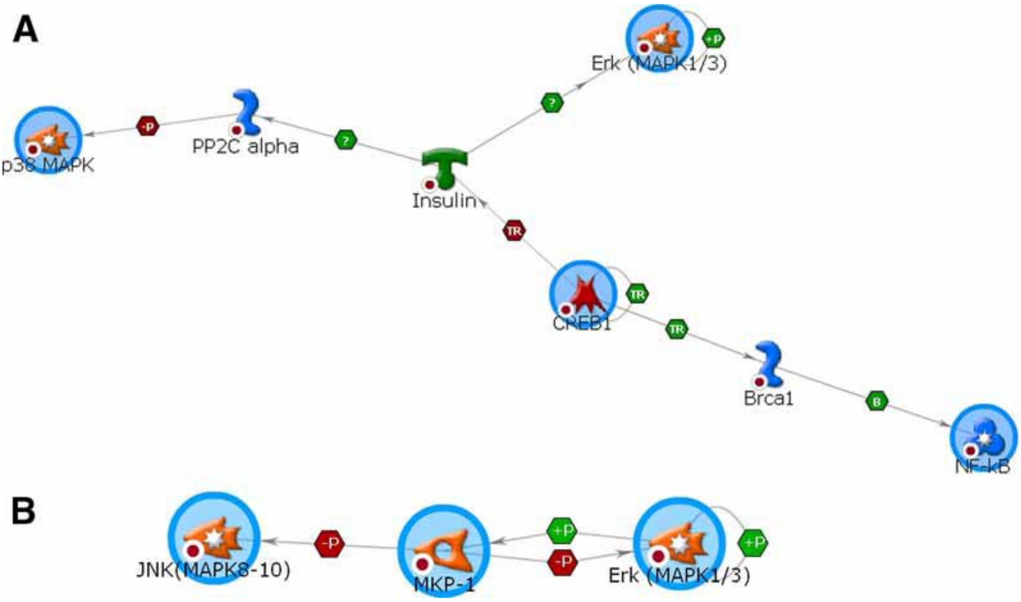


Fig. 6. Visualizing pathways from cellomics data. **(A)** Assays used in the FluoroTox system *(14)* visualized with MetaCore using the Analyze Network algorithm (*G*-score = 91.2, *p* = 1.70 e^{-14}), **(B)** Assays used with the alkaloid sanguinarine *(12)* visualized with MetaCore using the Analyze Network algorithm (*G*-score = 121.19, *p* = 1.81 e^{-12}). Nodes surrounded by a blue circle indicate those from the input list corresponding to therapeutic targets. Small colored hexagons indicate the type of interaction between two nodes, for example, Tr, transcriptional regulation; P, phosphorylation; B, binding. (Please *see* the companion CD for the color version of this figure.)

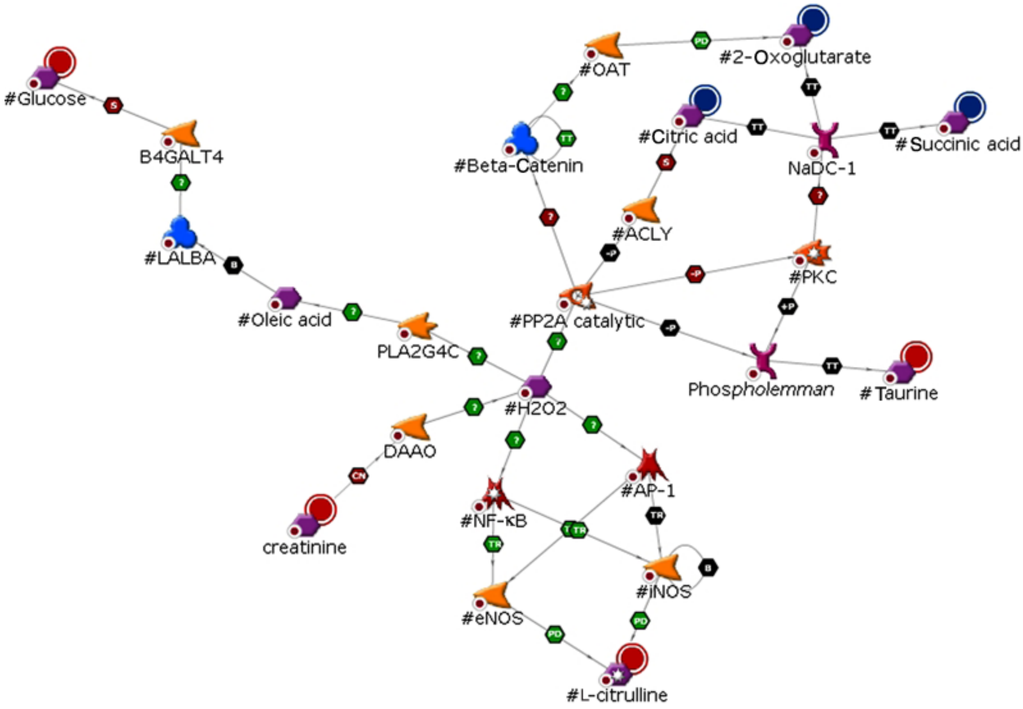


Fig. 7.

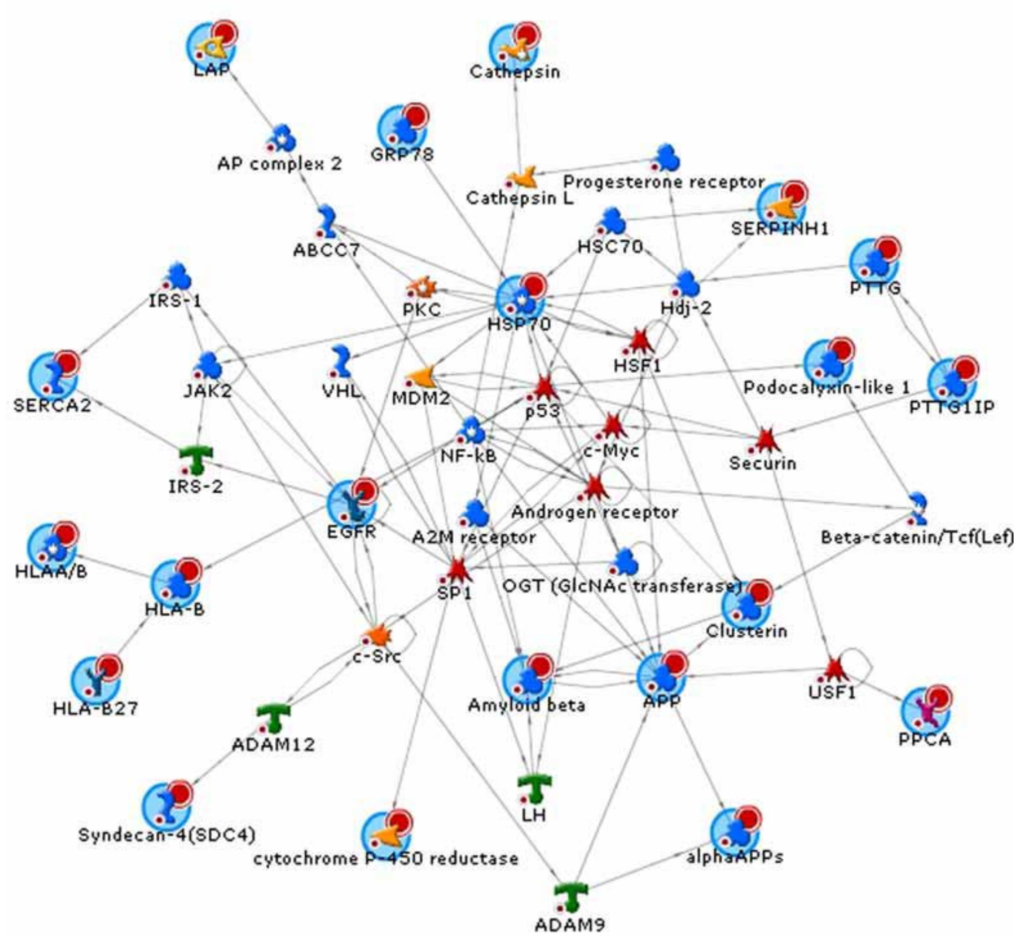


Fig. 8. MetaCore Analyze network analysis for cells treated with HIV-Tat ***(88)***. Genes upregulated at the G_1/S phase (*G*-score = 34.03, *p* = 1.56 e^{-38}), interactions hidden for clarity. Nodes surrounded by a blue circle indicate those from the input list that are upregulated. (Please *see* the companion CD for the color version of this figure.)

data set of genes in HeLa CD4+ cells that were significantly upregulated by HIV-1 Tat after self organizing map and *K*-means analysis ***(88)***. A list of upregulated genes at the G_1/S phase was analyzed with MetaCore using the Analyze Network algorithm (**Fig. 8**) and had statistics indicative of a significant network (*G*-score 34.03, *p* = 1.56 e^{-38}). The GO processes that mapped to this included response to unfolded protein (*p* = 1.01 e^{-7}), protection from natural killer cell mediated cytotoxicity (*p* = 1.17 e^{-7}), regulation of cell cycle (*p* = 2.35 e^{-7}), cell death (*p* = 4.88 e^{-7}), and antigen presentation, endogenous peptide antigen (*p* = 1.02 e^{-6}) which corresponds with Tat regulating the expression of a variety of GOs. A gene network like this might also suggest possible alternative targets for therapeutic intervention. There are several genes that are integral to the network but which do not appear to be upregulated, or alternatively this data was perhaps missing from the microarray and represents an area for further study.

Fig. 7. (*Opposite page*) Endogenous urinary metabolites observed in rat following hydrazine treatment ***(80)*** analyzed with MetaCore using the Analyze network algorithm (*G*-score = 46.31, *p* = 9.30 e^{-18}). Red or blue circles next to metabolites indicates the increase or decrease of the metabolites, respectively. (Please *see* the companion CD for the color version of this figure.)

4.3.2. Signature Networks for Radiosensitive Cervical Cancer Patients

There are many diseases for which microarray data sets have been generated with cells or human tissues. For example, there have been several studies that have assessed the effect of radiotherapy in cancer treatment, comparing individuals that were radiosensitive with those that were radioresistant at the level of gene expression, to derive a signature for successful response. Cervical cancer is a relatively common worldwide health concern for which human papillomaviruses are recognized as a causative agent with pleiotropic functions *(89)*. Currently, there is substantial treatment related morbidity and therefore new clinical options that modulate specific pathways to increase tumor cell death are urgently required *(90)*. Although the current treatment for cervical cancer is a combination of cisplatin chemotherapy and radiography, only the effect of radiography has been assessed at the level of gene expression to date. To our knowledge, there are at least three studies using small numbers of radiosensitive and radioresistant samples from patients with different radiotherapy protocols for cervical cancer that have undergone microarray analysis (**Table 1**). In each of these studies hierachical clustering was used to find differentially expressed genes. We have used these sets of 16–62 discriminating genes to build networks with MetaCore and assess whether additional information could be generated in this way as a preliminary step toward a signature network *(77)* for differentiating between radiosensitive and radioresistant cervical cancer. Fifteen of 35 genes were uploaded in MetaCore for one of the data sets *(91)* and was followed by use of the Analyze Network analysis to result in a statistically significant network (*G*-score = 37.7, $p = 4.4\ e^{-29}$, **Fig. 9A**). The following GO processes could be mapped on this network; phosphocreatine metabolism ($p = 6.06\ e^{-10}$), signal transduction ($p = 7.05\ e^{-07}$), positive regulation of interleukin-12 biosynthesis ($p = 8.49\ e^{-07}$), negative regulation of cell cycle ($p = 1.10\ e^{-06}$), and caspase activation through cytochrome c ($p = 1.31\ e^{-06}$). Eight of 16 genes were uploaded in MetaCore for a second data set *(92)* (that represents upregulated genes in radioresistant patients) followed by Analyze Network analysis (*G*-score 33.9, $p = 2.63\ e^{-20}$, **Fig. 9B**). The following GO processes could be mapped on this network; regulation of transcription, DNA-dependent ($p = 1.89\ e^{-04}$), protein transport ($p = 3.69\ e^{-04}$), traversing start control point of mitotic cell cycle ($p = 3.84\ e^{-04}$), intracellular protein transport ($p = 6.51\ e^{-04}$), and negative regulation of cell cycle ($p = 8.72\ e^{-04}$). Twenty-four of 62 genes were uploaded in MetaCore for the third data set *(93)* followed by Analyze Network analysis (*G*-score 37.5, $p = 2.04\ e^{-42}$, **Fig. 9C**). The following GO processes could be mapped on this network; regulation of transcription, DNA-dependent $p = 4.223\ e^{-11}$, signal transduction ($p = 7.05\ e^{-07}$), positive regulation of interleukin-12 biosynthesis ($p = 8.49\ e^{-07}$), transcription from Pol II promoter ($p = 2.28\ e^{-06}$) and positive regulation of transcription, DNA-dependent ($p = 2.71\ e^{-06}$). Although, there is no apparent overlap in the gene lists between the two most comparable studies for the first *(92)* and third *(93)* data sets, the resulting two independent gene networks indicate common GO processes for signal transduction and positive regulation of interleukin-12 biosynthesis. Ideally it would have been advantageous to use the complete starting gene lists for the thousands of cDNAs used in the three microarray studies rather than using clustering first, as we have previously demonstrated *(77)*, but none of these were available. This limited our analysis to using the gene lists after hierachical clustering with all of the disadvantages implicit in this approach. We await the publication of gene expression studies from patients that have received combination therapy using cisplatin and radiotherapy to assess whether the gene networks are similar or different to those described earlier.

4.4. Mapping Proteomics Data

Currently there appear to be far fewer examples of proteomics studies. There are even fewer studies that combine multiple high-throughput data types such as proteomics and genomics analysis *(74)*. In these examples there might be some, little or no correlation between gene-expression and protein levels. Proteomics data generated by a number of techniques *(19)* can be visualized in MetaCore by uploading SwissProt identifiers, for example. A recent publication

Table 1
The Effect of Radiotherapy in Cervical Cancer Genomics Studies

Total number of patients	Treatment regimen	Gene chip	Number of genes after analysis	Reference
19 (8 radiosensitive, 11 radioresistant)	1.8 Gy, 5d/wk 30.6 Gy total to whole pelvis + 60 min hyperthermia using 8 MHz radiofrequency electromagnetic waves + an additional dose to 52,2 Gy along with 192 Ir intracavity brachytherapy given at 7.5 Gy once/wk	Custom chip with 23,040 cDNA clones	After hierachical clustering 156 genes differentially expressed, narrowed down to 35 discriminating genes	*91*
19 (9 radiosensitive, 10 radioresistant)	1.8 Gy, 5d/wk 30.6 Gy total to whole pelvis + an additional dose to 52,2 Gy along with 192 Ir intracavity brachytherapy given at 7.5 Gy once/wk	Custom chip with 23,040 cDNA clones	After hierarchical clustering 171 genes were differentially expressed, narrowed down to discriminating 62 genes	*93*
13 (radiosensitive 7, radioresistant 6)	Not described	Custom chips with 10,692 cDNAs	After hierachical clustering 300 genes were differentially expressed, 16 upregulated genes in radioresistant samples are described	*92*

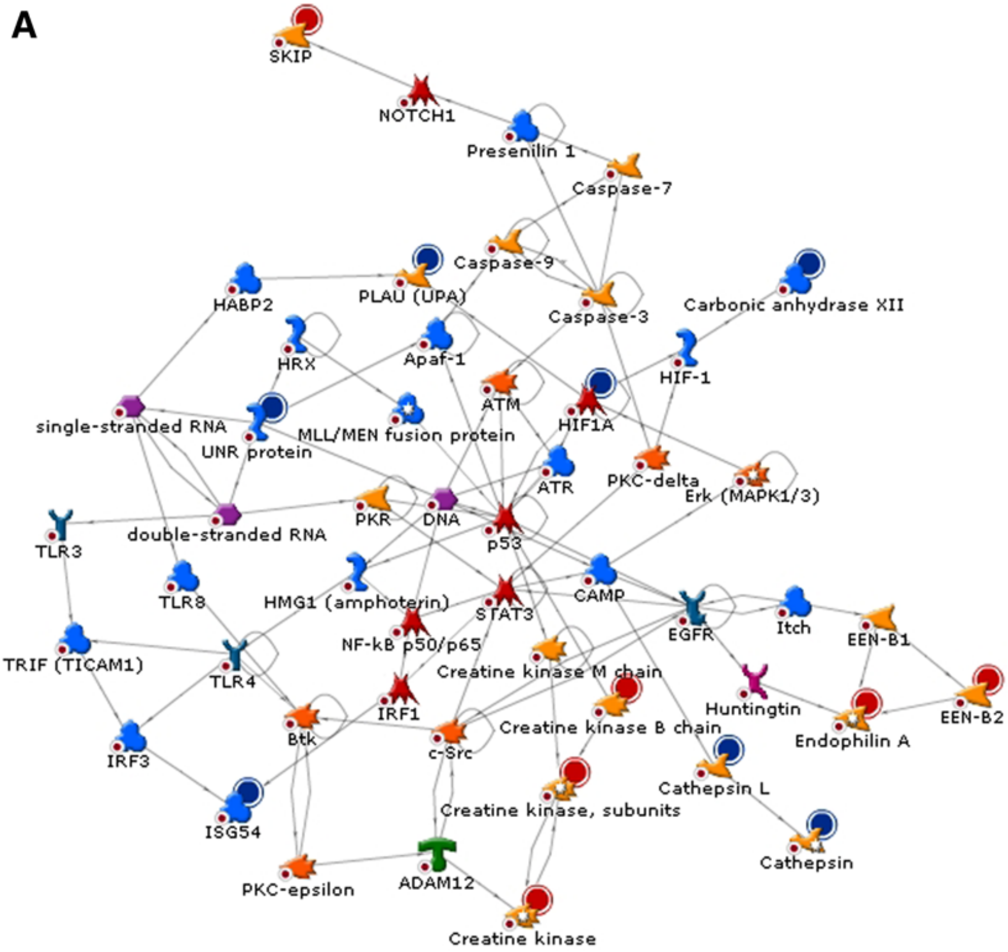


Fig. 9. **(A)** Differentiating radioresistant and radiosensitive patients with cervical cancer. MetaCore Analyze network analysis (G-score = 37.7, p = 4.4 e^{-29}) for discriminating genes from Harima et al. *(91)*, interactions hidden for clarity. **(B)** MetaCore Analyze network analysis (G-score 33.9, p = 2.63 e^{-20}) for initial set of 16 upregulated genes from Wong et al. *(92)*. **(C)** MetaCore Analyze network analysis (G-score 37.5, p = 2.04 e^{-42}) for descriminating genes from Kitahara et al. *(93)* interactions hidden for clarity. Red or blue circles next to genes indicate up or down regulation, respectively. (Please *see* the companion CD for the color version of this figure.)

described the proteins and genes modulated by the human papillomavirus 16 E7 oncogene in a cervical cancer cell line (C33A) *(89)*. Matrix-assisted laser desorption/ionization-time of flight mass spectrometry and microarrays (cytokine and apoptosis) were used to derive the protein and gene data, respectively. We have used MetaCore to analyze the data generated. We were able to recognize 25 of 50 proteins on networks and used the Analyze network algorithm to produce a network (G-score 24.49, p = 1.58 e^{-31}, **Fig. 10A**). This network contained transcriptional factors E2F, p53, and retinoblastoma protein, which have tumor suppressor functions and although these proteins did not increase or decrease in expression, connected downstream proteins did change their expression. These downstream effects are recognized by the following GO processes, which could be mapped on this network; transcription (p = 1.83 e^{-09}), regulation of transcription, DNA-dependent (p = 7.67 e^{-09}), regulation of cell cycle (p = 2.86 e^{-07}), transcription, DNA-dependent (p = 1.64 e^{-06}), and negative regulation of cell cycle (p = 4.02 e^{-06}). Hence, this network shows the likely pattern of disruption of normal physiological function that is

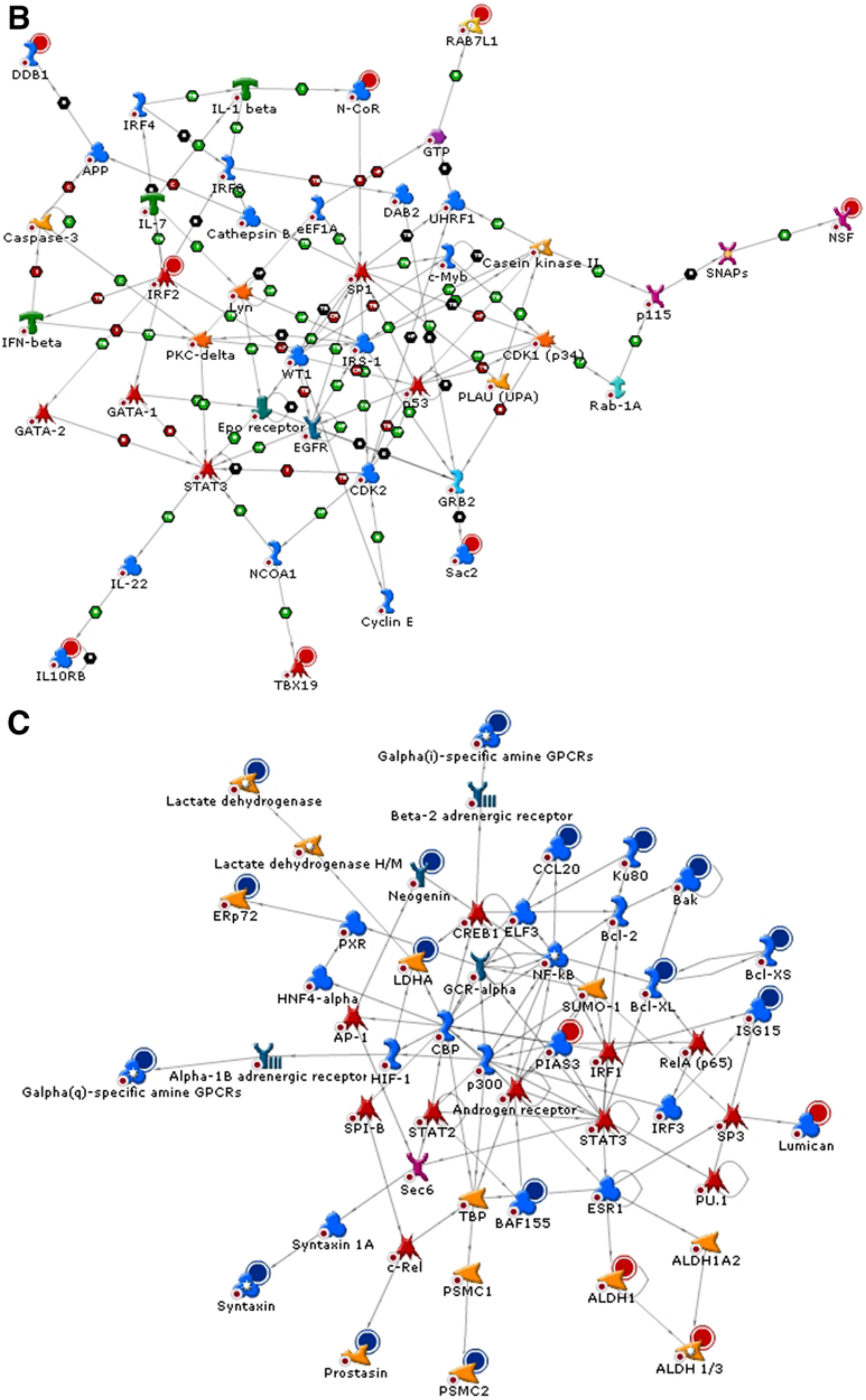


Fig. 9. *(Continued)*

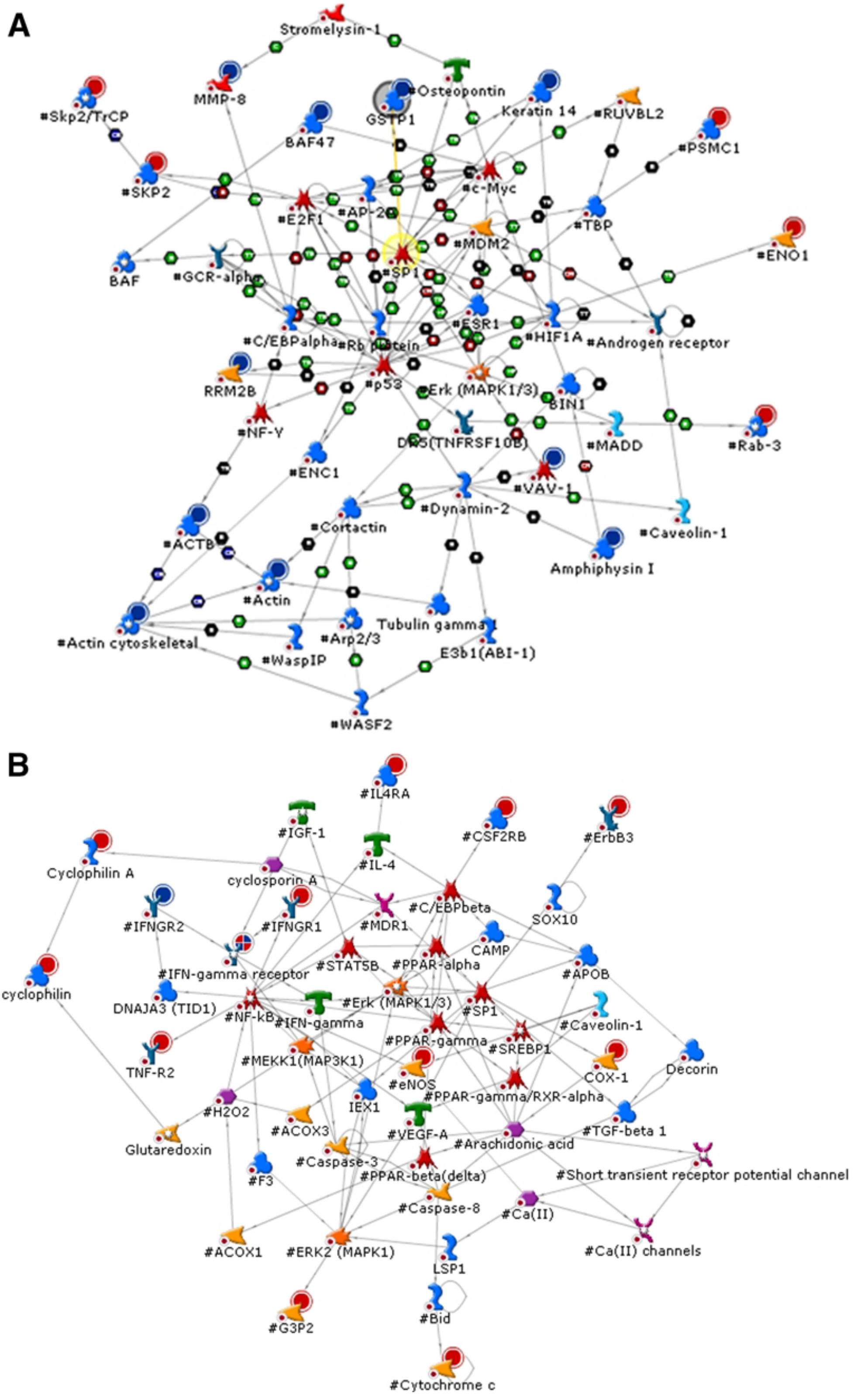

A
Stromelysin-1
MMP-8
#Osteopontin
GSTP1
Keratin 14
#RUVBL2
#Skp2/TrCP
BAF47
#PSMC1
#SKP2
#c-Myc
#AP-2
#E2F1
#TBP
#MDM2
#ENO1
BAF
#GCR-alpha
#SP1
#ESR1
#HIF1A
#Androgen receptor
#C/EBPalpha
#Rb protein
RRM2B
#p53
Erk (MAPK1/3)
BIN1
#NF-Y
DR5(TNFRSF10B)
#MADD
#Rab-3
#ENC1
#VAV-1
#Dynamin-2
#Caveolin-1
#Cortactin
#ACTB
Amphiphysin I
#Actin
Tubulin gamma 1
#Actin cytoskeletal
#Arp2/3
#WaspIP
E3b1(ABI-1)
#WASF2
B
#IL4RA
#IGF-1
#CSF2RB
#ErbB3
#IL-4
Cyclophilin A
cyclosporin A
#C/EBPbeta
SOX10
#IFNGR2
#IFNGR1
#MDR1
CAMP
#STAT5B
#PPAR-alpha
#APOB
#IFN-gamma receptor
cyclophilin
DNAJA3 (TID1)
#NF-kB
#IFN-gamma
#Erk (MAPK1/3)
#SP1
#Caveolin-1
#PPAR-gamma
#SREBP1
TNF-R2
#MEKK1(MAP3K1)
COX-1
Decorin
#eNOS
#PPAR-gamma/RXR-alpha
IEX1
#H2O2
#ACOX3
#TGF-beta 1
Glutaredoxin
#VEGF-A
#Arachidonic acid
#Caspase-3
#F3
#PPAR-beta(delta)
#Short transient receptor potential channel
#Caspase-8
#Ca(II)
#ACOX1
#ERK2 (MAPK1)
#Ca(II) channels
LSP1
#Bid
#G3P2
#Cytochrome c

Fig. 10.

achieved by the human papillomavirus 16 E7 oncogene. Fourteen of the 18 genes that were differentially expressed were able to map to networks in MetaCore using the Analyze network algorithm (*G*-score 35.69, $p = 1.48\ e^{-30}$, **Fig. 10B**). The following GO processes could be mapped on this network; apoptosis ($p = 1.964\ e^{-07}$), lipid metabolism ($p = 2.22\ e^{-07}$), response to pathogenic bacteria ($p = 2.64\ e^{-07}$), signal transduction ($p = 3.14\ e^{-06}$), and cholesterol metabolism ($p = 7.25\ e^{-06}$). Although, we have not been able to map all of the proteins and genes that were published to networks, interestingly there is no overlap between the gene and protein data and consequently the GO processes are different in both cases. The use of more expansive gene chips than the limited cytokine and apoptosis arrays would perhaps provide further insights into the complete picture of gene expression after exposure to the human papillomavirus 16 E7 oncogene than gathered here and perhaps provide some areas of overlap that are not presently apparent. Interestingly, this genomic and proteomic data from a cell line exposed to the human papillomavirus 16 E7 oncogene does have some GO processes in common with the gene networks that differentiate radiosensitive and radioresistant responses for cervical cancer (as described earlier), although the networks themselves appear to be different (**Subheading 4.3.2.**).

5. Discussion

Network analysis of experimental high-throughput data is a novel field and as such there are limitations and drawbacks. For one, the network validity depends on the interaction data source. The reliability of individual interactions, which represent the building blocks of networks is a fundamental prerequisite; otherwise errors multiply in multistep pathways and modules. To our knowledge, there have been few if any published comparisons of commercial databases or for that matter other manually curated databases. Another, as yet unsolved issue is the proper resolution of protein complexes, families and the appropriate representation of these protein groups on the networks. To date there has been no standardization of this vocabulary. There is also a perceived problem of compatibility between different types of protein interactions with metabolic pathways typically represented as linear and branched chains of consecutive reactions, which is in contrast to more complex signaling pathways.

We have described the MetaCore and MetaDrug databases, which address some of these issues and were designed and implemented with novel database architectures to allow the organization of relevant biological and chemical information around the molecular entities, genes, proteins, transcripts, and compounds by connecting them through functional processes: reactions, pathways, and ultimately networks. Additionally, the user can filter the networks by removing an object, filter by type of interaction, or by tissue type and the uploaded data can be filtered based on the desired fold change threshold, and so on.

The future development of methods to identify which gene networks are the most useful are likely to be important as are methods that highlight the critical genes for a disease or biological process. To date we have provided *Z*-score, *G*-score, and *p*-values to assess networks as described earlier, however, there might be many networks that are statistically significant and under these circumstances the user might have several choices of possible networks that are biologically reasonable. There has been a published example for the prioritization of gene candidates using molecular triangulation and this or a similar method might be considered for implementation in the types of software described previously *(64)*. Statistical tests in MetaCore might be applicable for two correlated but distinct tasks associated with functional data mining of large, high-throughput data sets. First, the "genome wide" data sets such as tens of thousands of differentially expressed genes connected in thousands of pathways are often too complex for functional analysis. Such data sets have to be reduced to small sets of several dozen genes that are most relevant to the

Fig. 10. (*Opposite page*) **(A)** Analysis of proteomics data for the E7 oncogene *(89)*, **(B)** Analysis of genomics data for the E7 oncogene, interactions hidden for clarity. Red or blue circles next to genes indicate up or down regulation, respectively. (Please *see* the companion CD for the color version of this figure.)

condition and can still be experimentally verifiable. The *p*-value, *Z*-score, and *G*-score procedures are designed exactly for this purpose. The networks built from an unabridged data set can be scored based on the relative saturation of their interaction space with experimental data (for instance, gene expression, protein abundance) using these *p*-values and *G*-scores. The higher scored networks can be considered as more relevant to the data set, and therefore, can be considered as a higher priority for further research. Such an approach is essentially different to standard clustering methods applied in high-throughput data analysis. Second, different categories of functional analysis (canonical pathways, GOs, signaling, and metabolic networks) can also be scored and prioritized relative to each other, which substantially enhances the flexibility of the analysis. For instance, a user can choose to start analysis by parsing the whole data set onto GO processes, applying *p*-value calculations; then build networks specific for the highest scored GO processes. Pathway maps scored in a similar way is another entry point for specific high-resolution network analysis. Both GO processes and pathways can, therefore, be aligned relative to each other (**Fig. 2**).

We have provided several examples of HCS data and how this might be analyzed on networks. A further recent alternative example derived multiparameter data for single human immune system cells using flow cytometry and produced causal protein signaling networks using a Bayesian network inference algorithm ***(94)***. A broader view of the use of Bayesian networks and other probabilistic graphical models suggests that their application could help with the data explosion we are seeing in biology ***(95)***. In contrast an algorithm for the reconstruction of accurate cellular networks (ARACNe) was recently described and used to reconstruct expression profiles of human B cells. ARACNe identified statistically significant gene–gene coregulation and eliminated indirect interactions. Using 336 expression profiles after perturbing B-cell phenotypes, a network was inferred. *MYC* appeared in the top 5% of cellular hubs and the network consisted of 40% of previously identified target genes ***(96)***. ARACNe was compared to Bayesian networks and found to be comparable. Such alternative algorithm approaches might be valuable for rapidly inferring dynamic networks and represent a useful adjunct to the database approaches we have described.

In summary, we have described various tools for the network analysis of different data types including high content and high-throughput data. We have provided several example applications of MetaCore and MetaDrug to the analysis of these different data types. In addition, we have also described multiple procedures for the evaluation of the statistical significance of the networks that are generated, linking the data to cellular processes and prioritization of the individual node networks with respect to relative saturation with experimental data. We believe these are universal approaches that are applicable to the network analysis of multiple data types in human and other eukaryotes. Network analysis of high-throughput and high-content data represents one of the first truly systems biology methods in the sense of representation and interpretation of the complete functional content of a cell, a tissue, and an organism. We believe the methods described in this chapter will therefore, have multiple applications in drug discovery and life science research in general ***(19)*** as part of a defined process with other computational tools to increase the chances of success for new pharmaceuticals ***(97)***.

Acknowledgments

We gratefully acknowledge our colleagues and the many scientists, who have provided constructive feedback for MetaCore and MetaDrug software development. The development of MetaCore was partially funded by a 2001 Michigan Life Science Corridor research grant and MetaDrug was partially funded by NIH NIGMS SBIR 2R44GM069124-02A1.

References

1. Nicholson, J. K. and Wilson, I. D. (2003) Understanding ‘global’ systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* **2,** 668–676.

2. Laghaee, A., Malcolm, C., Hallam, J., and Ghazal, P. (2005) Artificial intelligence and robotics in high-throughput post-genomics. *Drug Discov. Today* **10,** 1253–1259.
3. Hood, L. (2003) Systems biology: integrating technology, biology, and computation. *Mech. Ageing Dev.* **124,** 9–16.
4. Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999) From molecular to modular cell biology. *Nature* **402,** C47–C52.
5. Peri, S., Navarro, J. D., Amanchy, R., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13,** 2363–2371.
6. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95,** 14,863–14,868.
7. Butte, A. (2002) The use and analysis of microarray data. *Nat. Rev. Drug Discov.* **1,** 951–960.
8. Abraham, V. C., Taylor, D. L., and Haskins, J. R. (2004) High content screening applied to large-scale cell biology. *Trends Biotechnol.* **22,** 15–22.
9. Giuliano, K. A. and Taylor, D. L. (1998) Fluorescent-protein biosensors: new tools for drug discovery. *Trends Biotechnol.* **16,** 135–140.
10. Taylor, D. L., Woo, E. S., and Giuliano, K. A. (2001) Real-time molecular and cellular analysis: the new frontier of drug discovery. *Curr. Opin. Biotechnol.* **12,** 75–81.
11. Mitchison, T. J. (2005) Small-molecule screening and profiling by using automated microscopy. *Chembiochem* **6,** 33–39.
12. Vogt, A., Tamewitz, A., Skoko, J., Sikorski, R. P., Giuliano, K. A., and Lazo, J. S. (2005) The benzo[c]phenanthridine alkaloid, sanguinarine, is a selective, cell-active inhibitor of mitogen-activated protein kinase phosphatase-1. *J. Biol. Chem.* **280,** 19,078–19,086.
13. Mousses, S., Caplen, N. J., Cornelison, R., et al. (2003) RNAi microarray analysis in cultured mammalian cells. *Genome Res.* **13,** 2341–2347.
14. Tencza, S. B. and Sipe, M. A. (2004) Detection and classification of threat agents via high-content assays of mammalian cells. *J. Appl. Toxicol.* **24,** 371–377.
15. Vogt, A., Cooley, K. A., Brisson, M., Tarpley, M. G., Wipf, P., and Lazo, J. S. (2003) Cell-active dual specificity phosphatase inhibitors identified by high content screening. *Chem. Biol.* **10,** 733–742.
16. Simpson, P. B., Bacha, J. I., Palfreyman, E. L., Woollacott, A. J., McKernan, R. M., and Kerby, J. (2001) Retinoic acid evoked-differentiation of neuroblastoma cells predominates over growth factor stimulation: an automated image capture and quantitation approach to neuritogenesis. *Anal. Biochem.* **298,** 163–169.
17. Borchert, K. M., Galvin, R. J., Frolik, C. A., et al. (2005) High content screening assay for activators of the Wnt/Fzd pathway in primary human cells. *Assay Drug Dev. Technol.* **3,** 133–141.
18. Giuliano, K. A., Cheung, W. S., Curran, D. P., et al. (2005) Systems cell biology knowledge created from high content screening. *Assay Drug Dev. Tech.* **3,** 501–514.
19. Ekins, S., Bugrim, A., Nikolsky, Y., and Nikolskaya, T. (2005) Systems biology: applications in drug discovery. In *Drug Discovery Handbook* (Gad, S. C., ed.), Wiley, New York, pp. 123–183.
20. Ekins, S., Kirillov, E., Rakhmatulin, E., and Nikolskaya, T. (2005) A novel method for visualizing nuclear hormone receptor networks relevant to drug metabolism. *Drug Metab. Dispos.* **33,** 474–481.
21. Chaussabel, D. (2004) Biomedical literature mining: challenges and solutions in the ‘omics’ era. *Am. J. Pharmacogenomics* **4,** 383–393.
22. Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling. *Genome Biol.* **3,** 0055.1–0055.16.
23. Chen, H. and Sharp, B. M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5,** 147.
24. Blaschke, K. and Valencia, A. (2001) Can bibliographical pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Funct. Genomics* **2,** 196–206.
25. Santos, C., Eggle, D., and States, D. J. (2005) Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics* **21,** 1653–1658.
26. Grigorov, M. G. (2005) Global properties of biological networks. *Drug Discov. Today* **10,** 365–372.
27. Daraselia, N., Yuryev, A., Egorov, S., Novihkova, S., Nikitin, A., and Mazo, I. (2004) Extracting human protein interactions from Medline using a full-sentence parser. *Bioinformatics* **20,** 604–611.
28. Chien, C. T., Bartel, P. L., Sternglanz, R., and Fields, S. (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* **88,** 9578–9582.

29. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574.
30. Giot, L., Bader, J. S., Brouwer, C., et al. (2003) A protein interaction map of Drosophila melanogaster. *Science* **302,** 1727–1736.
31. Li, S., Armstrong, C. M., Bertin, N., et al. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* **303,** 540–543.
32. Mrowka, R., Patzak, A., and Herzel, H. (2001) Is there a bias in proteome research? *Genome Res.* **11,** 1971–1973.
33. Lehner, B. and Fraser, A. G. (2004) A first-draft human protein-interaction map. *Genome Biol.* **5,** R63.
34. Pagel, P., Kovac, S., Oesterheld, M., et al. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21,** 832–834.
35. Hughes, T. R., Marton, M. J., Jones, A. R., et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126.
36. van Noort, V., Snel, B., and Huynen, M. A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.* **19,** 238–242.
37. Kemmeren, P., van Berkum, N. L., Vilo, J., et al. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9,** 1133–1143.
38. Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell Proteomics* **1,** 349–356.
39. Ho, Y., Gruhler, A., Heilbut, A., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180–183.
40. Gavin, A. C., Bosche, M., Krause, R., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147.
41. Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.* **13,** 377–382.
42. Navarro, J. D. and Pandey, A. (2004) Unraveling the human interactome: lessons from the yeast. *Drug Discov. Today: Targets* **3,** 79–84.
43. Nikolsky, Y., Nikolskaya, T., and Bugrim, A. (2005) Biological networks and analysis of experimental data in drug discovery. *Drug Discov. Today* **10,** 653–662.
44. Peri, S., Navarro, J. D., Kristiansen, T. Z., et al. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* **32** (Database issue)**,** D497–D501.
45. Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22,** 78–85.
46. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.* **513,** 135–140.
47. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31,** 258–261.
48. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29.
49. Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000) Protein function in the post-genomic era. *Nature* **405,** 823–826.
50. Hood, L. and Perlmutter, R. M. (2004) The impact of systems approaches on biological problems in drug discovery. *Nat. Biotechnol.* **22,** 1215–1217.
51. Yu, H., Zhu, X., Greenbaum, D., Karro, J., and Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* **32,** 328–337.
52. Barabasi, A. -L. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–113.
53. Albert, R., Jeong, H., and Barabasi, A. L. (2000) Error and attack tolerance of complex networks. *Nature* **406,** 378–382.
54. Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4,** R7.
55. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411,** 41–42.

56. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., and Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet.* **20,** 227–231.
57. Han, J. D., Bertin, N., Hao, T., et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430,** 88–93.
58. Sharom, J. R., Bellows, D. S., and Tyers, M. (2004) From large networks to small molecules. *Curr. Opin. Chem. Biol.* **8,** 81–90.
59. Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5,** 262–275.
60. Csermely, P., Agoston, V., and Pongor, S. (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.* **26,** 178–182.
61. Parsons, A. B., Brost, R. I., Ding, H., et al. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.* **22,** 62–69.
62. Thornton-Wells, T. A., Moore, J. H., and Haines, J. L. (2004) Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.* **20,** 640–647.
63. Moore, J. H. (2005) A global view of epistasis. *Nat. Genet.* **37,** 13–14.
64. Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. USA* **101,** 15,148–15,153.
65. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14,** 283–291.
66. Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–255.
67. Wuchty, S. (2002) Interaction and domain networks of yeast. *Proteomics* **2,** 1715–1723.
68. Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14,** 292–299.
69. Aloy, P., Bottcher, B., Ceulemans, H., et al. (2004) Structure-based assembly of protein complexes in yeast. *Science* **303,** 2026–2029.
70. Ng, W. -L., Kazmierczak, K. M., Robertson, G. T., Gilmour, R., and Winkler, M. E. (2003) Transcriptional regulation and signature patterns revealed by microarray analyses of streptococcus pneumoniae R6 challenged with sublethal concentrations of translation inhibitors. *J. Bacteriol.* **185,** 359–370.
71. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000) The large-scale organization of metabolic networks. *Nature* **407,** 651–654.
72. Schilling, C. H. and Palsson, B. O. (1998) The underlying pathway structure of biochemical reaction networks. *Proc. Natl. Acad. Sci. USA* **95,** 4193–4198.
73. Ekins, S., Giroux, C. N., Nikolsky, Y., Bugrim, A., and Nikolskaya, T. (2005) A signature gene network approach to toxicity. *Toxicologist* **84,** meeting abstract.
74. Ekins, S., Nikolsky, Y., and Nikolskaya, T. (2005) Techniques: application of systems biology to absorption, distribution, metabolism, excretion, and toxicity. *Trends Pharmacol. Sci.* **26,** 202–209.
75. Korolev, D., Balakin, K. V., Nikolsky, Y., et al. (2003) Modeling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* **46,** 3631–3643.
76. Ekins, S., Andreyev, S., Ryabov, A., et al. (2005) Computational prediction of human drug metabolism. *Exp. Opin. Drug Metab. Toxicol.* **1,** 303–324.
77. Nikolsky, Y., Ekins, S., Nikolskaya, T., and Bugrim, A. (2005) A novel method for generation of signature networks as biomarkers from complex high-throughput data. *Toxicol. Lett.* **158,** 20–29.
78. Hodges, L. C., Cook, J. D., Lobenhofer, E. K., et al. (2003) Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells. *Mol. Cancer Res.* **1,** 300–311.
79. Hopkins, A. L. and Groom, C. R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* **1,** 727–730.
80. Bollard, M. E., Keun, H. C., Beckonert, O., et al. (2005) Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicol. Appl. Pharmacol.* **204,** 135–151.
81. Lu, B., Soreghan, B. A., Thomas, S. N., Chen, T., and Yang, A. J. (2005) *ACS*, San Diego. Meeting abstract.
82. Waters, K. M., Shankaran, H., Wiley, H. S., Resat, H., and Thrall, B. D. (2005) *Keystone Symposia*.
83. Lantz, R. C., Petrick, J. S., and Hays, A. M. (2005) *Society of Toxicology*. Meeting abstract.
84. Nie, A. Y., McMillian, M. K., Leone, A. M., et al. (2005) *Society of Toxicology*. Meeting abstract.

85. Meng, Q., Waters, K. M., Malard, J. M., Lee, K. M., and Pounds, J. G. (2005) *Society of Toxicology*. Meeting abstract.
86. Mason, C. W., Swaan, P. W., and Weiner, C. P. (2005) *Society of Gynecological Investigation*, Los Angeles, CA. Meeting abstract.
87. Weiner, C. P., Mason, C. W., Buhimschi, C., Hall, G., Swaan, P. W., and Buhimschi, I. (2005) *Society of Gynecological Investigation*, Los Angeles, CA. Meeting abstract.
88. Liang, W. S., Maddukuri, A., Teslovich, T. M., et al. (2005) Therapeutic targets for HIV-1 infection in the host proteome. *Retrovirology* **2,** 20.
89. Lee, K. A., Shim, J. H., Kho, C. W., et al. (2004) Protein profiling and identification of modulators regulated by the E7 oncogene in the C33A cell line by proteomics and genomics. *Proteomics* **4,** 839–848.
90. Hougardy, B. M., Maduro, J. H., van der Zee, A. G., Willemse, P. H., de Jong, S., and de Vries, E. G. (2005) Clinical potential of inhibitors of survival pathways and activators of apoptotic pathways in treatment of cervical cancer: changing the apoptotic balance. *Lancet Oncol.* **6,** 589–598.
91. Harima, Y., Togashi, A., Horikoshi, K., et al. (2004) Prediction of outcome of advanced cervical cancer to thermoradiotherapy according to expression profiles of 35 genes selected by cDNA microarray analysis. *Int. J. Radiat. Oncol. Biol. Phys.* **60,** 237–248.
92. Wong, Y. F., Selvanayagam, Z. E., Wei, N., et al. (2003) Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. *Clin. Cancer Res.* **9,** 5486–5492.
93. Kitahara, O., Katagiri, T., Tsunoda, T., Harima, Y., and Nakamura, Y. (2002). Classification of sensitivity or resistance of cervical cancers to ionizing radiation according to expression profiles of 62 genes selected by cDNA microarray analysis. *Neoplasia* **4,** 295–303.
94. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308,** 523–529.
95. Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science* **303,** 799–805.
96. Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37,** 382–390.
97. Swaan, P. W. and Ekins, S. (2005) Reengineering the pharmaceutical industry by crash-testing molecules. *Drug Discov. Today* **10,** 1191–1200.
98. Burk, O., Arnold, K. A., Nussler, A. K., et al. (2005) Antimalarial artemisinin drugs induce cytochrome P450 and MDR1 expression by activation of xenosensors pregnane X receptor and constitutive androstane receptor. *Mol. Pharmacol.* **67,** 1954–1965.

Reference Added in Proof

Ekins, S., Andreyev, S., Ryabov, A., et al. (2006) A combined approach to drug metabolism and toxicity assessment. *Drug Metabolism and Disposition*, in press.