4

Data Integration: Towards Understanding Biological Complexity

David Gomez-Cabrero and Jesper Tegner

Unit of Computational Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden

Systems biology will fully develop its potential when researchers are able to make use of all the available data. To fulfill this goal it is needed to overcome two major challenges: (i) how to store and make data and knowledge accessible; and (ii) how to integrate and analyze data from different sources and of different types. Hence, this chapter is divided into two sections. The first section describes public storage resources such as experimental data repositories, ontologies and knowledge databases; however we will not discuss the (relevant topic of) requirements for a data-warehouse capable of managing several databases in an interoperable manner (Kimball and Ross 2002). The second section reviews available tools and algorithms that are useful for integrating different types of data sets. However, since a comprehensive enumeration is beyond the scope of this chapter, we present some representative examples.

In this chapter we list integrative tools using open source codes and public repositories and databases (see Table 4.1). The relevant R packages (Ihaka and Gentleman 1996) are highlighted in bold (see Table 4.2).

4.1 Storing knowledge: Experimental data, knowledge databases, ontologies and annotation

Quantitative measurements of biological entities are data; organized data (e.g. by relational connections) becomes Information, and we gain Knowledge by the appropriate collection and analysis of Information. Biological research needs to find ways to store efficiently Data, Information and Knowledge. The term *efficient* denotes 'ways to allow integration among the different data types'.

Handbook of Statistical Systems Biology, First Edition. Edited by Michael P. H. Stumpf, David J. Balding and Mark Girolami.
© 2011 John Wiley & Sons, Ltd. Published 2011 by John Wiley & Sons, Ltd.

Experimental data are stored in data repositories. Information is stored in *Information Resources* such as Knowledge Databases and Ontologies. This section will describe each one of those terms, key examples of each one and the process of annotation (that acts as the bridge between experimental data and Information).

In this section we use protein p53 to provide relevant examples. p53 is a tumor repressor and a nuclear transcription factor that accumulates in response to cellular stress, including DNA damage and oncogene activation; therefore it guards the genome stability and normal cell growth. It was discovered in 1979 and it was then considered an oncogene (DeLeo *et al.* 1979; Lane and Crawford 1979; Linzer and Levine 1979), but it did not get the real attention of the research community until several years later when it was classified as a tumor suppressor that is mutated in most cancer cells (Jay *et al.* 1981; Mowat *et al.* 1985; Baker *et al.* 1989). The importance of this protein is highlighted by the 54 530 entries in PubMed (http://www.ncbi.nlm.nih.gov/pubmed, searching for 'p53', August 2010) where 6857 entries are reviews.

4.1.1 Data repositories

Over the years, the necessity for efficiently storing data has become clear. Sequence analysis provides us with a clarifying example: initially all new sequences obtained were submitted to journals; therefore it complicated its (efficient) use by researchers. To deal with this problem a database of sequences (EMBL Nucleotide Sequence Data Library), where all researchers would be able to submit their own sequences, was established at the European Molecular Biology Laboratory (EMBL); it was the first sequence database and its importance and utility became rapidly clear. Other initiatives, such as GenBank founded in 1982 and later absorbed into the National Center for Biotechnology Information (NCBI), followed and extended this pioneer experience.

These first initiatives have been growing more rapidly in requirements over recent years: from microarrays to RNA-Seq, the amount of data generated daily is overwhelming and it clearly points out the necessity of storage and analytical tools in order to use the data efficiently. Several efforts have been made over the years to provide ordered access to all public available data so that comparative studies can be performed. A relevant example is found in microarrays; two of the most relevant microarray repositories are: (i) ArrayExpress (Parkinson *et al.* 2007; http://www.ebi.ac.uk/arrayexpress) developed by European Bioinformatics Institute (EBI); and (ii) Gene Expression Omnibus (GEO) developed by National Center for Biotechnology Information (NCBI) (Edgar *et al.* 2002; Barrett *et al.* 2007; http://www.ncbi.nlm.nih.gov/geo). At GEO the basic repository unit is a sample record which describes condition, manipulation and abundance measurement of each element derived from it; samples can be grouped in Series by submitters and/or Data Sets by GEO curators. Both repositories are regularly updated.

A query of 'p53' within ArrayExpress (August 2010) returns 142 different experiments and 5948 microarray assays. For each experiment listed, data are available in a Raw format (usually CEL files) and/or in Processed format; however it is recommended to upload data in Raw format, as new (and hopefully better) methods to pre-process the data are continuously being developed. Each experiment has a unique identifier. For instance, E-GEOD-10795 is an identifier for an experiment that aims to elucidate the impact of TAF6d on cell death and gene expression; we also observe related information such as links to publications (see Lesne and Benecke 2008; Wilhelm *et al.* 2008), files (e.g. data archives, experiment design and array design), a description of the experiment and links to other databases. Among the links there is a reference to the same data set available at GEO where its identifier is GSE10795; following this link we observe similar information to the information found in ArrayExpress. A p53 query in GEO search returns 31 data sets and 257 series.

Repositories can store different types of data in different formats for the same defined measure, or different types of quantitative data for the same system. For example microarrays can be used to measure as many different things as (i) mRNA expression, (ii) exon differential expression (to compare isoforms) (Clark *et al.* 2007), (iii) as part of the ChIP-chip assay (Horak and Snyder 2002) and (iv) SNPs (Hacia *et al.* 1999). The need to integrate these different data sets highlights the need for standards; as an example, the Microarray

Gene Expression Data (MGED) Society defined the Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.* 2001) that corresponds to the minimum information that must be reported about a microarray experiment to enable its unambiguous interpretation and reproduction.

There are also repositories specific for certain diseases or biological systems: RefDIC (Hijikata *et al.* 2007) is a public compendium of quantitative mRNA/protein profile data obtained from microarray and two-dimensional gel electrophoresis based proteome experiments specifically for immune cells; Oncomine initiative collects and standardizes all published (and publicly available) cancer microarray data (http://www.oncomine.org), as described in Rhodes *et al.* (2007). There are also repositories for most biological data types. For instance Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) includes mass spectrometry experiments.

Data repositories are dealing continuously with new challenges. For instance, as High Throughput Sequencing (HTS) techniques provide tools that will replace microarrays (e.g. RNA-Seq provides a more precise assessment of differential mRNA expression), it highlights the necessity of new data types and new storage structures. Another necessary characteristic of repositories is 'updatability': *repositories should be able to accommodate new types of data, and link them to the previous ones if possible* as it will allow the integration of old and new data types.

4.1.2 Knowledge Databases

Experimental data need to be transformed into information. This information needs to be stored in an ordered way in (public) databases, named Knowledge Databases (KDs). Researchers use KDs to test their experimental results or to validate their hypotheses. Because KDs present the processed data extracted from experiments and are specific for different types of entities (e.g. genes, proteins or diseases) or even variations of the same topic (e.g. DNA sequence: transcripts or SNPs) they need specific structures for each case. However, relevant characteristics (and challenges) that all KDs share are:

1. Interconnection: every KD is providing a narrow view of a certain topic; to provide a better integrative view interconnection between KDs is needed.
2. Public access: all information in major KDs must be publicly available. this policy is enforced by entities such as International Nucleotide Sequence Database Collaboration (INSD, http://www.insdc.org/policy.html): '...no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party...'; 'All database records submitted to the INSD will remain permanently accessible as part of the scientific record'.
3. Updatability: KDs must remain flexible enough in order to be able to include new data types.
4. Standardization: in order to fulfill Interconnection and Updatability challenges it is necessary to define information storage standards.

KDs represent the state-of-knowledge in biology. However KDs are based on *current* knowledge, with different degrees of certainty.

4.1.2.1 p53 KD tour

As there is not space to enumerate all KDs we present the most important ones in Table 4.1. However, we find it useful to provide an example: we describe several p53 queries among different KDs. We begin by searching p53 at the NCBI, which allows the user to search over different databases. A search of 'p53 *Homo sapiens*' in the Gene Database redirects the query to Entrez Gene Database (Maglott *et al.* 2007) and returns 'TP53'

Table 4.1 *Repositories (R), Knowledge Databases (KDs) and Ontologies (O)*

	Type	Field	Reference/URL
ArrayExpress	R	Microarray	Parkinson *et al.* (2007)
CCDS	KD	Gene and Proteins	Pruitt *et al.* (2009)
ChEBI	O	Chemical Entities	Matos *et al.* (2009)
dbSNP	R KD	SNPs	Wheeler *et al.* (2007)
DIP	KD	Protein Interaction	Salwinski *et al.* (2004)
Entrez Gene	KD	Gene	Maglott *et al.* (2007)
FlyBase	KD	Fly Genome	http://flybase.org
GEO	R	High-throughput	Barrett *et al.* (2007)
GO	O	Gene	The Gene Ontology Consortium (2000)
HUGO	KD	Gene	http://www.genenames.org
JASPAR	R	PWM	Portales-Casamar *et al.* (2009)
KEGG	KD	Pathways	Kanehisa *et al.* (2006)
MGI	R+KD	Mouse Genome	http://www.informatics.jax.org
Negatome	KD	Protein Interaction	Smialowski *et al.* (2010)
OBO	O	Ontologies	Smith *et al.* (2007)
Oncomine	R	Cancer, microarray	Rhodes *et al.* (2007)
PDB	KD	Protein & Metabolomic	http://www.rcsb.org/pdb/home/home.do
PubMed	R	Journals	http://www.ncbi.nlm.nih.gov/pubmed
RefDIC	R	Immune system	Hijikata *et al.* (2007)
RefSeq	KD	Genes and Proteins	http://www.ncbi.nlm.nih.gov/RefSeq/
TRANSFAC	R	PWM	Windenger (2008)
UniProt	KD	Protein	The Uniprot Consortium (2010)

that links to http://www.ncbi.nlm.nih.gov/gene/7157 as the first result. Entrez Gene is a NCBI database for gene-specific information that focuses on genomes 'that have been completely sequenced, that have an active research community to contribute gene-specific information or that are scheduled for intense sequence analysis' and provides unique integer identifiers (GeneID) for genes and other loci. In *H. sapiens* the p53 GeneID is 7157; this identifier will help if any query is redirected to any other NCBI database. The official symbol, official name provided by a nomenclature authority, is TP53 and it was provided by the HUGO Gene Nomenclature Committee (HGNC; http://www.genenames.org/aboutHGNC.html). Other information available is: (1) TP53 has other aliases (P53, LFS1, TRP53, FLJ92943) that can be used to identify the gene in older references, (2) TP53 is a protein coding gene and (3) its HGNC identifier is HGNC:11998. Entrez Gene Database integrates information from RefSeq database (Pruitt *et al.* 2007; http://www.ncbi.nlm.nih.gov/RefSeq/), where RefSeq is 'a curated non-redundant collection of sequences representing genomes, transcripts and proteins' that integrates information from multiple sources therefore adding descriptions as: coding regions, conserved domains, gene and protein product names and (again) database cross-references. From a TP53 Entrez Gene database query the following RefSeq information can be obtained: (1) RefSeqGene identifier (for well-characterized genes to be used as reference standards) and (2) different transcripts and proteins, whose identifier names begin by NM_ and NP_, respectively. Among the transcripts and proteins the query returns the cellular tumor antigen *p53 isoform a*, that has two transcripts+protein related: (i) NM_000546.4 + NP_000537.3 and (ii) NM_001126112.1 + NP_001119584.1; in both cases there is a common reference, CCDS11118.1, to a consensus coding sequence (CCDS; Pruitt *et al.* 2009). The CCDS database annotates identical proteins on the reference mouse and human genomes with a stable identifier (CCDS ID) ensuring consistency between NCBI, Ensembl (Flicek *et al.* 2010), and UCSC Genome Browsers.

Following the links to Ensembl we arrive at Ensembl.org, a project that generates databases for chordates. The identifier for 'TP53 *H. sapiens*' is ENSG00000141510, and again at this page the original source, HGCN, and the access number on it, 11998, are shown. The web page provides the location of the gene (Chromosome 17: 7,565,257-7,590,856 reverse strand, GRCh37 human assembly) and different transcripts related to the gene, the length of each one of them and the protein products related. There are also the references to the CCDS if they are available. The CCDS database is only one of the examples of how different entities collaborate in standardizing annotations. Another major example is INSD (http://www.insdc.org/index.html) that combines the effort of the DNA Data Bank of Japan, GenBank (Benson *et al.* 2009) and the European Nucleotide Archive (European Molecular Biology Laboratory) to collect and disseminate DNA and RNA sequence data.

Other types of knowledge databases are not focused on individual terms but instead target physical interactions. Reactome (Matthews *et al.* 2009) is a curated knowledge database of biological pathways that includes cross-references to other biological databases such as Ensembl and Gene Entrez. A search for 'p53' within Reactome returns 123 terms (e.g. 'Transcriptional activation of p53 responsive genes' pathway, uniquely identified as REACT_202.2). The link to this pathway includes information about the preceding (e.g. 'Stabilization of p53', REACT_309.2) and following (e.g. 'Translocation of p27 to the nucleoplasm', REACT_9043.1) events. KEGG (Kanehisa *et al.* 2006), is a collection of manually drawn pathway maps from metabolism and cellular processes. A p53 query in KEGG PATHWAYS returns the KEGG 'p53 signaling pathway', identified as map04115; it includes (i) information about evidence used to develop the pathway, (ii) related pathways and (iii) links to other databases.

4.1.3 Ontologies

KDs present the processed data extracted from experiments. As we have observed in p53 we can obtain the sequence of the gene, the different exons and introns and other relevant data but we are missing answers to relevant questions such as 'Does p53 work alone or is it included in a gene module?' and 'Is p53 involved in any metabolomic pathway?'. These questions needed to be addressed from a different perspective than a KD. Therefore it was necessary to develop ways of storing relational information; Biological Ontologies (BOs) are one response to this need.

BOs represent the entities of biomedical interest and their relations and categories. Ontologies can be domain-specific (e.g. Chemical Entities of Biological Interest, ChEBI) (Matos *et al.* 2009) or level-specific [e.g. *Gene Ontology* (GO) has biological processes, cellular component and molecular function levels, see The Gene Ontology Consortium (2000)]. Ontologies can overlap, and can reuse elements from other ontologies. Ontologies are tools used to (i) integrate different meta-data, answering questions such as the existence of groups of entities and the possible hierarchical orders and relationships between them; and (ii) provide resource interoperability. In order to fulfill these tasks there are some prerequisites: high-quality, free availability, and re-distributiveness. *Open Biomedical Ontologies* (OBOs) are a collection of controlled vocabularies (ontologies) freely available to the biomedical community. Within OBOs, OBO Foundry (Smith *et al.* 2007) regulates the development of new ontologies by defining principles. Many new ontologies are defined to communicate with already available ones, e.g. *PRotein Ontology* (Natale *et al.* 2007) includes connections to *GO*, *OBO Disease Ontology* and several others. Until the mature development of federated biomedical ontologies different bridges are being created between ontologies. Two relevant examples are: (1) the Unified Medical Language System developed by the US National Library of Medicine whose Metathesaurus integrates more than 1.4 million concepts from over one hundred terminologies (http://www.bioontology.org/); and (2) the 'Minimal Information Requested In the Annotation of biochemical Models,' (MIRIAM) (Laibe and Le Novére 2007), that presents a set of guidelines for the annotation and curation of processes in computational systems biology models. MIRIAM Resources are being developed to support the use of Uniform Resource Identifiers, a useful tool for inter-operability.

4.1.3.1 p53 ontology tour

As in KDs we provide an example of ontology characteristics and structure by querying p53 in GO. GO contains a specific and curated (selected, collected and maintained by expert users) vocabulary for (i) the entities within the ontology, (ii) for terms related to entities (such as genes pointing to a biological process) and (iii) the terms related to the description of the entities. It is organized in three domains (cellular component, molecular function and biological process) and each domain is structured as a directed acyclic graph (DAG). The main page http://www.geneontology.org/ acts as a web browser that allows searching in the GO database.

A 'p53' query filtered by, '*H. sapiens*' in the biological process domain returns the term classified with the symbol 'TP53' and with the name 'Cellular tumor antigen p53'. Within the link to this term it is possible to retrieve information about the gene product (that offers different synonyms such as 'p53'), the peptide sequence, the sequence information and links to different Knowledge and Experimental Databases such as DIP, EMBL, GenBank, UniProt and PDB. Most importantly, the 'TP53 *H. sapiens*' page shows links to 60 different terms in the GO Database (e.g. Tp53 is related to the GO biological process term *apoptosis*, GO:0006915). All relations between genes and GO entities must be evidence based. There are two types of evidence: (a) Experimental Evidence that can be inferred from (i) Direct Assay, (ii) Physical Interaction, (iii) Mutant Phenotype, (iv) Genetic Interaction and (v) Expression Pattern; and (b) Computational Analysis Evidence that can be further classified as evidence inferred from (i) Sequence or Structural Similarity (Sequence Orthology, Sequence Alignment or Sequence Model), (ii) Genomic Context and (iii) Reviewed Computational Analysis. Evidence can be assigned by curators or assigned by automated methods; in all cases a clear trace of how the relation was generated must be provided. The relation between GO:0006915 and Gene Symbol 'TP53' is classified as Inferred from Direct Assay, it was assigned by UniProtKB and an identification for the reference is provided PMID:7720704 [the identifier in PubMed for the reference Eizenberg *et al.* (1995)]. Further exploration of the term GO:0006915 provides: (i) a definition of the term ('A form of programmed cell death that begins when...') and a reference (PMID:18846107), (ii) the relations to other GO terms in the DAG structure, such as 'apoptosis' *is a* 'programmed cell death' (GO:0012501), (iii) external references (e.g. links to Reactome), and (iv) a list of genes related to apoptosis (there are 1130 gene product associations).

4.1.4 Annotation

Annotation is the process of assigning properties to a given bioentity or the process of relating bioentities. For instance, if the entity is a gene the annotation process can (i) assign the gene to a gene set, (ii) classify the gene as constitutive or not constitutive and (iii) link the gene to other genes it regulates. Annotation is therefore a necessary process in the creation and updating of *Knowledge and Ontology* databases.

Annotation is based on evidence (as we observed previously in GO) that can be classified as Experimental Evidence or Computational Analysis Evidence. In this section we present some of the methods used to annotate databases and, at the end, we include a subsection that briefly reviews the R tools available.

4.1.4.1 Annotation by similarity

The very classic idea of annotation is based on the idea of 'similarity': *those elements that are similar in one aspect maybe be similar in other aspects*, therefore we can describe (functionally annotate) one gene by those genes that are similar to it; however the term 'similar' is specified differently in different approaches. Following this idea, high-throughput data have become a tool to functionally annotate genes and proteins (Kasif and Steffen 2010) by: (i) automated prediction of the function of genes based on homology and sequence similarity to genes of known function, (ii) organization of proteins (and genes) into clusters [PFAM (Finn *et al.* 2008 and US National Center for Biotechnology Information Protein Clusters (Klimke *et al.* 2009)] (iii) extending (i) by including further information such as phylogenetic profiles, coexpression, chromosomal gene clustering

and gene fusion. This information can be integrated with machine learning algorithms that are able to predict gene functions (Jansen *et al.* 2003). Automated annotation is a very active research field.

4.1.4.2 Annotation by Protein Binding Sites

Protein Binding Site (PBS) annotation is based on the identification of those sequences of nucleotides (protein-binding motifs) where a given Transcription Factor (TF, proteins that bind to promoter and/or enhancer regions of a gene regulating its expression) would bind. Protein Weight Matrices (PWMs) store information regarding which sequences are bound by a given TF; the PWM associated with a TF is a 4-row and n-column matrix that, for a nucleotide sequence of length n, each column i depicts the probability of the TF binding to a sequence that has the nucleotide in position i. TRANSFAC (semi-public; Matys *et al.* 2003; Windenger 2008) and JASPAR (public; Portales-Casamar *et al.* 2009; http://jaspar.genereg.net) are two PWM repositories. A search for TP53 in JASPAR database returns the MA0106.1 identifier, a *H. sapiens* zinc coordinating transcription factor that pertains to the Loop-Sheet-Heliz family and whose PWM is provided.

4.1.4.3 Annotation by Temporal Series

The use of Temporal Series (TS) in Annotation is that those bioentities that share the same expression pattern over time are regulated together and therefore, they would be expected to have the same functional group (cluster). Several statistical tools have been developed for the analysis of temporal series. Here we review some of the tools developed for microarray data analysis; however the main ideas can be extended to other data types (that also consider a small number of samples and a huge number of variables). The clustering methods can be classified by (i) the nature of the clusters they identify and (ii) the searching strategy. The searching algorithms can be grouped into three sets [from Krishna *et al.* (2010)]: (i) Pointwise distance based methods: grouping genes by minimizing an objective function generated by the distance (measure of similarity or dissimilarity) between pairs of genes. The description of this set can be found in Chapters 2 and 7 (see k-means and hierarchical clustering). (ii) Feature based clustering methods: grouping genes by using the general shape (local or global characteristics) of an expression profile, therefore they detect more complicated relations such as time-shifted or inversion relations. (iii) Model based clustering methods: based on statistical mixture models, which consider data to be generated from a finite mixture of underlying probability distributions, therefore each component corresponds to a different cluster. A very useful tool of this group is **MaSigPro** (Conesa *et al.* 2006), a statistical procedure for multi-series time-course microarray experiments. It is available as a Bioconductor package and from http://www.ivia.es/centrogenomica/bioinformatics.htm. Recent methodologies combine different strategies as in Krishna *et al.* (2010) where the authors combine pairwise distances (based on the Granger distance) with network clustering.

4.1.4.4 Experimental Design to improve annotation

Many experiments are designed to increase our knowledge of the relationship between bioentities. We discuss (and extend our previous approach to) PWMs as an example. The generation of JASPAR and TRANSFAC PWM is widely questioned because : (i) those matrices are constructed using a median of 18 individual sequences therefore they are expected to capture only a subset of the permissible range of binding sites; (ii) the accuracy of PWM models has been questioned (Benos *et al.* 2002); (iii) there are many examples in which transcription factors bind sets of sequences that cannot be described by standard PWMs (Chen and Schwartz 1995); and (iv) it is possible that PWMs are specific for different conditions (Berger *et al.* 2008). A more original approach that makes use of high-throughput technologies is Protein Binding Microarray (PBM) (Mukherjee *et al.* 2004); the authors used PBMs containing 41 944 60-mer probes in which all possible 10-base sequences were represented to analyze the

DNA-binding specificity. The specific construction of the microarrays provides a way to robustly estimate the binding preference of each protein to all 8-mers (Berger *et al.* 2006). Berger *et al.* (2006) provided a database of newly generated PWMs that does not solve all the problems stated but gives a major improvement.

4.1.4.5 Annotation by Text-Mining

Researchers whose experiments highlight new relationships between elements are encouraged to insert this information in publicly available curated databases. However because this is not always the case and due to the amount of information stored in journals, text-mining tools have been developed (Renear and Palmer 2009; Attwood *et al.* 2010). Text-mining is *the process of extracting information from text*, therefore it can be used as a tool for annotation in biology where text is extensively available in journals. Some relevant text-mining tools are:

1. iHOP (information Hyperlinked Over Proteins) (http://www.ihop-net.org/; Hoffman and Valencia 2004). It allows the search for biomedical terms that are mentioned together in the same sentence with a gene or protein of reference; the query returns all biomedical terms and for each one of them the references and sentences where they were both, term and gene/protein, mentioned.
2. PubGene (www.pubgen.org; Jenssen *et al.* 2001). It extends the query options offered by iHOP and it also generates a network that relates all single elements returned by the query and provides statistical significance of every relation in the network. For instance, a *TP53 H. sapiens* query indentifies tp53 in 10 265 documents and returns a list of elements (BAX, BCL2, CDKN1A, MKI67 and TCEAL1) organized within a network.
3. GRAIL (http://www.broadinstitute.org/mpg/grail/; Raychaudhuri *et al.* 2009). It integrates published scientific text with SNPs. Given a set of SNPs or genomic regions, a set of relevant genes is generated. From this gene set GRAIL searches in the literature for similarities in the associated genes. This tool can be understood as a *SNP to Gene set selection tool*, where usually SNPs are coming from the output of genome wide association studies.

4.1.4.6 Annotation in Bioconductor

Bioconductor uses the R programming language to develop 'tools for the analysis and comprehension of high-throughput genomic data' (www.bioconductor.org). It contains more than 380 packages and it is periodically updated. Within Bioconductor there are many tools that allow researchers to use annotations and ontologies. Each package is updated periodically and full descriptions of them can be found at the website.

Regarding annotation, there is a set of resources that allows programmers and users to map between probes, genes, proteins, pathways and ontology terms. Bioconductor has built-in representations of major ontology databases and data resources as:

1. GO: GO.db is a set of annotation maps that describes the entire Gene Ontology. GO, within each of its categories, is conceived as a DAG; within GO.db there is a set of datasets that specify those relations. This package is updated biannually.
2. KEGG: KEGG.db package provides information about the latest version of the KEGG pathway databases. It is updated biannually and it maps KEGG identifiers and elements within them to other databases such as GO terms or Entrez Gene.
3. Microarrays: there are packages that annotate the different microarray platforms and versions to different gene identifiers. An example is the classical Affymetrix Mouse Genome 430 2.0 Array, where in

Table 4.2 *R packages for Data Integration*

Package	Description	Reference
CCA	Canonical correlation analysis	González *et al.* (2008)
Gostats (B)	Tools for interacting with GO and microarray data (including functional enrichment)	Falcon and Gentleman (2007)
GSEA	Functional enrichment by Gene Set Enrichment Analysis	Subramanian *et al.* (2005)
IntegrOmics	Integration of different types of omic data	Lê Cao *et al.* (2009)
LRPath	Functional Set Enrichment by logistic regression	Sartor *et al.* (2009)
MaSigPro (B)	Analysis of multi-series time-course microarray experiments	Conesa *et al.* (2006)
RankAggreg	Tool that allows the combination of ordered lists using Rank aggregation	Pihur *et al.* (2009)

(B), package included in Bioconductor.

Bioconductor the annotation data (mouse4302.db, that provides mappings between manufacturer identifiers and other identifiers such as Entrez Gene and Ensembl), the cdf file (mouse4302.cdf, used to convert between (x,y)-coordinates on the chip to single-number indices and back) and the probe sequence data (mouse4302probe, the probe sequence data in a data-frame R object) are available.

4. Full genomes: there are packages that contain the different genomes sequenced such as *H. sapiens, Mus musculus and Saccharomyces cerevisiae*.

Most of the previous packages depend on the AnnotationDBi package that provides user interface and database connection code for annotation data packages using SQLite data storage.

4.2 Data integration in biological studies

This section reviews relevant examples of integrating Data repositories, KDs and Ontologies. We divide this section into two parts. We first review the integration of different experimental data types; the second part reviews the integration of meta-data and experimental data. Finally, we review how network and visualization tools have been used in data integration.

4.2.1 Integration of experimental data

In order to provide a unifying view of a biological system through all experimental data available it is needed to develop techniques to overcome the problem of integrating different data types and the different experimental conditions. From the literature we can extract two approaches. One considers (R) generic tools to integrate different omic data types. For instance **IntegrOmics** (Lê Cao *et al.* 2009), is a package developed to integrate different datasets, even if they are of different types. To deal with the problem of the large number of elements and the reduced number of measures ($p >> n$), the authors developed and implemented two different approaches: (i) a regularized canonical correlation analysis (**CCA**) (González *et al.* 2008) in the case of $p >> n$ (González *et al.* 2009) and (ii) a sparse partial least squares regression (Lê Cao *et al.* 2008) to simultaneously integrate and select variables using Lasso penalization. **RankAggreg** (Pihur *et al.* 2009) provides two methods (a Cross-Entropy method and a Genetic Algorithm) to combine ordered lists using Rank aggregation; the strenght of this approach is that Rank aggregation allows the combination of lists from different sources (e.g. data types). A second approach is to review practical cases whose methodologies can be standardized. Below we provide some key examples.

4.2.1.1 Microarrays sampled from different experimental designs

Microarray repositories are a major resource as thousands of microarray experiments have been stored over the last decade. However the challenges to use this resource in an integrative way are: (i) experimental designs are as a rule very different (e.g. different animal models and different experimental conditions), (ii) the necessary large number of comparisons (usually more than 10 000) (that will result in many false-positives unless appropriately stringent thresholds are employed, see Chapter 2) and (iii) different sources of variability (Jarvinen *et al.* 2004; Thompson *et al.* 2004); for example in some cases variability between technologies (such as the use of different microarray technologies) is lower than variability between laboratories (the same experiment, with the same technology in different laboratories) (Wang *et al.* 2005).

Two main approaches have been considered to deal with those challenges. One approach is to avoid individual-level comparison between datasets and use only data summaries. Oncomine (Rhodes *et al.* 2007) is a cancer microarray database that integrates a web-based data-mining platform. Researchers are expected to first select properly among the datasets available in the database and then to use meta-analysis to identify the genes that are significantly over-expressed or under-expressed across multiple independent studies. However other approaches have been evaluated that avoid the 'selection' step: in Submap, Hoshida *et al.* (2007) developed a method for integrating and comparing data from different datasets. The method begins with a set of datasets and a pre-gene grouping within each dataset, then it compares the relationship between the different groups by Gene Set Enrichment Analysis (see Chapter 7) and it summarizes the relations between different dataset clusterings by a matrix. This method has been validated against different sets and it is robust against different DNA microarray platforms and laboratories.

A second approach is to consider the low probability for multiple transcripts to follow a complex pattern of expression across dozens or hundreds of conditions by chance. Therefore, if those sets exist they may constitute coherent and biologically meaningful transcriptional units. However, transcriptional units *must* be validated by the use of other techniques and experimental designs. Chaussabel *et al.* (2008) designed a methodology to identify transcriptional modules formed by genes co-ordinately expressed in multiple microarray disease datasets. They tested the methodology over microarray datasets from blood samples and used the obtained modules to provide a set of biomarkers that were able to indicate the disease progression in patients with lupus erythematosus. Following the same idea but considering a predefined set of genes, Nilsson *et al.* (2009) developed a large-scale computational screen to identify mitochondrial proteins whose transcripts consistently co-express with the core machinery of heme biosynthesis. The idea is that interesting genes are those that are correlated with the gene set of reference only when the gene set is acting as a functional unit. The authors succeeded in proving (by experimental validation) that several top-ranked genes not previously related to heme biosynthesis and mitochondrial iron homeostasis were actually related.

A third approach is the use of clustering algorithms which are of major relevance in integrating datasets from different samples. However the algorithm classification provided in Chapters 2 and 7 need to be extended by the nature of the clustering. Methods can be further classified as: (i) one-way clustering, to find either gene clusters or sample clusters; (ii) two-way clustering, to find both gene clusters and sample clusters in a combined approach; and (iii) bi-clustering methods, gene clusters defined only over a sample cluster that is found simultaneously (Getz *et al.* 2000; Hägg *et al.* 2009).

4.2.1.2 Comparing and/or integrating different technologies

When a new experimental technique is developed it is tested against the well known results from previous techniques; this allows an evaluation of the weaknesses and strengths of new methodologies. Therefore validation can be considered as Data Integration because different data types should be compared. We compare

High Throughput Sequencing (HTS, see Chapter 7) versus microarray in Chromatin Immunoprecipitation (ChIP) analysis.

ChIP is a procedure used to determine whether a given protein binds to or is localized to a specific DNA *in vivo*. A short introduction on how it works would be: first a target [such as protein or chromatin mark, see for instance Barski *et al.* (2007) and Wang *et al.* (2008)] is selected, then an antibody that attaches to the selected target is added to the sample and it is used to purify selected DNA. The ChIP essay returns an amount of 'marked' DNA for a further study; the two major analysis methodologies applied to analyze it are HTS (ChIP-Seq) and microarray (ChIP-chip). HTS technology needs to map all readings to the genome of reference (generating a coverage) and then statistical tools are used to search for regions that are differentially expressed [such as MACS (Zhang *et al.* 2008) for TF binding sites, and SICER (Zang *et al.* 2009) for histone modification profiling] and therefore are considered to contain marks.

Park (2009) compares ChIP-Seq and ChIP-chip. The conclusion is that comparative analysis between technologies must take into account that: (i) microarrays are only measuring over predefined regions and no new region will be found; however, in ChIP-Seq new binding regions can be discovered; (ii) in microarrays there is cross-hybridization between probes and nonspecific targets, where in ChIP-Seq some GC bias can be present; (iii) the amount of DNA required is higher in ChIP-chip analysis; and (iv) the necessity of amplification steps is less required in ChIP-Seq. ChIP profiles are definitely more defined in ChIP-Seq however, if a nucleotide window is selected, there is a correlation between the profiles obtained by both technologies.

4.2.1.3 Genetics and Epigenetics

The importance of epigenetic modifications (Flintoft 2010) has been extensively shown in recent studies such as Barski *et al.* (2007), Jothi *et al.* (2008), Schones *et al.* (2008), and Wang *et al.* (2008, 2009) where the CD4+ T cell was extensively studied. The conclusion is that gene expression needs to account for epigenetic modifications as they modify the availability of the genes to be transcribed. Recently Karli *et al.* (2010) showed that it is possible to predict the expression level of a single gene by using a maximum of three histone modifications as predictors. Gene expression was measured by normalized microarray data, while for histone modifications the log tag number from ChIP-Seq experiments (one per modification) were used. The authors were able to export the models to other cells that were not trained for these data showing a clear validation of their initial assumptions. Recently, Artyomov *et al.* (2010), developed the first mathematical model that considers genetic and epigenetic regulatory networks, describing the transformations resulting from expression of reprogramming factors. However the major approaches that use epigenetic data are now based on statistical models.

4.2.1.4 Transcriptomics and Metabolomics

Jozefczuk *et al.* (2010) compare and integrate gene expression and metabolomic measurements over different stress conditions and over a period of time. For five conditions (oxidative stress, glucose-lactose diauxic shift, heat, cold and unperturbed culture as control) Gas Chromatography Mass Spectrometry (GC-MS) measurements were made in three different samples, for three different technical replicates and for 12 different times points. Microarray analysis was performed for three samples each without technical replicates and for two time points under each condition; except for oxidative stress condition for which 12 samples were measured. Individually both types of data return the same type of conclusions: in both cases, metabolites and mRNA expression, it is possible to group the time profiles by the type of perturbations. However, the authors show that the profiles generated by the metabolomic data are much more specific (to the perturbations) than the profiles generated by the transcriptomic data, showing that even if both omics are related, metabolites were more sensitive to the perturbations. Also condition-dependent associations between metabolites and transcripts were identified by co-clustering and canonical correlation analysis on combined metabolite and

transcript datasets. Therefore the authors were able to confirm existing models for co-regulation between gene expression and metabolites.

4.2.2 Ontologies and experimental data

A key example of integration of *knowledge and ontology databases* and *experimental gene expression data* is Gene-set analysis. The basic idea is to identify predefined (biologically relevant) gene sets (PGSs) enriched with differentially expressed (DE) genes. Sets associated with Gene Ontology terms (The Gene Ontology Consortium 2000) and KEGG pathways (Kanehisa *et al.* 2006) are of common use. Given two experimental conditions (control and disease), the identification of enriched GO biological terms points out possible biological processes that are involved in the disease development. A description of this technique and some variations are included in Chapter 7.

4.2.3 Networks and visualization software as integrative tools

Since the seminal paper on network analysis (Barabási and Albert 1999) and subsequent application to biological systems (Jeong *et al.* 2000, 2001), there has been a revolution on how we understand and analyze molecular biology data. One basic idea that makes networks a powerful tool is that any biological entity and their relations can be analyzed through them. More important, biological entities of different types can be compared through them. For instance we present an example in Figure 4.1; let us consider that Figure 4.1 shows a set of genes (named A, B, C and D) that can be considered as genes or as their related proteins depending on the context. In Figure 4.1(a) each node denotes a gene; there is an edge between genes if both pertain to the same module by clustering analysis of transcriptomic data sets (see Section 4.2.1). Figure 4.1(b) considers each node as a protein and shows a link between two nodes if there is experimental validation by yeast two-hybrid interaction of physical interaction [such as binding, see Steltlz *et al.* (2005)]. In Figure 4.1(c) each node denotes both a gene and its related protein; it is shown a direct edge (x,y) if protein x binds to the promoter region of y by using for instance PWMs. We can compare networks by observing which relations are unique to each one of the networks and which are common to all and by comparing network properties (such as degree distribution and distribution of the shortest paths; see Chapters 14 and 15 for greater detail). All networks can also be merged for further analysis [see Figure 4.1]. Chapters in Part C develop these ideas and show their integrative power.

However, networks need to be visualized. One visualization key tool is Cytoscape, which is an open source software that allows visualization of molecular interaction networks (Shannon *et al.* 2003) and includes tools to integrate these interactions with experimental data (Cline *et al.* 2007). As a visualization tool it includes: (i) Data Integration that supports many standards such as Simple Interaction Format, GML, BioPAX, SBML and OBO and it allows the importing of data files; (ii) Visualization manager (VizMapper); and (iii) Network Analysis tools. One of the major resources of Cytoscape is the number of plug-ins available; two plug-ins of interest are BinGO (Maere *et al.* 2005) that checks for representation of Gene Ontology categories in biological networks and CABIN (Collective Analysis of Biological Interaction Networks) (Singhal and Domicob 2007) that enables analysis and integration of interactions evidence obtained from multiple sources. Cytoscape is able to import data files generated by R.

4.3 Concluding remarks

Biological sciences are in a revolutionary phase fueled by recent advances in technologies for measuring biological entities and states. New data types are being produced and the volume of data that requires storage is rapidly increasing. This situation presents new challenges, not only in terms of data storage but perhaps primarily in the sense of data integration. Furthermore data integration is becoming a must as it is generally admitted that no data type provides a complete vision of any biological system.

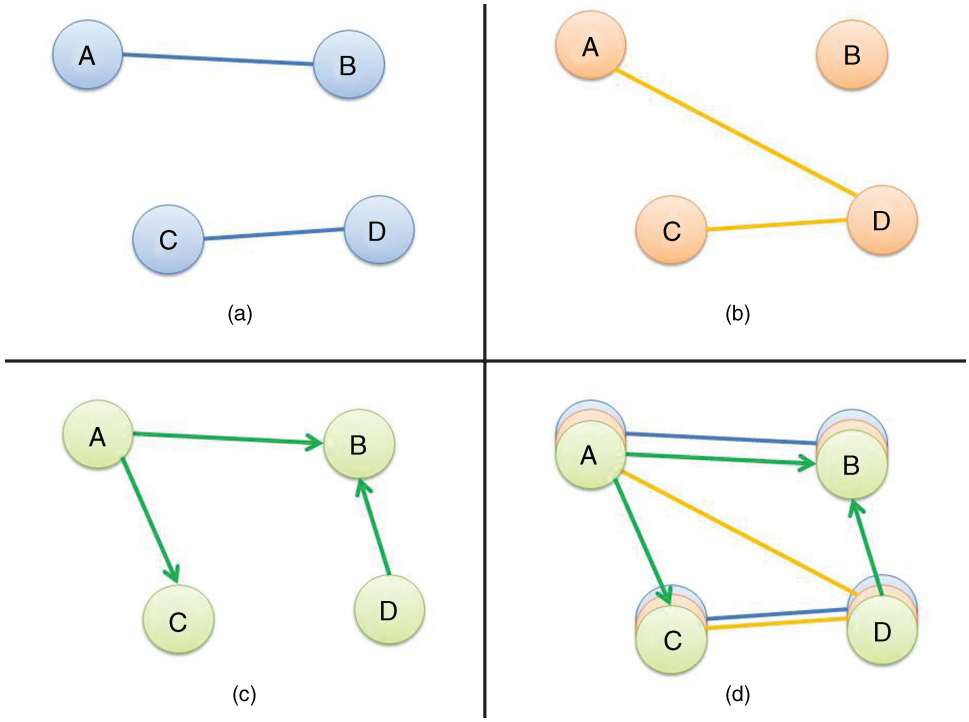

Figure 4.1 *Examples of gene and protein networks. (a) Gene association by clustering algorithms. (b) Protein association by physical interaction such as binding. (c) Node association by transcription factor binding to a promoter sequence. (d) Network merging*

We have shown that KDs and Ontologies are storage structures that provide an organized view of current knowledge, but it is necessary to further develop these structures to increase their scope and integration. On the other hand, the use of the experimental data and/or knowledge stored will make necessary the development of (statistical) generic tools able to integrate different data types; we can expect both metabolomics and lipidomics to increase in volume and quality and thereby increase the need for integrative tools beyond current transcriptomics and proteomics applications.

Equally important to development and use of standards and new tools for data integration will be the development of tools for scientific visualization. This area and its application to systems biology is still rather underdeveloped with the exception of Cytoscape and a few others. This opens up the possibility for exciting projects involving computer scientists with expertise in visualization, computational biologists and experimental and medical researchers.

Finally we find that data integration will be the key challenge in systems biology; statistics, networks, mathematical analysis and data structures will be key technologies in the success of this integrative approach.

References

Attwood TK, Kell DB, McDermott P, *et al.* 2010 Utopia documents: linking scholarly literature with research data. *Bioinformatics* **26**, i568-i574.

Artyomov MN, Meissner A and Chakraborty AK 2010 A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency. *PLoS Computational Biology* **6**(5), 1–14.

Baker SJ, Fearon ER, Nigro JM, *et al.* 1989. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244**, 217–221.

Barabási AL and Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512.

Barrett T, Troup DB, Wilhite SE, *et al.* 2007 NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research* **35**, 760–765.

Barski A, Cuddapah S, Cui K *et al.* 2007 High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4), 823–837.

Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al.* 2009 GenBank. *Nucleic Acids Research* **37**, 26–31.

Benos PV, Bulyk ML and Stormo GD. 2002 Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Research* **30**, 4442–4451.

Berger MF, Philippakis AA, Qureshi AM, *et. al.* 2006 Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology* **24**, 1429–1435.

Berger MF, Badis G, Gehrke AR, *et al.* 2008 Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276.

Brazma A, Hingamp P, Quackenbush J, *et al.* 2001 Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* **29**, 365–371.

Chaussabel D, Quinn C, Shen J, 2008 A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150–164.

Chen CY and Schwartz RJ 1995 Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. *Journal of Biological Chemistry* **270**, 15628–15633.

Clark TA, Schweitzer AC, Chen TX, *et al.* 2007 Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biology* **8**, 64.

Cline MS, Smoot M, Cerami E, *et al.* 2007 Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**, 2366–2382.

Conesa A, Nueda MJ, Ferrer A and Talo M 2006 maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**(9), 1096–1102.

DeLeo AB, Jay G, Appella E, *et al.* 1979 Detection of a transformation related antigen in chemically induced sarcomas and other transformed cells of the mouse. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 2420–2424.

Edgar R, Domrachev M and Lash AE 2002 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210.

Eizenberg O, Faber-Elman A, Gottlieb E, *et al.* 1995 Direct involvement of p53 in programmed cell death of oligodendrocytes. *EMBO Journal* **14**(6), 1136–44.

Falcon S and Gentleman R 2007 Using GOstats to test genelists for GO term association. *Bioinformatics* **23**(2), 257–258.

Finn RD, Tate J, Mistry J, *et al.* 2008 The Pfam protein families database. *Nucleic Acids Research* **36**, 281–288.

Flicek P, Aken BL, Ballester B, *et al.* 2010 Ensembl's 10th year. *Nucleic Acids Research* **38**, 557–562.

Flintoft L 2010 Complex disease: Adding epigenetics to the mix. *Nature Reviews Genetics* **11**, 94–95.

Getz G, Levine E and Domany E 2000 Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 12079–12084.

González I, Dejean S, Martin PGP and Baccini A 2008 CCA: an R package to extend canonical correlation analysis. *Journal of Statistical Software* **23**, 1–14.

González I, Dejean S, Martin PGP, *et al.* 2009 Highlighting relationships between heteregeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems* **17**, 173–199.

Hacia JG, Fan J, Ryder O, *et al.* 1999 Determination of ancestral alleles for human single nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Letter* **22**, 164–167.

Hägg S, Skogsberg J, Lundström J, *et al.* 2009 Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) Study. *PLoS Genetics* **5**(12), e1000754.

Hijikata A, Kitamura H, Kimura Y, *et al.* 2007 Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. *Bioinformatics.* **23**(21), 2934–2941.

Hoffmann R and Valencia A 2004 A gene network for navigating the literature. *Nature Genetics* **36**, 664.

Horak CE and Snyder M 2002 ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods in Enzymology* **350**, 469–83.

Hoshida Y, Brunet JP, Tamayo P, *et al.* 2007 Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS ONE* **2**(11), e1195.

Ihaka R and Gentleman R 1996 R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Jansen R, Yu H, Greenbaum D, *et al.* 2003 A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453.

Jarvinen A, Hautaniemi S, Edgren H, *et al.* 2004 Are data from different gene expression microarray platforms comparable? *Genomics* **83**, 1164–1168.

Jay G, Khoury G, DeLeo AB, *et al.* 1981. p53 transformation-related protein: detection of an associated phosphotransferase activity. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 2932–2936.

Jenssen T, Laegreid A, Komorowski J and Hovig E 2001 A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* **28**, 21–28.

Jeong H, Tombor B, Albert R, *et al.* 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

Jeong H, Mason S, Barabási AL and Oltvai ZN 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42.

Jothi R, Cuddapah S, Barski A, *et al.* 2008 Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research* **36**(16), 5221–5231.

Jozefczuk S, Klie S, Catchpole G, *et al.* 2010 Metabolomic and transcriptomic stress response of *Escherichia coli*. *Molecular Systems Biology* **6**, 364.

Kanehisa M, Goto S, Hattori M, *et al.* 2006 From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* **34**, 354–357.

Karli R, Chung HR, Lasserrea J, *et al.* 2010 Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **107**(7), 2926–2931.

Kasif S and Steffen M 2010 Biochemical networks: the evolution of gene annotation. *Nature Chemical Biology* **6**(1), 4–5.

Kimball R and Ross M 2002 *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling.*, John Wiley & Sons, Ltd, New York.

Klimke W, Agarwala R, Badretdin A, *et al.* 2009 The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Research* **37**, 216–223.

Krishna R, Li C and Buchanan-Wollaston V 2010 A temporal precedence based clustering method for gene expression microarray data. *BMC Bioinformatics* **11**, 68.

Laibe C and Le Novére N 2007 MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology* **1**, 58.

Lane DP and Crawford LV 1979. T antigen is bound to a host protein in SV40-transformed cells. *Nature* **278**, 261–263.

Lê Cao KA, Rossouw D, Robert-Grani C and Besse P 2008 A sparse PLS for variable selection when integrating Omics data. *Statistical Applications in Genetics and Molecular Biology* **7**, Article 35.

Lê Cao K, Gonzalez I and Dejean S 2009 IntegrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **21**, 2855–2856.

Lesne A and Benecke A 2008 Feature context-dependency and complexity-reduction in probability landscapes for integrative genomics. *Theoretical Biology and Medical Modelling* **5**, 21.

Linzer DI and Levine AJ 1979 Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell* **17**, 43–52.

Maere S, Heymans K and Kuiper M 2005 BiNGO: a Cytoscape plugin to assess over representation of Gene Ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449.

Maglott D, Ostell J, Pruitt KD and Tatusova T 2007 Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **35**, 26–31.

Matthews L, Gopinath G, Gillespie M, *et al.* 2009 Reactome knowledge base of human biological pathways and processes. *Nucleic Acids Research* **37**, 619–22.

Matos P, Alcantara R, Dekker A, *et al.* 2009 Chemical Entities of Biological Interest: an update. *Nucleic Acids Research* **38**, 249–254.

Matys V, Fricke E, Geffers R, *et al.* 2003 TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**(1), 374–378.

Mowat M, Cheng A, Kimura N, *et al.* 1985 Rearrangements of the cellular p53 gene in erythroleukaemic cells transformed by Friend virus. *Nature* **314**, 633–636.

Mukherjee S, Berger MF, Jona G, *et al.* 2004 Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**, 1331–1339.

Natale DA, Arighi CN, Barker WC, *et al.* 2007 Framework for a protein ontology. *BMC Bioinformatics*, **27**(8),S1.

Nilsson R, Schultz IJ, Pierce EL, *et al.* 2009 Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. *Cell Metabolism* **10**, 119–130.

Park PJ 2009 ChIP-seq: advantages and challenges of a maturing technology. 2009 *Nature Reviews Genetcs* **10**(10), 669–680.

Parkinson H, Kapushesky M, Shojatalab M, *et al.* 2007 ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research* **35**, 747–750.

Pihur V, Datta S and Datta S 2009 RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **10**, 62.

Portales-Casamar E, Thongjuea S, Kwon AT, *et al.* 2009 JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research* 1–6.

Pruitt KD, Tatusova T and Maglott DR 2007 NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35**, 61–65.

Pruitt KD, Harrow J, Harte RA, *et al.* 2009 The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* **19**(7), 1316–23.

Raychaudhuri S, Plenge RM, Rossin EJ, *et al.* 2009 Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLOS Genetics* **5**(6), e1000534.

Renear AR and Palmer CL 2009 Strategic reading, ontologies, and the future of scientific publishing. *Science* **325**(5942), 828–832.

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, *et al.* 2007 Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**(2), 166–180.

Salwinski L, Miller CS, Smith AJ, *et al.* 2004 The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, 449–451.

Sartor MA, Leikauf GD and Medvedovic M 2009 LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* **25**(2), 211–217.

Schones DE, Cui K, Cuddapah S, *et al.* 2008 Dynamic regulation of nucleosome positioning in the Human Genome. *Cell* **132**(5), 887–898.

Shannon P, Markiel A, Ozier O, *et al.* 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11), 2498–2504.

Singhal M and Domicob K. 2007 CABIN: Collective Analysis of Biological Interaction Networks. *Computational Biology and Chemistry* **31**, 222–225

Smialowski P, Pagel P, Wong P, *et al.* 2010 The Negatome database: a reference set of non-interacting protein pairs *Nucleic Acids Research* **38**, 540–544.

Smith B, Ashburner M, Rosse C, *et al.* 2007 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25**(11), 1251–1255.

Stelzl U, Worm U, Lalowski M, *et al.* 2005 A human protein-protein resource interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968.

Subramanian A, Tamayo P, Mootha VK, *et al.* 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**(43), 15 545–15 550.

The Gene Ontology Consortium 2000 Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29.

The UniProt Consortium 2010 The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* **38**, 142–148.

Thompson KL, Afshari CA, Amin RP, *et al.* 2004 Identification of platform-independent gene expression markers of cisplatin nephrotoxicity. *Environmental Health Perspectives* **112**, 488–494.

Wang H, He X, Band M, *et al.* 2005 A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics* **6**(1), 71.

Wang Z, Zang C, Rosenfeld JA, *et al.* 2008 Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* **40**(7), 897–903.

Wang Z, Zang C, Cui K, *et al.* 2009 Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**(5), 1019–1031.

Wheeler DL, Barrett T, Benson DA, *et al.* 2007 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **35**, D5–12

Wilhelm E, Pellay F, Benecke A and Bell B 2008 TAF6d controls apoptosis and gene expression in the absence of p53. *PloS One* **3**(7), e2721.

Windenger E 2008 TheTRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics* **9**(4), 326–332.

Zang C, Schones DE, Zeng C, *et al.* 2009 A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958.

Zhang Y, Liu T, Meyer CA, *et al.* 2008 Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**(9), 137.