Int. J. Dev. Biol. 56: 859-866 (2012)
doi: 10.1387/ijdb.120138fm

THE INTERNATIONAL JOURNAL OF DEVELOPMENTAL BIOLOGY
www.intjdevbiol.com

Knowledge-based bioinformatics for the study of mammalian oocytes

FRANCESCA MULAS*[,1], LUCIA SACCHI[2], LAN ZAGAR[3], SILVIA GARAGNA[1,4], MAURIZIO ZUCCOTTI[5], BLAZ ZUPAN[1,3] and RICCARDO BELLAZZI*[,1,2]

[1]Centre for Tissue Engineering, University of Pavia, Pavia, Italy, [2]Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Pavia, Italy, [3]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, [4]Laboratorio di Biologia dello Sviluppo, Dipartimento di Biologia e Biotecnologie 'Lazzaro Spallanzani', Università degli Studi di Pavia, Pavia, Italy and [5]Dipartimento di Scienze Biomediche, Biotecnologiche e Traslazionali, Università degli Studi di Parma, Italy.

ABSTRACT Bioinformatics tools have been recently applied to study the differentiation of the mammalian oocyte during folliculogenesis. In this review, we will summarize our knowledge of 1) the use of biological databases for the extraction of relevant information, 2) bioinformatics methods for knowledge extraction and representation, 3) the application of these methods to the study of mammalian oocyte differentiation and 4) state-of the-art prediction approaches for the assessment and estimation of the cell differentiation status.

KEY WORDS: *knowledge extraction, database, oocyte, stem cell*

Introduction

Change is the main theme that describes developmental and differentiation processes. For example, beyond the morphological modifications that occur during the development of the mammalian zygote into a blastocyst or the *in vitro* differentiation of embryonic stem cells (ESCs) into cardiomyocytes, a multitude of changes occur in the molecular backstage. Networks of genes are switched on and off, are down- or up-regulated along pathways that for many of these processes still remain unknown.

One of these yet unexplored developmental processes is the differentiation of the mammalian oocyte during folliculogenesis. Although the paucity of this biological material has made high throughput studies difficult to perform, appropriate bioinformatics tools have now been made available to bring to light the underlying molecular changes and to provide us predictive models for the cell differentiation status.

In this review we will present our knowledge on 1) the utility of biological databases for the extraction of relevant information, 2) bioinformatics methods and tools for knowledge representation, 3) applications of these methods to the study of the mammalian oocyte differentiation, and 4) state-of the-art prediction approaches for assessing the differentiation stage of cells.

Knowledge bases for bioinformatics analysis

Biological function and pathway databases

One of the most important goals in systems biology is the identification of the functions of genes and their relationships. To this regard, genes are often profiled through their expression measured under different environmental or experimental conditions and those with similar transcriptional profiles are likely to participate in common processes or share common functions. Bringing to light the relationships among transcriptional products requires prior knowledge on gene function for at least a subset of similarly profiled genes. A primary source for such annotation is Gene Ontology (GO) (Ashburner *et al.*, 2000), a controlled vocabulary for describing the role of genes and gene products in a number of organisms. At the highest level, GO consists of three ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. A gene product might be located in one or more cellular components; it is active in one or

Abbreviations used in this paper: ESC, embryonic stem cell; GEO, Gene Expression Omnibus; GO, Gene Ontology.

***Address correspondence to:** Francesca Mulas. Centre for Tissue Engineering, University of Pavia, Via Ferrata, 1 - 27100, Pavia, Italy. e-mail: francesca.mulas@unipv.it and Riccardo Bellazzi, Dipartimento di Ingegneria Industriale e dell'Informazione, Università degli Studi di Pavia, Via Ferrata, 1 - 27100, Pavia, Italy. e-mail: riccardo.bellazzi@unipv.it

Final, author-corrected PDF published online: 5 February 2013.

ISSN: Online 1696-3547, Print 0214-6282

more biological processes, during which it performs one or more molecular functions.

Besides organizing a vocabulary of terms into an ontology, GO provides annotation data by assigning the vocabulary terms to genes and their products. Tools like DAVID (Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/) are available to find all the processes and functions in which genes of interest are known to be involved (Huang *et al.*, 2009b). GO can be useful also for the comparison between genes of different organisms and for studying new genomes. The controlled vocabularies are structured so that they can be queried at different levels: a user may exploit the GO to search for a specific gene or to find a particular process and explore its hierarchy of terms, as shown in Fig. 1.

Another source of functional annotations are pathway repositories. Pathways often provide a rich representation of the molecular mechanisms that are at the basis of biological functions. They may be represented as graphs with directional flow that summarize the current knowledge on how biomolecules work together. KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg) for example, is a collection of databases containing organism-specific networks of molecular interactions in the cells (Kanehisa *et al.*, 2012). The aim of KEGG repository is to link lower-level information (*e.g.* genes, proteins, enzymes, reaction molecules, etc.) with higher-level information (*e.g.* interactions, enzymatic reactions, pathways, etc.). Another repository with annotated pathways and reactions is Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html), a comprehensive and integrated source of information about human biological processes (Matthews *et al.*, 2009). The Reactome web portal provides a collection of tools that allow researchers to browse and visualize pathway models, and to carry out pathway-based analyses of complex experimental and computational data sets.

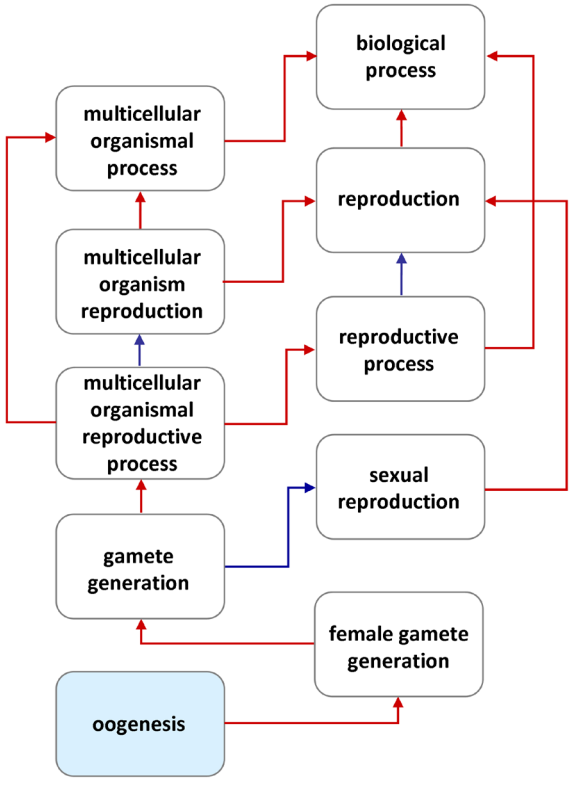


Fig. 1. Hierarchy of oogenesis-related Gene Ontology terms *(http://www.geneontology.org/). Red and blue arrows indicate "is a" and "part of" relations between the linked terms, respectively.*

Pathways, together with other functional and textual annotations, are exploited to infer molecular interactions and, as such, rendered in emerging Protein-Protein Interaction (PPI) databases. Among them, the Biomolecular Interaction Network Database (BIND, http://metadatabase.org/wiki/BIND_-_Biomolecular_Interaction_Network_Database) collects hypothesized interactions between two or more biological entities (*e.g.*, DNA, molecular complexes) (Bader *et al.*, 2003). Another resource is the Molecular INTeraction database (MINT, http://mint.bio.uniroma2.it/mint/Welcome.do), which contains information about interactions obtained from work published in peer-reviewed journals, excluding genetic or computationally inferred interactions from the database (Licata *et al.*, 2012).

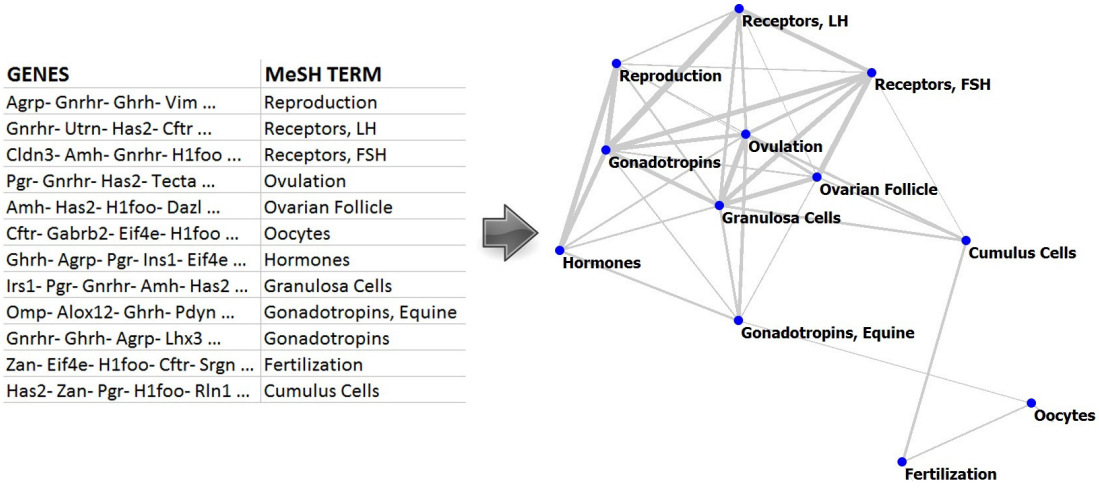
Literature databases

Very useful resources for automated exploration and inference of biomedical relations are literature repositories. Perhaps the most widely used is Entrez by National Library of Medicine, that also aims at integrating other types of information from the databases maintained by National Center for Biotechnology. These databases include nucleotide sequences, protein sequences, macromolecular structures, genomes and MEDLINE articles. Access to these resources is granted through the web-based interface of PubMed and through Entrez application program interfaces. Besides containing links to full text articles, PubMed also provides links to many other databases such as Nucleotide, Protein, Structure, Taxonomy, Genome, Expression, and Chemical Databases. Entries in these databases are often linked to terms in several ontologies. PubMed citations, for example, have been assigned MeSH terms and publication types from the Medical Subject Headings (MeSH; http://www.ncbi.nlm.nih.gov/mesh). MeSH is a controlled vocabulary thesaurus crafted by the National Library of Medicine and is used for indexing, cataloguing, and searching for biomedical and health-related information and documents. It consists of sets of terms organized in a hierarchical structure that permits searching at various levels of specificity. MeSH refers to the domain of medicine and includes different kinds of concepts that indicate the subject of an indexed article.

Articles present in the MEDLINE database are annotated with MeSH terms by expert curators, who summarize the presented information and the described genes. This tight association with the MEDLINE database, in addition to the hierarchical structure, is what makes MeSH vocabulary a very useful tool for searching and indexing journal citations and other data. Being structured and organized in hierarchical trees, MeSH terms can be very useful to identify relevant genes in the field of interest on the basis of the available literature. An example of this annotation is shown in Fig. 2, where the genes annotated to oocytes-related MeSH terms are used to assess a measure of the similarity among the terms. More complex approaches, that will be discussed in the following section, deal with both functional and textual annotation of genes.

Data repositories

Researchers are increasingly encouraged by scientific journals



GENES	MeSH TERM
Agrp- Gnrhr- Ghrh- Vim ...	Reproduction
Gnrhr- Utrn- Has2- Cftr ...	Receptors, LH
Cldn3- Amh- Gnrhr- H1foo ...	Receptors, FSH
Pgr- Gnrhr- Has2- Tecta ...	Ovulation
Amh- Has2- H1foo- Dazl ...	Ovarian Follicle
Cftr- Gabrb2- Eif4e- H1foo ...	Oocytes
Ghrh- Agrp- Pgr- Ins1- Eif4e ...	Hormones
Irs1- Pgr- Gnrhr- Amh- Has2 ...	Granulosa Cells
Omp- Alox12- Ghrh- Pdyn ...	Gonadotropins, Equine
Gnrhr- Ghrh- Agrp- Lhx3 ...	Gonadotropins
Zan- Eif4e- H1foo- Cftr- Srgn ...	Fertilization
Has2- Zan- Pgr- H1foo- Rln1 ...	Cumulus Cells

Fig. 2. Literature-based gene expression analysis. *Medical subject heading (MeSH) terms and their annotated genes are used to create a network of keywords where the terms that share a significant number of annotated genes are linked.*

to deposit their data on freely available community resources, which store them in appropriate formats for comprehensive analysis. For instance, Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) provided by the National Center for Biotechnology (NCBI), serves as a public repository for a wide range of high-throughput experimental data, including single and dual channel microarray-based experiments measuring mRNA and protein abundance, as well as non-array techniques, such as mass spectrometry peptide profiling and various types of quantitative sequence data (Edgar *et al.*, 2002). Currently, GEO contains about 10000 experiments provided by 5000 different research groups, and it includes a pipeline for data analysis, named GEOR.

Another well-known database of functional genomics experiments, including gene expression, is ArrayExpress, provided by the European Bioinformatics Units (Parkinson *et al.*, 2009). Data stored in ArrayExpress can be analyzed through a web-based application named Gene Expression Atlas (http://www.ebi.ac.uk/gxa). This tool relies on a subset of re-annotated data, which can be queried for gene expressions under a specific biological condition and across sets of experiments.

One utility of data repositories is comparative analysis of experiments. For instance, a researcher interested in the transcriptomic signature of the developmental competence of oocytes may compare his experimental results to results from other available data sets on the same process. Once such data have been selected from GEO or ArrayExpress, these resources enable the application of bioinformatics analysis on the raw gene expression data (Fig. 3).

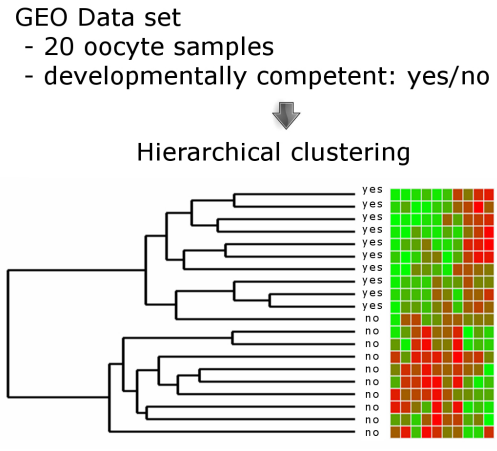


Fig. 3. Example of a pipeline for analysis of Gene Expression Omnibus (GEO) data from developmentally competent ('yes') and not competent ('no') oocytes. *A heat map reflecting microarray gene expression values across 20 samples belonging to these two different conditions can be obtained with bioinformatics tools that usually provide a measure to group similar samples (clustering).*

Methods for knowledge extraction and representation

Enrichment analysis

One of the initial steps of any bioinformatics analysis is often related to the inspection of a subset of interesting genes (or gene list) in order to extract any available information about the functions and pathways such genes are involved into. Besides considering each gene as a single entity, it is also important to understand which are the biological functions that are enriched in a set of candidate genes.

The annotation enrichment is a procedure for statistical analysis of biological annotations that are over-represented (enriched) in a candidate subset and that are thus indicative of a particular experiment or phenotype (Huang *et al.*, 2009a). An annotation term is significantly over-represented if the probability of finding by chance the same or a higher number of genes associated to that term is low. The significance of a term depends on the proportion of genes annotated to it in the list of interest and on the number of genes belonging to a specific genome and associated to that term. The hypergeometric or the binomial probability distributions are commonly used to determine the significantly enriched terms and their biological functions.

While these strategies require a fixed subset of candidate genes, a more complex and popular approach is the Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea/index.jsp). GSEA eliminates the need to impose specific thresholds for gene subset selection by proposing an approach based on the

scoring and ranking of genes (Subramanian *et al.*, 2005). Gene scores may be computationally obtained by, for example, comparing case-control transcriptions, making GSEA ideal for analysis of quantitative data. The genes ranked according to the selected score are compared to gene sets from KEGG, GO or other sources and enrichment is reported on the basis of computation of probability that the top-ranked genes belong to the target gene set.

Enrichment analysis can simplify the interpretation of large-scale experiments as researchers can focus on gene sets, which tend to be more interpretable. Thanks to these strategies, the perspective of the analysis is moved from single genes to larger groups of molecules that represent biological functions at a cellular level.

Several enrichment analysis tools have become popular for the study of oocytes and embryos. DAVID is one of them; it integrates annotation terms from several sources and besides enrichment analysis enables the search for related genes or terms and rendering of genes on pathway graphs to facilitate biological interpretation.

DAVID is an example of a web application that integrates best statistical practices and most reliable knowledge repository in a fixed, predefined analytics pipeline. An alternative approach is that of data exploration environments, which allow users to construct analytics pipelines on their own. The benefits are a gain in flexibility and an interactive data analytics, at the expense of higher complexity of the user's interface. An example of such tools is Orange (http://orange.biolab.si/) which uses the visual programming paradigm to build of analytics procedures that combine data mining and bioinformatics components (Curk *et al.*, 2005). An example of a data analysis pipeline constructed with this tool is shown in Fig. 4, where differentially expressed genes from a GEO experiment are identified and then analyzed for term enrichment in GO Browser. Another example of a toolbox that combines several functionalities for analysis is Babelomics (http://babelomics.bioinfo.cipf.es/), which is conceived for the analysis of transcriptomics, proteomics and genomics data and allows interpreting the results through different functional enrichment or gene set methods (Al-Shahrour *et al.*, 2005). Such interpretation can be performed using both functional annotations from GO and KEGG and regulatory information from PPI databases. Functional classification of genes and visualization of the related categories is nicely supported in PANTHER (Protein ANalysis THrough Evolutionary Relationships, http://www.pantherdb.org/) Classification System (Thomas *et al.*, 2003). This tool takes into account the gene families, the GO classes and the Pathways, to assign a functional class to each gene.

The toolboxes described in this section are just some few examples examples of a growing set of general or dedicated software platforms for bioinformatics data analysis. These are regularly reported in special application sections of journals such as Bioinformatics, BMC Bioinformatics and Nucleic Acids Research.

Text mining approaches

With the increasing volume of biomedical publications, the automated analysis and mining on literature is becoming an essential part of biomedical data exploration pipelines. One of the most challenging goals is the extraction of specific embedded information, such as associations of genes with their related diseases and treatments, and experiment-specific retrieval of related literature and its summarization.

To this end, a number of Text Mining and Natural Language Processing methods have been developed for the automated elaboration of textual content of biomedical publications (Krallinger *et al.*, 2008, Weeber *et al.*, 2005). The main efforts in this area have been related to the identification of biological entities (Name Entity Recognition) in free text. The most straightforward literature analysis approach for this task is the extraction of keywords and their rep-

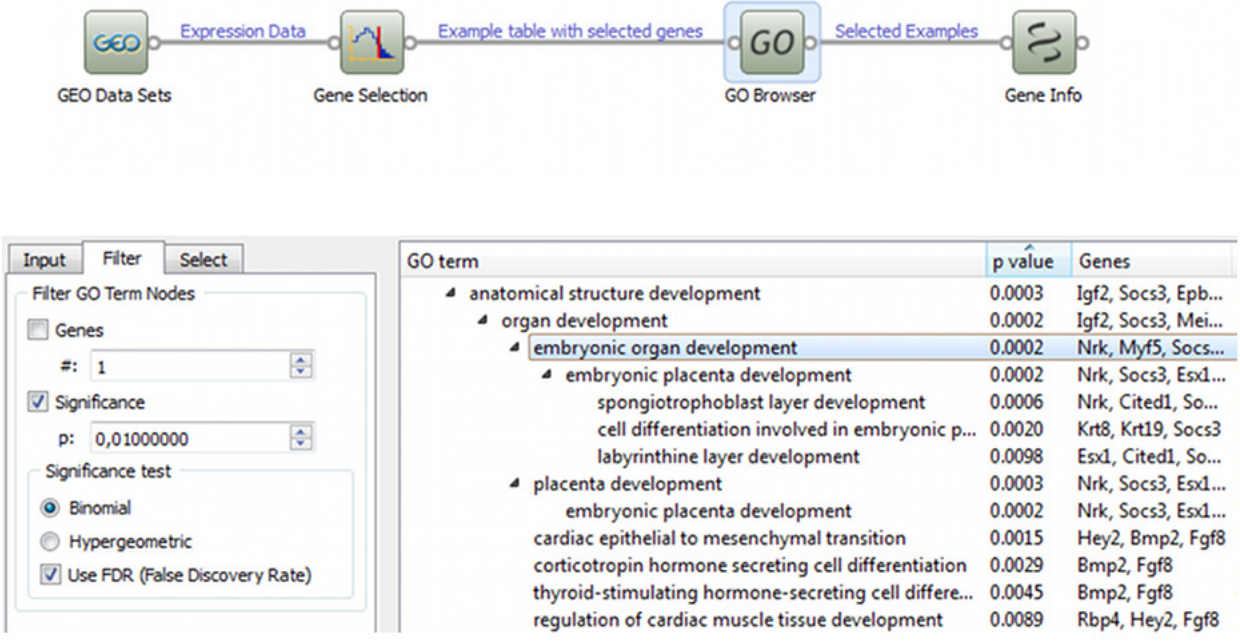


Fig. 4. An example of microarray data analysis schema in Orange. *Significantly expressed genes selected from a Gene Expression Omnibus (GEO) data set are analyzed with Gene Ontology (GO) enrichment. The 'GO Browser' enables to set the parameters of the analysis (*e.g. *statistical test and p-value thresholds) and to visualize the enriched terms and the corresponding genes within the hierarchy of GO terms. Categories of genes may be selected for further bioinformatics analysis.*

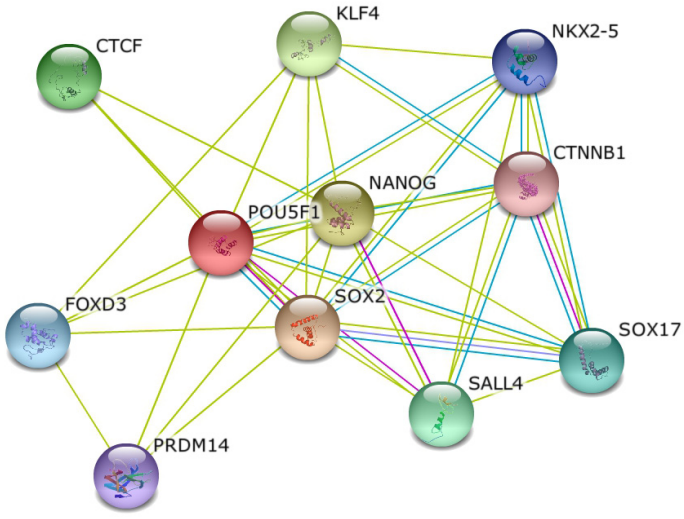


Fig. 5. Network of OCT4 interacting proteins obtained with STRING *(http://string-db.org/).*

resentation through frequency patterns. After some pre-processing steps required for the text decomposition, such as lemmatization and stemming, words representing specific kinds of entities (e.g., gene names) are identified from free text. This step must deal with a problem of ambiguity, due to the existence of different names for the same entity and to abundance of acronyms and abbreviations in biomedical literature, which can result in a misinterpretation of words. For this reason, Name Entity Recognition strategies usually rely on other databases, such as the Entrez Gene NCBI database, to derive the correct names (Nuzzo *et al.*, 2010). In a separate step, a set of keywords is analyzed to retrieve significant co-citation with the entities of interest. A crucial aspect of the analysis is about the type of considered keywords that may change according to the functional context in which an entity applies. Although every word from the text can be analyzed, the most common way to consider functional context in text mining is the introduction of structured, hand-curated information about biological entities. For instance, MeSH vocabulary assigns domains like diseases or anatomy to publications. Such annotations can be used to establish different 'lines of evidence' for an entity relation derived from text. A certain disease can be assigned to a gene if the paper where the entity was identified had been assigned to the disease via MeSH. The level of association is then assigned. Here the simplest approach is to consider their occurrence in the text or the number of publications that contain the same evidence. More sophisticated methods take into account the position of the word in the sentence and weight each term according to a measure of the importance of the term for a particular gene. A recently proposed method associates genes based on the disease-related UMLS annotations extracted from PubMed. According to this procedure, terms about diseases or symptoms that are frequent in all the publications and not specifically assigned to the publications related to the considered gene are rated as less important (Nuzzo *et al.*, 2010).

An example of a software that nicely integrates data analysis and literature mining is Pathway Studio (http://www.ariadnegenomics.com/products/pathway-studio/). It enables the analysis of biological data, including gene expression or proteomics experiments (Nikitin *et al.*, 2003). The MedScan module of Pathway Studio aims at extracting biomedical information by using dictionaries to identify biological terms (*e.g.* pathways, proteins, etc) and extracting the relationships with natural language processing methods. Other text mining software suites specifically designed for biomedical data analytics are reviewed in (Lu *et al.*, 2011).

From gene lists to association networks

While enrichment analysis and text mining are helpful in unveiling the most updated information on a set of genes, they shed little light on the complex interacting groups of molecules that constitute living systems. When analyzing experimental data, we are not only interested into entities that are related to observed processes, but in their relations and networks that would reveal the underlying biological mechanisms.

For a number of well-studied model organisms such as budding yeast, it is possible to use validated PPI and highlight the associations most related to the process of interest. When considering highly complex species, including mouse and humans, researchers have to take into account other types of associations, due to the lack of functionally validated interactions. In gene networks, the concept of interaction is thus expanded as the findings from physical interaction experiments are combined with any potentially useful assertion on associations of biomedical entities from available knowledge-sources. Text mining and co-occurrence in the publications are evaluated for this purpose and the evidence coming from the literature is used by popular tools such as STRING (Szklarczyk *et al.*, 2011) and IPA (Ingenuity Systems, www.ingenuity.com) for constructing protein or gene networks. Unlike primary PPI interaction databases, STRING combines and weights information automatically extracted from high-throughput data, literature data mining, signaling and transcriptional pathways and organism-specific databases. STRING also automatically transfers PPI-validated interactions among organisms if a orthologous protein pair is present in another species. Within a web interface, STRING can render the networks related to a selected protein, where each link is enhanced with the information regarding the source of knowledge from which the interaction was inferred (Fig. 5). A score indicating the confidence of the predicted interaction is also provided, with higher values assigned when the link has been inferred from multiple sources.

A very popular commercial tool for exploration of biological networks is the software Ingenuity Pathway Analysis (IPA). It provides an environment for explorative analysis over biological interactions obtained from manually curated relationships among proteins, RNA, genes, metabolites, protein complexes, drugs and diseases. Every relation displayed can be traced back to the sources from where it was inferred and the software provides detailed related information that are manually curated and updated by experts. When run on a dataset, IPA Core Analysis enables the user to visualize and analyze the metabolic and signaling pathways, and to inspect cellular processes and transcription factors related to the genes of interest.

Applications to –omics analysis in oogenesis and early embryogenesis

Recently, there has been an increasing effort in leveraging the available knowledge for studying developmental processes during oogenesis and early embryogenesis. In particular, results from several studies using –omics technologies, such as microarrays

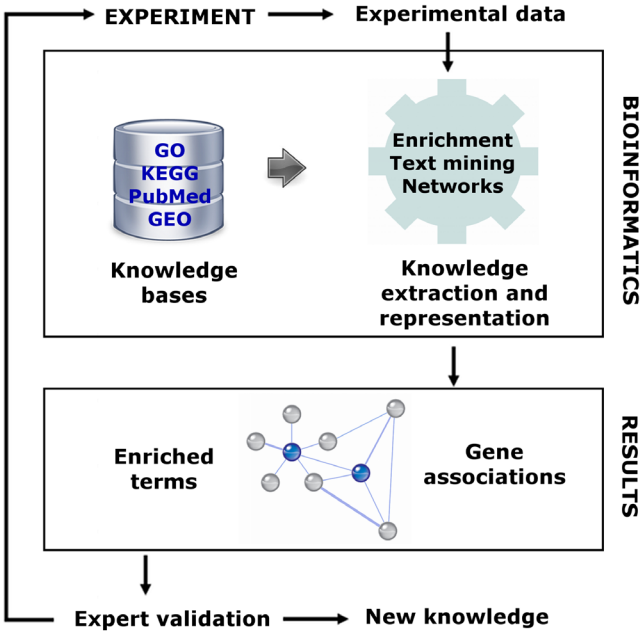


Fig. 6. The process of knowledge-based bioinformatics analysis. *Data coming from high-throughput experiments are analyzed with methods that extract the available knowledge from different biological repositories. Results in terms biological annotation terms and networks connecting genes or proteins are further analyzed and refined based on the knowledge of the expert and on the evidence coming from the data. For instance, genes that were not previously included in the lists would be considered in light of the knowledge retrieved by means of the bioinformatics methods. The final aim of the process is to promote knowledge discovery.*

and mass spectrometry, have been analyzed using the bioinformatics tools for annotation and association of molecules, with the common aim of hypothesizing unknown entities that play important role in cell differentiation. For instance, researchers in the field of oocytes differentiation may consider a subset of genes known as maternal-effect genes. In the analysis of data from oocytes, an added value is represented by the evidence that a known maternal-effect gene is often related to another gene that has not been previously considered. Such evidence can be obtained from gene annotations and association networks.

Most of the current works in this area focus their bioinformatics analysis on gene set enrichment, with particular reference to GO processes and KEGG pathways. In one of these studies, thanks to the analysis performed with DAVID, it was possible to highlight the major differences in terms of activated pathways between *in vitro* matured and immature bovine oocytes, thus confirming the transcriptional changes during oocyte growth (Mamo *et al.*, 2011). The results provided insights into the metabolic pathways whose activation prevent the maturation of developmentally competent oocytes. An integrated pipeline of proteomics experiments coupled with GO enrichment and IPA networks was also applied to bovine oocytes at the germinal vesicle stage (Peddinti *et al.*, 2010). The methods allowed identifying the signaling and the processes that may have an impact on oocytes developmental competence and maturation.

With a similar aim, MII developmentally competent oocytes (MII^{SN}) were compared with oocytes that cease development at the 2-cell stage (MII^{NSN}) (Zuccotti *et al.*, 2011, Zuccotti *et al.*, 2008). The GO enrichment analysis performed with DAVID and the networks obtained using IPA allowed the identification of gene expression networks involved into the regulation of biochemical pathways representative of the adverse biological status of MII^{NSN} oocytes. In addition, the exploited knowledge-based methods highlighted the essential role of the OCT4 transcription factor, whose down-regulation in incompetent oocytes induces the up-regulation of a group of genes involved in the activation of adverse pathways, such as oxidative phosphorylation, mitochondrial dysfunction and apoptosis. More recently, the same authors provided a thorough analysis of the identified genes in preimplantation embryos derived from MII^{NSN} and MII^{SN} oocytes (Zuccotti *et al.*, 2011). The transcriptional link between eggs, early preimplantation embryos and embryonic stem cells (ESCs) was assessed thanks to a comprehensive bioinformatics analysis that took advantage of the functions provided by the Orange software. First, the enriched GO terms were identified through the GO Browser widget, as shown in Fig. 4. A measure of the connection of these genes with cancer-related processes was assessed thanks to a statistical analysis of the gene expression data stored in ArrayExpress. Moreover, a literature-based search strategy and text mining techniques, described in (Nuzzo *et al.*, 2010), were used to gather the available knowledge from PubMed about a set of OCT4-regulated genes whose transcripts were detected in both MII oocytes and 2-cell embryos. The text mining methods allowed retrieving a set of MeSH terms from the publications, which were combined with information from the Gene Ontology to assign to each gene a set of weighted annotation terms. The obtained annotation profiles were then used to assess a measure of the similarity among the genes, that allowed the construction of a OCT4-transcriptional network, aimed at exploring the contribution of not previously considered molecular factors to the mammalian egg developmental competence.

Although the use of data automatically extracted from the literature is not very common in the publications related to oocyte development, other works have shown the benefit of the obtained information. Novel proteomic technologies coupled with an exhaustive bioinformatics analysis were used to uncover the

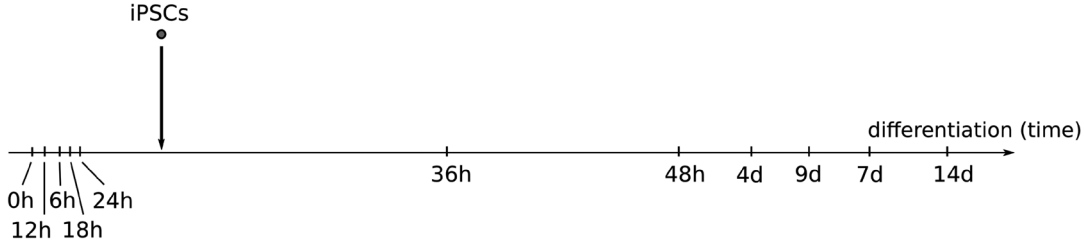


Fig. 7. Example of a differentiation scale *where a sample obtained from induced pluripotent stem cells (iPSCs) is projected to the scale to uncover its real stage of development with respect to standard conditions of differentiation.*

maternal proteins that play a role in mature oocytes (Zhang *et al.*, 2009). In this work, Babelomics allowed the identification of the over-represented and under-represented GO terms, while the text mining tool Pathway Studio was used for retrieving the entries in PubMed related to the proteins of interest. The results of the text mining procedure were visualized in a graph where the proteins were linked to their most frequently associated pathway-related keywords (e.g., oogenesis).

In most of the presented approaches, researchers follow a data analysis pipeline that relies on the steps shown in Fig. 6.

For example, Zuccotti *et al.,* (2011) expanded the set of genes initially selected relying on the expression values considering the information extracted from GO and the literature. The knowledge bases and the methods for extracting and representing the knowledge were then used to produce a OCT4 transcriptional network. This result was validated by the authors in order to identify biologically relevant clusters in the network, and to define the most interesting genes (*i.e.* an additional set of OCT4-regulated genes) that were unveiled by means of the knowledge-based methodologies. The knowledge discovery process continued with the revision of the gene expression data from oocytes and embryos in order to highlight the change of transcription of the selected genes.

Similarly, the enrichment and literature-based analysis of the proteomics experiments conceived by Zhang *et al.,* (2009) helped identifying a set of proteins of interest, whose gene names were used to query an additional database containing mRNA expression profiles in different tissues. This procedure allowed the identification of a group of genes of the T-cell leukemia family that can be classified as oocyte-specific based on their expression patterns.

Knowledge-based bioinformatics have been successfully applied for the identification of the stem cells regulatory network (Campbell *et al.*, 2007, Pardo *et al.*, 2010), which shares a large part of the genes with the OCT4-transcriptional network active during oocyte development (Ding *et al.*, 2012, Zhang *et al.*, 2009, Zuccotti *et al.*, 2011). In stem cell research, the reliability of the identified networks is usually increased by integrating information coming from different knowledge repositories. To discover the functions of the predicted list of OCT4 interactors, Pardo *et al.,* (2010) used a set of bioinformatics resources, including GO, KEGG, PANTHER, MINT and other PPI databases, to retrieve the previously known interactions (Pardo *et al.*, 2010). Similar studies have applied innovative pathway enrichment methods (Babaie *et al.*, 2007) or approaches for analyzing the literature in PubMed (Campbell *et al.*, 2007) achieving the same goal in the characterization of ESCs networks.

Future directions and challenges

The results of the studies presented in the previous section point out the evidence of a relationship between some transcription factors that play a key role in oogenesis and a number of genes taking part also in the regulatory networks active in ESCs. However, the link between eggs, early preimplantation embryos and ESCs should be further investigated in light of the great number of experiments on stem cells available in the biological data repositories.

One of the first efforts made in this direction is represented by the Embryonic Stem Cells Database (ESCDb, http://biit.cs.ut.ee/escd/). Based on human and mouse ESC-related datasets, this resource enables retrieving information on the interaction between a gene of interest and the most known transcription factors in stem cell development (Jung *et al.*, 2010). Links to other databases allow searching for genes with a specific behavior in selected tissues or pathways.

Recently, there has been an increasing effort in exploiting the available genome-wide expression data to highlight the transcriptional changes that occur during the differentiation of stem cells (Aiba *et al.*, 2009, Dutkowski and Ideker, 2011). In particular, some of these studies have proposed specific tools that infer on the pluripotency status of cells. Muller *et al.,* (2011) developed a classification method able to distinguish pluripotent from differentiated samples (Muller *et al.*, 2011). With a similar aim, a recently proposed bioinformatics pipeline was used to transform genome wide expression data into a graphical device that predicts the differentiation status of ESCs (Zagar *et al.*, 2011). Using this method, the stages corresponding to normal development are displayed in a one-dimensional ruler, referred to as *differentiation scale*. Given the transcriptional phenotype of a sample from new experimental conditions (e.g., induced pluripotent stem cells (iPSCs) or cellular lines treated with chemical agents), the model is used to project the experimental sample on the scale in order to determine its actual differentiation stage with respect to that of the control sample (Fig. 7). Interestingly, a reduced set of genes was identified in each stage as responsible for determining the transcriptional signature of the cell in that stage, which included known pluripotency and differentiation markers as well as genes not previously studied in this field (Mulas *et al.* 2012). The results of these studies suggest that the signature of pluripotency is hidden in the transcriptome and may be unveiled by bioinformatics approaches. These data-driven predictive tools ought to be integrated with the available knowledge, including the established stem cells regulatory networks and the PPI databases. Differentially expressed genes contribute, only for a limited part, to the entire network that determines developmental changes. PPI resources may be useful to identify other relevant genes, i.e., those that have not been selected as differentially regulated in the experiment, but are surrounded by a number of regulated genes in the interaction network (Nitsch *et al.*, 2010). Once the expanded network of genes that play a role for determining the signature of a specific pluripotency status have been derived, the results may be used to obtain a more precise picture of the links existing between ESCs and the mammalian oocytes.

Acknowledgments

This work was supported by the Fondazione Cariplo grant (2008–2006), the Italian Ministry of Health ITALBIONET project, CARE-MI EU FP7 project (Health-F5-2010-242038), UNIPV-Regione Lombardia, Fondazione Alma Mater Ticinensis and by the grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

AIBA, K., NEDOREZOV, T., PIAO, Y., NISHIYAMA, A., MATOBA, R., SHAROVA, L.V., SHAROV, A.A., YAMANAKA, S., NIWA, H. and KO, M.S. (2009). Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells. *DNA Res* 16: 73-80.

AL-SHAHROUR, F., MINGUEZ, P., VAQUERIZAS, J.M., CONDE, L. and DOPAZO, J. (2005). BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res* 33: W460-464.

ASHBURNER, M., BALL, C.A., BLAKE, J.A., BOTSTEIN, D., BUTLER, H., CHERRY, J.M., DAVIS, A.P., DOLINSKI, K., DWIGHT, S.S., EPPIG, J.T. *et al.,* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat*

Genet 25: 25-29.

BABAIE, Y., HERWIG, R., GREBER, B., BRINK, T.C., WRUCK, W., GROTH, D., LEHRACH, H., BURDON, T. and ADJAYE, J. (2007). Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells* 25: 500-510.

BADER, G.D., BETEL, D. and HOGUE, C.W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 31: 248-250.

CAMPBELL, P.A., PEREZ-IRATXETA, C., ANDRADE-NAVARRO, M.A. and RUDNICKI, M.A. (2007). Oct4 targets regulatory nodes to modulate stem cell function. *PLoS One* 2: e553.

CURK, T., DEMSAR, J., XU, Q., LEBAN, G., PETROVIC, U., BRATKO, I., SHAULSKY, G. and ZUPAN, B. (2005). Microarray data mining with visual programming. *Bioinformatics* 21: 396-398.

DING, J., XU, H., FAIOLA, F., MA'AYAN, A. and WANG, J. (2012). Oct4 links multiple epigenetic pathways to the pluripotency network. *Cell Res* 22: 155-167.

DUTKOWSKI, J. and IDEKER, T. (2011). Protein networks as logic functions in development and cancer. *PLoS Comput Biol* 7: e1002180.

EDGAR, R., DOMRACHEV, M. and LASH, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30: 207-210.

HUANG, D.W., SHERMAN, B.T. and LEMPICKI, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.

HUANG, D.W., SHERMAN, B.T. and LEMPICKI, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44-57.

JUNG, M., PETERSON, H., CHAVEZ, L., KAHLEM, P., LEHRACH, H., VILO, J. and ADJAYE, J. (2010). A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS One* 5: e10709.

KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M. and TANABE, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40: D109-114.

KRALLINGER, M., VALENCIA, A. and HIRSCHMAN, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* 9 Suppl 2: S8.

LICATA, L., BRIGANTI, L., PELUSO, D., PERFETTO, L., IANNUCCELLI, M., GALEOTA, E., SACCO, F., PALMA, A., NARDOZZA, A.P., SANTONICO, E. *et al.,* (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40: D857-861.

LU, Z., KAO, H.Y., WEI, C.H., HUANG, M., LIU, J., KUO, C.J., HSU, C.N., TSAI, R.T., DAI, H.J., OKAZAKI, N. et al. (2011). The gene normalization task in BioCreative III. *BMC Bioinformatics* 12 Suppl 8: S2.

MAMO, S., CARTER, F., LONERGAN, P., LEAL, C.L., AL NAIB, A., MCGETTIGAN, P., MEHTA, J.P., EVANS, A.C. and FAIR, T. (2011). Sequential analysis of global gene expression profiles in immature and *in vitro* matured bovine oocytes: potential molecular markers of oocyte maturation. *BMC Genomics* 12: 151.

MATTHEWS, L., GOPINATH, G., GILLESPIE, M., CAUDY, M., CROFT, D., DE BONO, B., GARAPATI, P., HEMISH, J., HERMJAKOB, H., JASSAL, B. *et al.,* (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622.

MULAS, F., ZAGAR, L., ZUPAN, B. and BELLAZZI, R. (2012). Supporting regenerative medicine by integrative dimensionality reduction. *Methods Inf Med* 51: 341-347.

MULLER, F.J., SCHULDT, B.M., WILLIAMS, R., MASON, D., ALTUN, G., PAPAPETROU, E.P., DANNER, S., GOLDMANN, J.E., HERBST, A., SCHMIDT, N.O. *et al.,* (2011). A bioinformatic assay for pluripotency in human cells. *Nat Methods* 8: 315-317.

NIKITIN, A., EGOROV, S., DARASELIA, N. and MAZO, I. (2003). Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics* 19: 2155-2157.

NITSCH, D., GONCALVES, J.P., OJEDA, F., DE MOOR, B. and MOREAU, Y. (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 11: 460.

NUZZO, A., MULAS, F., GABETTA, M., ARBUSTINI, E., ZUPAN, B., LARIZZA, C. and BELLAZZI, R. (2010). Text Mining approaches for automated literature knowledge extraction and representation. *Stud Health Technol Inform* 160: 954-958.

PARDO, M., LANG, B., YU, L., PROSSER, H., BRADLEY, A., BABU, M.M. and CHOUDHARY, J. (2010). An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell* 6: 382-395.

PARKINSON, H., KAPUSHESKY, M., KOLESNIKOV, N., RUSTICI, G., SHOJATALAB, M., ABEYGUNAWARDENA, N., BERUBE, H., DYLAG, M., EMAM, I., FARNE, A. *et al.,* (2009). ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 37: D868-872.

PEDDINTI, D., MEMILI, E. and BURGESS, S.C. (2010). Proteomics-based systems biology modeling of bovine germinal vesicle stage oocyte and cumulus cell interaction. *PLoS One* 5: e11240.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., PAULOVICH, A., POMEROY, S.L., GOLUB, T.R., LANDER, E.S. *et al.,* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545-15550.

SZKLARCZYK, D., FRANCESCHINI, A., KUHN, M., SIMONOVIC, M., ROTH, A., MINGUEZ, P., DOERKS, T., STARK, M., MULLER, J., BORK, P. *et al.,* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39: D561-568.

THOMAS, P.D., CAMPBELL, M.J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. and NARECHANIA, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13: 2129-2141.

WEEBER, M., KORS, J.A. and MONS, B. (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform* 6: 277-286.

ZAGAR, L., MULAS, F., GARAGNA, S., ZUCCOTTI, M., BELLAZZI, R. and ZUPAN, B. (2011). Stage prediction of embryonic stem cell differentiation from genome-wide expression data. *Bioinformatics* 27: 2546-2553.

ZHANG, P., NI, X., GUO, Y., GUO, X., WANG, Y., ZHOU, Z., HUO, R. and SHA, J. (2009). Proteomic-based identification of maternal proteins in mature mouse oocytes. *BMC Genomics* 10: 348.

ZUCCOTTI, M., MERICO, V., BELLONE, M., MULAS, F., SACCHI, L., REBUZZINI, P., PRIGIONE, A., REDI, C.A., BELLAZZI, R., ADJAYE, J. *et al.,* (2011). Gatekeeper of pluripotency: a common Oct4 transcriptional network operates in mouse eggs and embryonic stem cells. *BMC Genomics* 12: 1-13.

ZUCCOTTI, M., MERICO, V., SACCHI, L., BELLONE, M., BRINK, T.C., BELLAZZI, R., STEFANELLI, M., REDI, C.A., GARAGNA, S. and ADJAYE, J. (2008). Maternal Oct-4 is a potential key regulator of the developmental competence of mouse oocytes. *BMC Dev Biol* 8: 97.

Further Related Reading, published previously in the *Int. J. Dev. Biol.*

Rediscovering pluripotency: from teratocarcinomas to embryonic stem cells
Ivana Barbaric and Neil J. Harrison
Int. J. Dev. Biol. (2012) 56: 197-206

Impaired meiotic competence in putative primordial germ cells produced from mouse embryonic stem cells
Marianna Tedesco, Donatella Farini and Massimo De Felici
Int. J. Dev. Biol. (2011) 55: 215-222

In vitro germ cell differentiation during sex differentiation in a teleost fish
Tohru Kobayashi
Int. J. Dev. Biol. (2010) 54: 105-111

Differentiation of mouse primordial germ cells into female or male germ cells.
N Nakatsuji and S Chuma
Int. J. Dev. Biol. (2001) 45: 541-548

A bioinformatics approach to investigating developmental pathways in the kidney and other tissues.
J B Bard
Int. J. Dev. Biol. (1999) 43: 397-403

5 yr ISI Impact Factor (2011) = 2.959