Gene Ontology Based Meta Analysis of Genome Scale Experiments

Chad A. Shaw, Ph.D.

Contents

0.1 Introduction and Background

Genome scale experiments and microarray expression studies have become increasingly common in biological and medical science. The proliferation of data makes the integration of results both a pressing challenge and an opportunity for discovery. The identification of commonalities between studies can add confidence to the interpretation of results, while the identification of differences gives insight to the distinctions between treatments and experimental conditions. In this chapter we explore the utility of gene annotations – specifically Gene Ontology (GO) annotations – as a resource to facilitate meta-analysis across a family of experiments.

We consider GO based analysis at the level of gene lists. Gene lists are the most commonly accessible summary of results from expression array or library sequencing experiments. Experimental results for microarray studies are sometimes available as numerical summaries. However, in this chapter we will focus only on gene lists.

The chapter is organized to give the reader an overview of the Gene Ontology system and its implmentation. The chapter then considers the statistical methods used to analyze single lists. We then extend these methods to consider groups of lists which arise from many studies. Finally we consider an actual data analysis problem using a family of lists from reports in the stem cell community.

0.2 Ontologies

Ontologies are a growing and important formalism in the world of computing. The term ontology originates in philosophical literature where it means system of knowledge. More recently, computational authors have defined an ontology: “An ontology is an explicit specification of a conceptualization” [8]. In the current context we intend the term ontology to mean a generalized taxonomy for information in some field of study. Ontologies present a way to organize information to facilitate storage, retrieval, and general computing against the knowledge structure. Ontologies currently exist within a large number of different domains: engineering, medical science, linguistics, and of course molecular biology and genetics.

In practice an ontology is usually implemented as a controlled vocabulary with well defined relationships connecting the terms. Although the number and variety of terms as well as the types of relationships will vary across fields, most ontologies can be represented mathematically as graph structures where nodes are terms and edges connect the related terms. Ontologies are most often represented as Directed Acyclic Graphs or DAGs. A DAG is a graph with directed edges where a child can have more than 1 parent. The acyclic character of the graph comes from the directionality of edges. The graph has a direction or flow from a root node to terminal nodes. Although there may be more than one path from the root to descendant nodes, there are no cycles or loops strating from ancestral nodes which lead down through the graph and then back up to ancestor nodes.

The internet and the flood of available online information has made ontologies increasingly important. Ontologies represent a rational and flexible way to unify information within a field of activity. Ontologies have increasingly been seen as a way to provide more intelligent data retrieval through smart internet searches [5]. Ontologies enhance information retrieval because query terms in a search can be placed into a context in a field of knowledge. Instead of pattern matching a string, the query term can be identified in the sematically structure ontology and results can be returned based on the relationship of the query term with other elements of the ontology. All terms in an ontological neighborhood of the query may provide useful results. Because the concept of an ontology is so general, it is likely that many of the tools and techniques which have been developed for analysis in the GO community have broader application. As well, many of the concepts and methodologies developed in the general ontology community have application within the biological spehere [7],[9].

0.3 The Gene Ontology

The Gene Ontology (GO) is a highly developed, active, community supported, species independent annotation system for describing genes. Information about the GO can be found at the well maintained and informative GO website (http://www.geneontology.org) and in numerous informative articles describing the development of the GO [3],[4]. The formal structure of the GO is quite simple at first glance, but the GO also has some fairly sophisticated features and it continues to evolve. In this section we examine the history and motivation for the GO. We also consider the implementation of GO and its various features. Finally, we discuss software tools used to manipulate GO information and to perform calculations with GO data.

0.3.1 History and Motivation for GO

To appreciate the GO project one should have an idea of the historical context in which the GO developed. The GO consortium emerged in the late 1990s in a period of enormous productivity but also fragmentation within molecular biology. Several parallel projects to completely sequence the genomes of a number of organisms were ongoing. As well, genome scale experiments such as gene expression microarray studies had just become technically feasible. In this context several model organism communities began a large scale initiative to share information on the genes whose sequence was revealed by the ongoing genome projects and whose expression profiles were being measured with microarrays.

Among the main insights derived from the sequencing work was the clear shared evolutionary history of the genes observed across the various organisms. A large number of genes – and the exact numbers and proportions are still debated – reappear with largely the same sequence among highly divergent organisms. The shared gene sequences arise because of the common evolutionary origin of the genomes of these species. This phenomena of common evolutionary origin is termed homology in the biological sciences, and homology is a central tool to make sense of the commonalities among the life on earth.

Common evolutionary origins appear at virtually every scale at which a biological system can be described. Common origins underly developmental patterns of animals and morphological features such as the limb bones of land animals. As well, all verterbrate neurological systems share homologous patterns of development. The maternal care behaviors of mammals are also an example of evolutionary homology. At the molecular scale, homology is manifest in the highly similar nucleotide and amino acid sequence found in the genomes and proteins of various taxa.

For the GO, the fortunate consequence of evolutionary homology is the shared framework evolution provides for describing what genes do. The various organisms may differ greatly in size, shape and behaviors, yet the common origin of life means that these organisms can be thought of as variations on a theme. The consequence of evolution is that a single vocabulary might be able to characterize the activity of genes and their role in the underlying life processes. It is possible to contemplate a single vocabulary applying universally to all organisms.

A central difficulty in creating such a shared terminology is not the nature of biology, but rather the nature of human beings. Despite the unifying principles of evolutionary science, the human communities of molecular biologists and geneticists – the scientific communities – have a sociologic tendency to become more specialized and divergent. For this reason the terminology – the names given to genes and their functions – have tended to differ, sometimes considerably. Although a gene might consistently appear in humans, mice and even yeast, that gene would likely have a different name in each taxon, and the jargon used to describe its biological properties would also be somewhat distinct.

The difficulty of divergent vocabularies only became clear in the late 1990s. At that time three separate model organism communities attempted to share the data archived in their species specific databases. The three separate databases were the Sachromyces Genome Database for Yeast, Flybase for FruitFlies and the MGI Database for Mouse musculus [3],[4]. The GO was initially developed in 1998 as a unification strategy to share information contained in just these three databases. The GO rapidly adopted a more lofty ideal to create a unified annotation system for all genes in all organisms. The feasibility of such a unified annotation system rested and continues to depend on the essential commonality and shared evolutionary origins of life on earth.

0.3.2 GO Specification and Implementation

Creating a shared ontology for describing all genes in all organisms is clearly a daunting challenge. To accomplish this goal, the GO consortium settled on a three-fold collection of structured vocabularies. The three separate vocabularies were adopted to essentially report on three distinct aspects of each gene's properties. These separate vocabularies are Biological Process, Molecular Function and Cellular Component. Genes are annotated into each of these three vocabularies.

Biological Process

The Biological Process ontology aims to capture the biological objective to which a gene's product contributes. The process ontology contains many terms which capture a notion of biological state change or transition. For instance, there are large sections of the ontology dedicated to the cell-cycle. The ontology also represents the processes of catabolism and metabolic re-arrangement. This ontology also describes cellular communication processes such as signal transduction and the process of DNA transcription. Importantly, the process ontology does not seek to represent the biochemical events necessary to accomplish a process; rather, the process ontology merely tries to describe the variety of processes available to a cell. This ontology has seen the most development is the most widely used in computational science.

Molecular Function

Molecular Function represents the biochemical aspect of what a gene product does. The function ontology describes only what is done without attempting to express the context in which reactions occur or what purpose they might serve. Examples of the function ontology include terms like “kinase” or “adenylate cyclase.” The terms in molecular function can be more difficult to comprehend without extensive training in the biological sciences.

Cellular Component

The Cellular Component ontology describes the physical localization of a gene product within cells. The terms in this ontology include cellular regions like “cell membrane” or “nucleus”. This ontology also includes terms that represent multi-protein complexes such as entities like “ribosomes” or “proteosomes” [3].

DAG

The terms in each ontology are organized as a directed acyclic graph or DAG. Again, a DAG is a graph where edges have directionality or flow, and a DAG differs from a tree in that a node can have more than 1 parent. The terms in each ontology DAG are arranged in a pattern of high generality to increasing specificity – from sweeping concepts to refined detail. The figure (0.1) presents a subset of the GO BiologicalProcess ontology.

Relationships

There are two main types of relationships between parental nodes and their descendants in the GO. These two relationships are the is-a relationship and the part-of relationship. The is-a relationship indicates that the descendant term is a sub-class or sub-type of the parental term. The part-of relationship is more complex. This relationship indicates that the descendant term is actually a component of the parental term. This relationship is especially important for the Cellular Component branch of the GO. In computational work, the distinction between these relationships is sometimes ignored. All parent-child

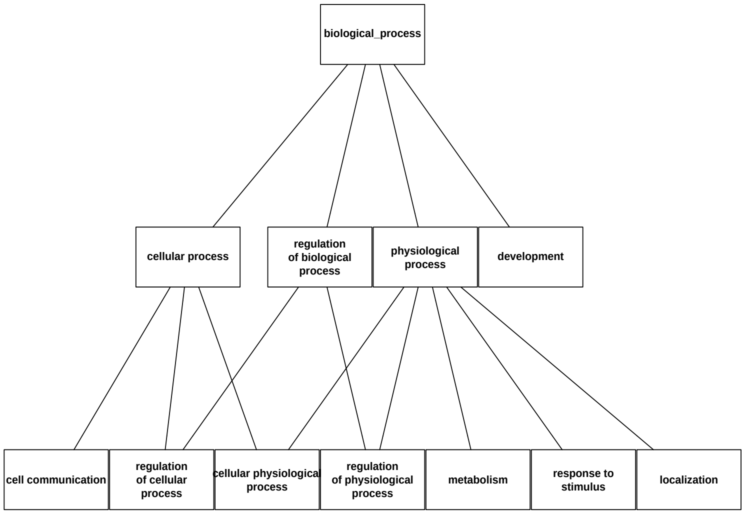

biological_process
cellular process
regulation of biological process
physiological process
development
cell communication
regulation of cellular process
cellular physiological process
regulation of physiological process
metabolism
response to stimulus
localization

Figure 0.1 The figure shows an abbreviated view of the Biological Process arm of the GO. The figure shows how terms are organized from generality to increasing specificity. As well, the figure shows descendant terms with numerous parents. This feature of the GO is essential to represent the semantics of molecular biology. Not all relationships or terms actually present in the ontology are depicted.

relationships in the GO are expected to abide the “true-path” rule. The “true-path” means that the heirarchical structure of the GO must hold from any term to its top level parent. The “true-path” rule must apply for relationships of either the is-a or part-of type (see http://www.geneontology.org/GO.usage.shtml for more information.)

Annotations

Genes are not nodes in the ontology. Rather, genes are annotated to terms within the ontology. When genes are assigned annotation within the GO, those annotations are expected to adhere to the “true-path” rule. If this property does not hold for some gene, then it is the responsibility of a curator to add the necessary GO terms to make a new path so that a gene may be annotated to the GO in such a way that the “true-path” rule will hold. Genes can be annotated to any number of different terms within the GO. There is no limit to the number of terms to which a gene may be assigned, nor is

Species	Biological Process	
	All codes	non-IEA codes
SGD Saccharomyces cerevisiae	6457	6457
FlyBase Drosophila melanogaster	10033	6834
MGI Mus musculus	13002	8773
GO Annotations @ EBI Human	21703	8331

Table 0.1 The table shows the counts of current annotations to the GO Biological-Process Ontology in a variety of species. Distinctions are made between electronically derived annotations and those which result from other sources.

their a constraint on the relationships between the terms. A table showing the number of annotations available is given in (0.1).

Evidence Codes

All annotations within GO have evidence codes. There are 20 evidence codes now in use. Some evidence codes refer to “electronically derived” information – which usually means sequence similarity. Importantly, a large number of GO annotations are based on real experiments. The number of such experimentally derived annotations continues to grow quite rapidly, and these “real-world” based annotations are highly valuable. Annotations based on sequence data are often somewhat less trustworty. A table listing the evidence codes appears in (0.2).

Species Specificity

The GO is designed to be species independent. This means that the vocabulary should be a common platform for analysis of genes from any taxa. The species independence makes the development of annotation resources faster – so that the effort can leverage work across communities. It also means that the GO can facilitiate multi-species comparisons. Species specific data exists, though, and there must be additional ontologies beyond the GO to analyze such data.

0.3.3 Software Tools

There are many softwares in use to analyze GO data. The various software serve a variety of purposes. Some tools, such as AmiGO, are purely for browsing the GO data structure and for comprehensively viewing all gene annotations. Other tools provide methods for enrichment analysis. Three such enrichment analysis tools are [1],[2], and [6]. Any software that is used for GO

Code	Description
IC	Inferred by Curator
IDA	Inferred from Direct Assay
IEA	Inferred from Electronic Annotation
IEP	Inferred from Expression Pattern
IGI	Inferred from Genetic Interaction
IMP	Inferred from Mutant Phenotype
IPI	Inferred from Physical Interaction
ISS	Inferred from Sequence or Structural Similarity
NAS	Non-traceable Author Statement
ND	No biological Data available
RCA	Inferred from Reviewed Computational Analysis
TAS	Traceable Author Statement
NR	Not Recorded

Table 0.2 Gene Ontology Evidence Codes

Software	Purpose	Language
AmiGO	browse, query, visualize	C, Java, Perl
DAVID	enrichment analysis	ASP, Java
FatiGO	enrichment analysis	unknown
Ontology Traverser	enrichment analysis	Java and R

Table 0.3 Gene Ontology Software Tools

content analysis must solve 2 fundamental problems:(a) the software must represent the GO data structure as a DAG and (b) the system must be capable of mapping experimentally generated lists to the DAG and computing the necessary count data.

The ontology itself is stored in a relational database. Many people curate the ontology and there are extensive resources to aid in the annotation process. The ontology data structure can be downloaded in at least three formats: MySQL tables, text, and XML. The XML format bears mentioning. XML is the most widely used language for sharing information of the web, and web based representations of the GO and derivative analyses can facilitate web based data sharing. A table describing a few GO software tools is provided in (0.3).

	Annotated to Term	Not Annotated to Term
In Gene List	X	$L - X$
Not In Gene List	$N_C - X$	$N - N_C - (L - X)$

Table 0.4 A two-way table representation of counts at a single GO node. The quantity of interest is X which is the number of genes which are both found in the list and which are annotated to the GO term.

0.4 Statistical Methods

The GeneOntology (GO) is well suited to computational analysis. The fundamental approach we consider is to tabulate experimentally derived sets of genes – gene lists – against the GO data structure. Once the tabulation is accomplished, we analyze the counts of genes annotated at or below each GO node. This procedure can be performed for single lists as well as collections of lists. When analyzing a single gene list, GO tabulation can rapidly summarize the list and reduce its content from a large number of often obscure individual genes to a smaller number of biologically relevant and interpretable categories. This form of GO directed data reduction is a central benefit of GO analysis.

When analyzing a collection of lists, the GO can generate exploratory comparisons of the lists for data visualization. Exploratory analysis provides a method to compare lists which differ markedly in size, experimental platform, or taxonomic source. In addition to exploratory methods, testing can be used to identify those GO categories and subgraphs with unusual gene-counts across the family of lists.

0.4.1 Analyzing a Single Gene List

The analysis of a single list concerns the count statistics determined when the list is tabulated against the GO structure. The random quantities to consider are the counts of genes annotated at or below each GO node.

Content Analysis in a Single Category

The simplest problem to consider is the analysis of counts at a single GO node. The representation of the problem in terms of a 2-way table is shown in (0.4).

The statistical question revolves around the magnitude of X. Under the null hypothesis that the list is randomly generated, the null distribution for X is usually considered to be the hypergeometric distribution. The hypergeometric distribution is used to model counts of objects drawn from a finite collection without replacement. The hypergeometric is often presented as a description of drawing balls from an urn. The total number of balls in the urn is N, and the balls come in 2 colors, red and black. Our interest is in the number of red balls we obtain in L sequential draws without replacement. The total number of red balls in the urn is N_C, and the number of black balls is $N - N_C$. The number of red balls in our sample of size L is the random outcome X. The distribution of X is given:

$$P(X = k|N, N_C, L) \quad = \quad \frac{\begin{pmatrix} L \\ k \end{pmatrix}\begin{pmatrix} N - N_C \\ L - k \end{pmatrix}}{\begin{pmatrix} N \\ L \end{pmatrix}} \tag{0.1}$$

Quite naturally this probability mass function is parametrized by the values of N,N_C and L. In our problem of assessing gene counts at a single GO node, we must make an analogy between the genes derived from our experiment and the imaginary balls drawn from the urn.

In our model, the objects being "drawn" are genes annotated to the GO. The universe of objects is the collection of all genes possibly observable in our assay which have some annotation to the GO. The marked objects are those annotated to the node under consideration. The assumptions we make in applying the hypergeometric model are detailed below:

- N is the total number of genes on the microarray or within our universe of possibly observable genes which have annotations to the GO major branch where the node we consider resides (eg. we are considering a node within the BiologicalProcess arm of GO).
- N_C is the number of genes in the universe of possibly observable genes which are annotated to the GO category (GO term) under consideration. Often this is a number far smaller than N.
- L is the number of genes in the list with annotations to the GO branch of the node
- X will represent the number of genes in our list which are also annotated to this GO node.
- We proceed under the null hypothesis that the list was randomly assembled, so that all N genes with GO annotations are a priori equally likely to appear in our assembly of L genes.

Under these assumptions, we can suppose that the null distribution for the number of genes annotated at or below a single node follows the probability law given in (0.1). When the number of annotations N grows large and the number of genes L is relatively small, the hypergeometric can be approximated by the binomial distribution:

$$P(X = k) \quad \approx \quad \begin{pmatrix} L \\ k \end{pmatrix}\left(\frac{N_C}{N}\right)^k\left(1 - \frac{N_C}{N}\right)^{L-k} \tag{0.2}$$

Also of great interest is the standardized gene count at a GO node. This standardized count is merely the observed count less the expected count under randomness divided by the standard deviation of the count under randomness. For the gene count Y this quantity is:

$$Z \quad = \quad \frac{Y - L\left(\frac{N_C}{N}\right)}{\left(L\left(\frac{N_C}{N}\right)\left(1 - \frac{N_C}{N}\right)\frac{(N-L)}{(N-1)}\right)^{\frac{1}{2}}} \tag{0.3}$$

When L is relatively small with respect to N, the quantity in the denominator is well approximated by the binomial variance estimate. In all situations the scale change from the binomial variance to the hypergeometric variance is constant across nodes for it does not depend on N_C but only depends on the total number of annotations for the universe defined by the assay and the size of the list under consideration.

The Multivariate Distribution of GO Counts

There is clearly more information in the tabulation of list counts than the observation at any single GO node. The specification of the joint distribution of counts is therefore of great interest. Under the null hypothesis that the genes in a list are randomly drawn from the collection of all annotated genes – meaning that all genes are equally likely to appear – analysis of the joint distribution follows the combinatorial arguments for the multivariate hypergeometric distribution. However, this joint distribution is complicated by the overlap in the genes annotated to GO nodes in distantly related branches of the GO structure.

The strict inheritance of annotations for nodes arranged in an unbranched path of GO nodes makes analysis of counts in unbranched paths possible. Recall that genes annotated to any GO node are by implication also annotated to their parental nodes according to the "true-path" rule. The annotations at ancestral nodes propogate up through the ontology by recursive application of the rule. Although distinctions can be made based on the is-a and part-of parent-child relationships or by evidence codes, in principle all child-node gene annotations are a proper subset of the annotations to each parent node. As will be shown, the analysis of the joint distribution of counts in unbranched paths is an important special case of a tractable joint distribution of GO counts.

Unfortunately, the joint distribution of counts for GO nodes which are linked through branched relationships can be quite complicated. The complication arises because a single gene can be annotated to any number of GO terms throughout the GO topology. When the joint distribution of GO counts in a subgraph containing arbitrarily related nodes is considered, the analysis should take into account the annotation overlap and the consequence for the distribution of counts. The analysis must consider the number of genes which share annotations at or below the nodes under consideration. Although it proves possible to give the joint distribution for pairs of nodes, the complete joint distribution for the entire graph is difficult to specify.

Unbranched Paths

We consider first the joint distribution of counts at nodes in an unbranched path stretching from the root GO term to a terminal descendant node. The derivation of the joint distribution for counts in an unbranched path appears below.

- Denote the list counts for genes annotated at each node along a path of $l+1$ nodes stretching from a root node to a descendant node $l+1$ steps away with the count vector $\mathbf{X} = (X_0, X_1, \ldots X_l)^t$. X_0 is the count of the number of genes in the list itself with annotations to the GO branch. For each $i \in 1 \ldots l$ the $i-1$ node in the path is a parent of the i^{th} node, so that the inequality $X_i \leq X_{i-1}$ holds.

More strongly, because of the "true-path" rule, the genes being counted at level i are shared by the parent at the level $i-1$.

- The maximum number of possible annotations at each node – the maximum possible gene counts – along the path follows a similar nested pattern. Denote the vector of the maximum possible annotation counts along the path: $\mathbf{N} = (N_0, N_1, \ldots N_l)^t$. Again, we have $N_{i+1} \leq N_i$ because of the nested structure of annotations along the path.
- The annotation overlap – or genes in common between ancestral nodes and descendant nodes – for genes in our list along the path is denoted $X_{i,i+1}$. Again, because all annotations at the child are shared by the parents, we have $X_{i,i+1} = X_i - X_{i+1}$.
- The universe of possible gene annotations along a path also have an overlap structure. For a parent node with count N_i, denote the number of gene annotations not shared by the child node as $N^c_{i,i+1}$.

With this notation we can write down the joint distribution of counts along an unbranched path under the null model that the GO annotated genes in our list are randomly sampled from the universe of genes with annotations:

$$P\left(\mathbf{X} = \mathbf{x}\right) = \frac{\begin{pmatrix} N_l \\ x_l \end{pmatrix}\begin{pmatrix} N_{p-1} - N_l \\ x_{p-1} - x_l \end{pmatrix}\cdots\begin{pmatrix} N_0 - N_1 \\ X_0 - X_1 \end{pmatrix}}{\begin{pmatrix} N \\ X_0 \end{pmatrix}} \tag{0.4}$$

As with the counts at a single node, the without-replacement exact null distribution for counts along an unbranched path can be approximated by a sampling with replacement model, and in this case the appropriate distribution to consider is the multinomial distribution. It is still important to consider the overlap in counts between parent and child nodes.

Beyond the joint distribution given above, the exact covariance structure for counts along an unbranched path can also be analyzed. For an ancestral node X_a and a descendant node X_d linked by an unbranched path, the covariance between X_a and X_d is:

$$Cov(X_a, X_d) = \frac{N_d}{N_a}\sigma_a^2 \tag{0.5}$$

This formula can be arrived at using the definition of covariance and the joint distribution for the node pair. An identity for summing binomial coefficients is important to simplify the joint expectation of counts. In the expression above, σ_a^2 is the variance of X_a, N_a is the count of total annotations to the ancestral node and N_d is the count of total annotations to the descendant node.

Since both the mean and covariance matrix for counts along an unbranched path are directly calculable, it is possible to arrive a simple statistic to calculate the total enrichment along an unbranched path. If we consider a single unbranched path of nodes, denoted p, and we denote the vector of counts along the path as $\mathbf{X_p}$ with expected counts $\mu_\mathbf{p}$, then enrichment statistic for the full path is the quadratic form:

$$Q_p = \left(\mathbf{X_p} - \mu_\mathbf{p}\right)^\mathbf{t}\mathbf{\Sigma_p^{-1}}\left(\mathbf{X_p} - \mu_\mathbf{p}\right) \tag{0.6}$$

The distribution of Q_p can be approximated with a chi-square distribution on l degrees of freedom, where l is the length of the path. It should be noted that in circumstances where GO counts are completely inherited between a parental node and a descendant node so that $\frac{N_d}{N_a} = 1$, a linear dependence will be introduced into the path counts. In this circumstance, one of the completely dependent counts should be removed from the vector $\mathbf{X_p}$ so that the covariance matrix is non-singular.

Bivariate Distribution of Arbitrary GO Terms

Before we proceed to describe methods for analyzing multiple lists, it is useful to discuss the bivariate distribution for counts at a pair of GO nodes related by a branched relationship. Without loss of generality, label the two nodes A and B, and denote the two counts X_A and X_B and denote by X_{AB} the number of genes in the list which are shared by nodes A and B . The distribution can be analyzed by taking expectations over the number of common genes.

$$f\left(x_A, x_B\right) \quad = \quad E_{X_{AB}}\left(f\left(x_A, x_B | X_{AB}\right)\right) \tag{0.7}$$

In the expression $f\left(x_A, x_B\right)$ is the bivariate mass function. The difficulty in the expression is the overlap X_{AB}. The overlap is potentially different for all pairs of nodes. Although the bivariate mass function is simple to write, summaries such as the covariance of X_A and X_B must be calculated separately for each pair and cannot easily be expressed in a simple formula such as (0.5).

The joint distribution for an arbitrary pair of nodes, applying the expression (0.7) is given by:

$$\sum_{x_{AB}=0}^{min(x_A, x_B, N_{AB})} \frac{\begin{pmatrix} N_{AB} \\ x_{AB} \end{pmatrix}\begin{pmatrix} N_A - N_{AB} \\ x_A - x_{AB} \end{pmatrix}\begin{pmatrix} N_B - N_{AB} \\ x_B - x_{AB} \end{pmatrix}\begin{pmatrix} N - (N_A + N_B) + N_{AB} \\ X_0 - (x_A - x_B) + x_{AB} \end{pmatrix}}{\begin{pmatrix} N \\ X_0 \end{pmatrix}}$$

0.4.2 Multiple Lists

With the tools for analyzing single lists in hand, the focus can be turned to multiple lists. The GO provides a content-based mechanism to compare lists that differ in size, platform or species origin. In the multi-list context, a useful goal is to create visual summaries which describe the relationships between lists based on their GO tabulation. Visual summaries reveal broad scale structure between results and can be useful to orient further work. An additional goal is to create a testing framework to identify the GO categories with unusual enrichment across a cohort of lists.

Exploratory Data Analysis and List Distances

The most direct way to compare a collection of lists is to count the overlap in the members of the lists. Unfortunately, counting list overlaps is problematic for several reasons. First, the lists may differ markedly in size, so that list intersection is an asymmetric operation. Second, all lists generated from genome scale experiments suffer from imperfect statistical power and false negative results. The imperfect power of the list generation mechanism will inevitably result in missed genes which should be included in the results, and these genes will fail to appear in the intersection. Generally speaking, list intersection results in a geometric loss of power because the power to detect genes in both lists is the product of the power of the separate list generation schemes. Finally, all lists suffer from false positive contamination. No matter what rule was used to generate the lists, the likelihood is that members of the lists have been erroneously included.

Fortunately, the distributional results described in the previous section provide a more coherent approach to compare lists. The bivariate distribution described by (0.7) together with the explicit covariance formula given by (0.5) suggest a distance metric. Denote by $\mathbf{X}_{\mathbf{g_i}}$ the vector of counts at all GO nodes for list g_i, and the analogous quantity for list g_j. We can derive both the expectation for the counts: $\mu_{\mathbf{g_i}} = \mathbf{E}\big(\mathbf{X}_{\mathbf{g_i}}\big)$, $\mu_{\mathbf{g_j}} = \mathbf{E}\big(\mathbf{X}_{\mathbf{g_j}}\big)$ and the covariance matrices for counts $\mathbf{\Sigma}_{\mathbf{g_i}}$ and $\mathbf{\Sigma}_{\mathbf{g_j}}$. A natural choice of distance between the lists is then the Malhalanobis distance between $\mathbf{X}_{\mathbf{g_i}}$ and $\mathbf{X}_{\mathbf{g_j}}$. If we set $\mathbf{Y}_{\mathbf{g_i}} = \mathbf{X}_{\mathbf{g_i}} - \mu_{\mathbf{g_i}}$ and $\mathbf{Y}_{\mathbf{g_j}} = \mathbf{X}_{\mathbf{g_j}} - \mu_{\mathbf{g_j}}$ and $\mathbf{\Sigma}_* = \mathbf{\Sigma}_{\mathbf{g_i}} + \mathbf{\Sigma}_{\mathbf{g_j}}$ we have

$$D(g_i, g_j) \quad = \quad \big(\mathbf{Y}_{\mathbf{g_i}} - \mathbf{Y}_{\mathbf{g_j}}\big)^{\mathbf{t}} \mathbf{\Sigma}_*^{-\mathbf{1}} \big(\mathbf{Y}_{\mathbf{g_i}} - \mathbf{Y}_{\mathbf{g_j}}\big) \tag{0.8}$$

This distance metric is still somewhat difficult to compute. A simpler alternative is to disregard the off diagonal elements of $\mathbf{\Sigma}_{\mathbf{g_i}}$ and $\mathbf{\Sigma}_{\mathbf{g_j}}$. In this case we have a simple distance metric:

$$D(g_i, g_j) \quad = \quad \sum_{\text{all nodes}_k} \big(Z_{g_i,k} - Z_{g_j,k}\big)^2 \tag{0.9}$$

The $Z_{l,k}$ in the expression are given by (0.3). Both the distance metric (0.8) and (0.9) are an improvement over list intersections.

Tests for GO Enrichment Across a Family of lists

Analysis also suggests methods for testing list counts across a family of lists. If we suppose there are k lists under consideration, then analysis of counts across lists should determine whether certain categories are collectively enriched or suppressed across a family of lists.

The first method we consider is a multi-way exact test for counts at a single GO term. A large body of literature exists concerning exact analysis for counts in many 2-way tables. If we recall the representation for counts at a single GO term as given in (0.4), then the method described in [11] can be used to derive an overall measure of significance of counts across the k tables. Briefly, the method enumerates the total number of two-way tables with marginal totals equal to those observed across the k lists. A network algorithm is employed to determine the total number of such tables. The central quantity of interest is the total number of successes at GO term j:

$$S_j = \sum_{i=1}^{k} X_{i,j} \tag{0.10}$$

The work in [11] gives a method to compute the mass function of S_j. For any GO term with an observed total count of s hits across k lists, the p-value for counts is $P(S_j \geq s)$. This p-value should be corrected for the multiplicity of p-values generated when considering many GO terms.

If we recall the formula for measuring total path enrichment given in (0.6), then a simple method presents itself for combining information across k lists. The approach is simply to sum the scores determined by (0.6).

$$T_p = \sum_{i=1}^{k} Q_{i,p} \tag{0.11}$$

$$\tag{0.12}$$

Under the assumption described in the section on a single path score (0.6), each $Q_{i,p}$ will have an approximately chi-square distribution on l degrees of freedom. If we assume the k lists are derived independently the quantity T_p will have an approximate chi-square distribution on lk degrees of freedom. When multiple lists are derived from the same study, those lists will be dependent, violating the assumptions for the chi-square approximation. As long as the lists are small relative to the number of possible genes, the dependence will be weak and the approximation is useful.

0.5 Stem Cell Data

In this section we consider meta-analysis of microarray stem cell data from mouse stem cell experiments. The stem-cells being investigated include hematopoetic, skin, hair, neural and cells of embryonic origin. The goal in the analysis is to identify commonalities and differences among the stem cell types.

0.5.1 Background on Stem Cells

Stem cells are special cells which appear in all multi-cellular organisms and which have the potential to give rise to descendant lineages of more specialized cells. The specialized cells are necessary in multi-cellular organisms to perform particular functions or to participate in distinct organ systems. The ability of cells to specialize is remarkable because, by and large, all cells in any organism share the exact same DNA and therefore have the same genetic information. The process by which cells attain their specific activities is termed differentiation. Stem cells are the ancestral cells in the differentiation process. These special cells have the potential to give rise to a variety of different descendant cell lineages. Stem cells are highly abundant in the early growth and development of organisms, but they disappear as organisms mature and then age. Interestingly, reservoirs of stem cells are present in many if not all tissues of adult organisms, and these cells are called adult stem cells. Unfortunately, adult stem cells generally have far less potential to generate different types of descendant cells than embryonically derived stem cells which are, in principle, capable of generating all differentiated cell types. Because of the great potential of stem cells for generating replacement of damaged or lost tissue, much experimental effort now focuses on characterizing stem cell properties. Interest focuses on finding properties shared by all stem cells as well as on identifying particular features which make one stem cell population different from another.

0.5.2 Example Studies

We consider data from 5 published studies. A total of 13 gene lists are present in this collection, and these lists are all derived from Affymetrix microarray experiments on the mgu74av2 platform. The lists range in size from approximately 2000 genes to as few as about 10. Before we present analysis of these data, we take time to consider the scope of the various individual studies.

Ivanova

One of the first stem cell papers presented was by Ivanova in 2002 [10]. This paper describes analysis of transcriptional profiles of stem cells from the hematopoetic system. The paper consideres 4 types of cells: long term hematopoetic stem cells (LT-HSC), short term hematopoetic stem cells (ST-HSC), and their descendant cells early progenitors and late progenitors. Although the results of this study are complex, we consider 4 lists derived from the experiment which are pertinent to the HSC system. Long term HSC (203), Short term HSC (10), Early progenitors (134), Intermediate progenitors (44), Late progenitors (182).

Ramalho-Santos

The paper by Ramalho-Santos [13] from the Melton laboratory in Boston was one of the first attempts to identify a common transciptional profile across different types of stem cells. The paper considered 3 distinct types of stem cells in mice: embryonic stem cells (ESC), neural stem cells (NSC), and hematopoetic stem cells (HSC). The general experimental approach for the NSC and HSC cell types was to compare the transcriptional profile of the stem cell population to the profile of differentiated cells from the same system. Assays were performed using the Affymetrix mgu74av2 microarray. For NSCs, the reference population was taken to be lateral ventricle tissue from the brain. For HSC the reference population was taken to be whole bone marrow. For the ESC, which are ancestral to all mouse cell types, the reference was taken to be the average of the brain and bone marrow differentiated reference. Gene lists were derived for each stem cell type based on Affymetrix present/absent calls and on estimated fold changes between the stem cell population and the derived cells. No multiple testing corrections were made, and the lists are relatively large. The ESC list consisted of 1787 probe sets which are reduced to approximately 1335 annotated genes in our analysis. The HSC list is also large, consisting of 1977 probe sets which are reduced to 1479 annotated genes for our analysis. The NSC list is largest of all, consisting of 2458 probe sets which reduce to 1807 genes for our analysis.

Venezia

The paper by Venezia et al [15] concerns HSC only. In this paper, a time course experiment is performed in which quiescent HSC are stimulated to proliferate by treatment at time 0 with the powerful cytotoxic agent 5fu. The drug 5fu is a commonly used chemotherapy agent used to treat cancer. In this experiment, 5fu is being used to stimulate the normally quiescent HSC to divide; the hope is that the process of stimulation and return to quiescence will reveal important transcriptional properties of the normally quiescent HSC. In this study mRNA samples are collected from cells at days 0, 1,2,3,6,10,and 30 after treatment with 5fu. Peak cell division occurs between days 6 and 10. Expression curves were fit to each gene across the timecourse, and analysis revealed 2 major types of expression profiles. Some genes demonstrated repressed expression across the proliferative phase of the timecourse, and these genes are termed the quiescence signature or Qsig; there are 225 genes assigned to this class. Other genes have elevated expression across the timecourse, and these genes are termed the proliferation signature or Psig. There are approximately 265 genes in this cohort. As with the other studies, this experiment was performed on the Affymetrix mgu74av2 microarray platform.

Morris

Hair stem cells are of interest in the treatment of hair loss and other disorders of the hair and skin. The Morris paper [12] describes the isolation of putative stem cells from a distinctive region of the hair follicle in the epithelium. Analysis revealed a genelist of 93 distinct murine genes with GO annotations, and these genes are the raw material for GO analysis.

Tumbar

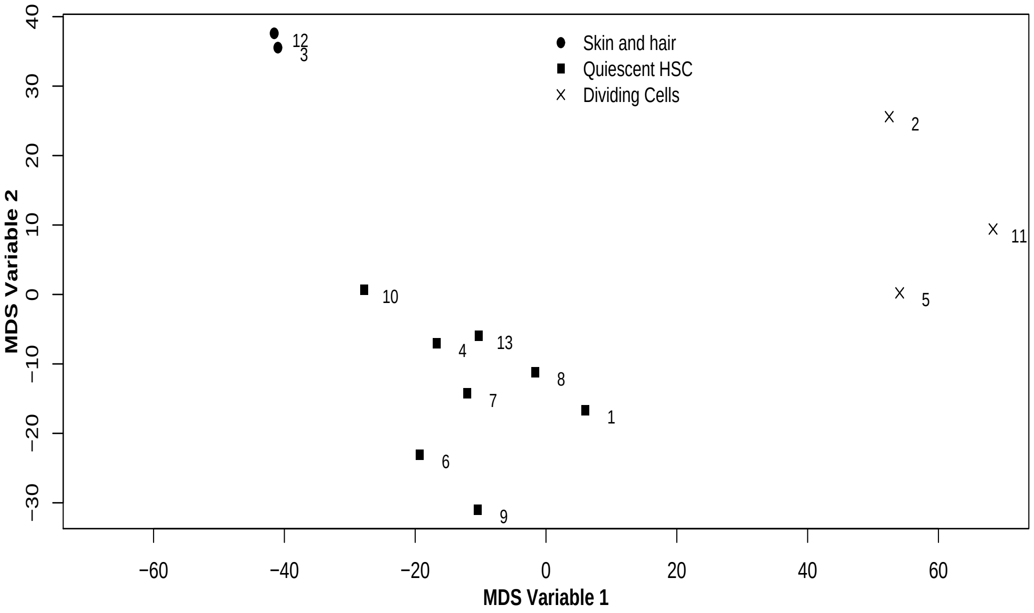
The paper by Tumbar et al.[14] is one of the first to describe skin stem cells in adult mice. The paper discusses the ambiguity surrounding skin stem cells, and it describes an experimental technique to derive the multi-potent cells from a slowly dividing compartment of the basal epithelium. The stem cells are thought to be found within a specialized area of the basal epidermis called the stem cell niche. The cells isolated in the paper are characterized by their ability to retain a flourescent label, and these cells are named “label retaining cells” or LRCs. The ability to retain the label shows the cells to be relatively quiescent. Affymetrix analysis of mRNA derived from these cells was performed using the Affymetrix mgu74av2 platform. These array data were compared against results from mRNAs derived from adjacent basal epidermal tissue. Analysis revealed a gene list of approximately 154 probe sets up-regulated in the LRCs when compared to more specialized, descendant cells from the basal epidermis (BL cells). Analysis shows that approximately 126 characterized mouse genes are represented by these 154 probe sets. In this chapter we consider the content of these 126 genes when analyzed against the geneontology data structure.

0.5.3 Meta-Analysis

The first step in the meta-analysis is to generate GO tabulations for each of the 13 lists available in the 5 studies. This step was accomplished using the R package developed in [2]. Once these results were computed, it is possible to proceed with exploratory analyses and then testing for enrichment to identify commonalities and differences between the lists.

Exploratory Analysis

To perform exploratory analysis a 13 × 13 distance matrix was constructed on the gene lists using the simple distance metric (0.9). A multi-dimensional scaling method was applied to this distance matrix to generate a bivariate scatter plot of the lists. The scatter plot is presented in figure 0.3.



1. Ivanova:HSC−EarlyProgenitor
2. Ramalho−Santos:ESC
3. Morris−HairFollicle
4. Ivanova:HSC−AdultandFetal
5. Venezia:HSC−PSig
6. Venezia:HSC−QSig
7. Ramalho−Santos:HSC
8. Ivanova:HSC−IntermediateProgenitor
9. Ivanova:HSC−LateProgenitor
10. Ivanova:HSC−LongTermT
11. Ramalho−Santos:NSC
12. Tumbar:Skin
13. Ivanova:HSC−ShortTerm

Figure 0.2 The figure shows a multi-dimensional scaling analysis of the 13 list distance matrix computed using (0.9). The results suggest the lists might be grouped into 3 families: skin and hair, quiescent HSC, and dividing stem cells.

Many useful features can be ascertained from the figure 0.3. First, there are three groups of lists. At least two of these groupings correspond to simple biological interpretations. The tightest group consists of the hair and skin. Both of these lists are derived from analysis of epithelial tissue. The other large group which correspond to a simple interpretation are the quiescent hematopoetic stem cells. Even lists of widely divergent size are grouped together in this cohort. The final group conists of the ESC list, the NSC list, and the Psig list from the Venezia study. The common biological thread among these three lists is that the genes in the lists represent actively dividing cell populations. The process of cell division strongly distinguishes this group from the other cohorts.

Commonalities and Differences

The statistics developed in (0.10) and (0.12) can be used to analyze the commonalities and differences between lists. To analyze the lists, we first group them into the cohorts made clear by the scatter plot (0.3). In each of the three groups we then perform the following analyses:

- Multi-way exact tests of each GO term in each of the three groups. This analysis results in a marginal p-value for each GO term in each group. In generating this analysis we also tabulate the total number of gene hits to each term within each group.
- Path enrichment analysis for each unbranched path in each group. Paths are enumerated by identifying all branches of the GO from the root term to terminal nodes.

Once the statistics have been calculated, the remaining task is to identify commonalities and differences between the lists. To identify commonalities we identify all those GO paths which are significantly enriched in all three groups. To make the analysis even more stringent, we make a further restriction to identify those GO classes which are also significantly enriched by marginal analysis of the term with the multi-way exact test.

The result of the commonality analysis suggests that there are some properties shared by these very different gene lists. All lists share enrichment for the cellular differentiation and cellular communication GO categories indicating that both the cell differentiation and the cell signaling system is a common feature of the stem cell lists. The regulation of transcription is also a strongly shared property of the lists.

In addition to examining commonalities, the results can also be used to distinguish differences between the groups. To distinguish lists, the following approach is useful:

- Identify GO paths which are significant in one list cohort, but which are less strongly enriched in the other two cohorts
- From amongst these significant GO paths, identify those nodes which have high gene counts and whose marginal p-values under the exact calculation are very strongly significant

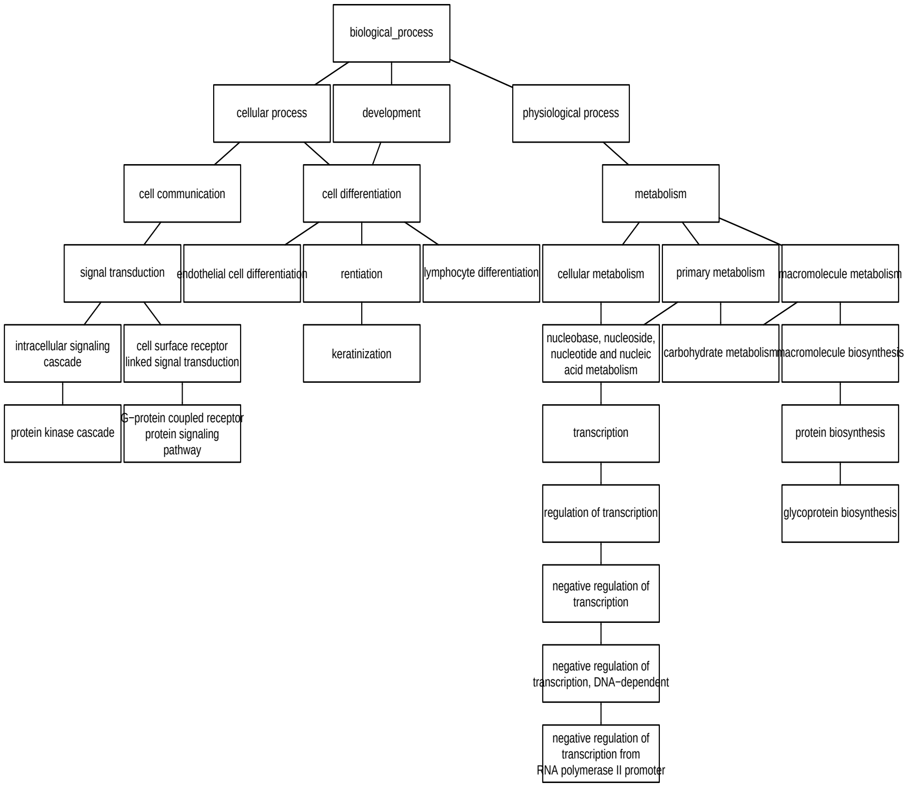

biological_process
cellular process
development
physiological process
cell communication
cell differentiation
metabolism
signal transduction
endothelial cell differentiation
rentiation
lymphocyte differentiation
cellular metabolism
primary metabolism
macromolecule metabolism
intracellular signaling cascade
cell surface receptor linked signal transduction
keratinization
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
carbohydrate metabolism
macromolecule biosynthesis
protein kinase cascade
G−protein coupled receptor protein signaling pathway
transcription
protein biosynthesis
regulation of transcription
glycoprotein biosynthesis
negative regulation of transcription
negative regulation of transcription, DNA−dependent
negative regulation of transcription from RNA polymerase II promoter

Figure 0.3 This figure demonstrates the content overlap amongst the three groups of gene lists. The analysis presented shows only results from the Biological Process ontology.

Analysis of the skin and hair lists (0.4) shows these to be meaningfully distinct from the other lists. Among the distinctions is the clear enrichment of these lists for the cell-migration and cell-motility system. This characteristic is interesting in light of the clear necessity for cell movement in skin and hair stem cells, as these cells must migrate out of their niche to produce descendants. The skin and hair lists also show enrichment for genes involved in cartiledge formation and in early neurogenesis. Both of these characteristics make sense in light of the fact that the skin cells arise from the same developmental lineage that gives rise to connective tissue and brain.

Analysis of the quiescent HSC lists (0.5) also show meaningful distinctions from the other lists. Among the distinctions is the clear enrichment of these lists for the cellular quiescence. Interestingly, the motility system is shut off in these cells. This characteristic is interesting in light of the fact that HSC by and large are not believed to migrate in order to generate their progeny. The HSC lists also show clear enrichment for the genes which are known to be involved in specification of the various lineages of cells descendant from the HSC. These GO categories suggest that the HSC cells have the potential to express all these lineage specifying genes.

Analysis of the final cohort of lists(0.6) shows a complex set of distinctions from the other groups. First and foremost, the genes in these lists reflect dividing cells, and there is a clear signature of cell division. The other distinctions are more difficult to decipher. Largely speaking, the genes in these lists reflect a heightened metabolic activity from the other 2 list groups. The metabolic activity also requires the cellular detoxification system, and these cellular processes are also shown to be enriched. It is not clear why immune response would be elevated in these lists.

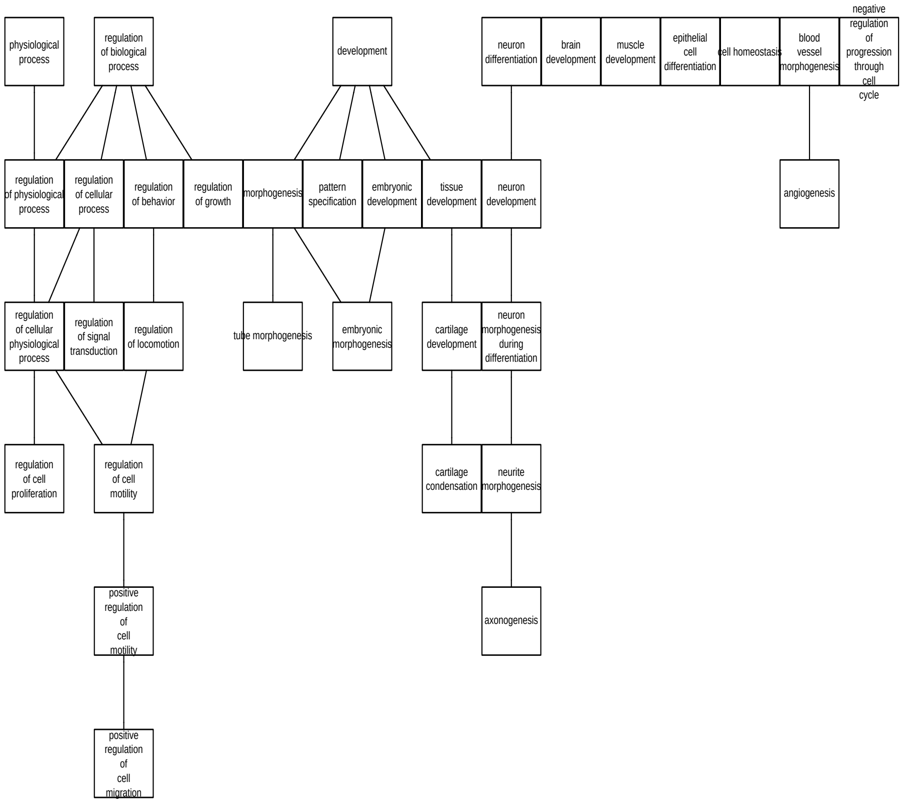

physiological process
regulation of biological process
development
neuron differentiation
brain development
muscle development
epithelial cell differentiation
cell homeostasis
blood vessel morphogenesis
negative regulation of progression through cell cycle
regulation of physiological process
regulation of cellular process
regulation of behavior
regulation of growth
morphogenesis
pattern specification
embryonic development
tissue development
neuron development
angiogenesis
regulation of cellular physiological process
regulation of signal transduction
regulation of locomotion
tube morphogenesis
embryonic morphogenesis
cartilage development
neuron morphogenesis during differentiation
regulation of cell proliferation
regulation of cell motility
cartilage condensation
neurite morphogenesis
positive regulation of cell motility
axonogenesis
positive regulation of cell migration

Figure 0.4 The figure depicts those subsets of the GO which strongly distinguish the skin and hair lists from the other cohorts. Only the BiologicalProcess ontology is depicted.

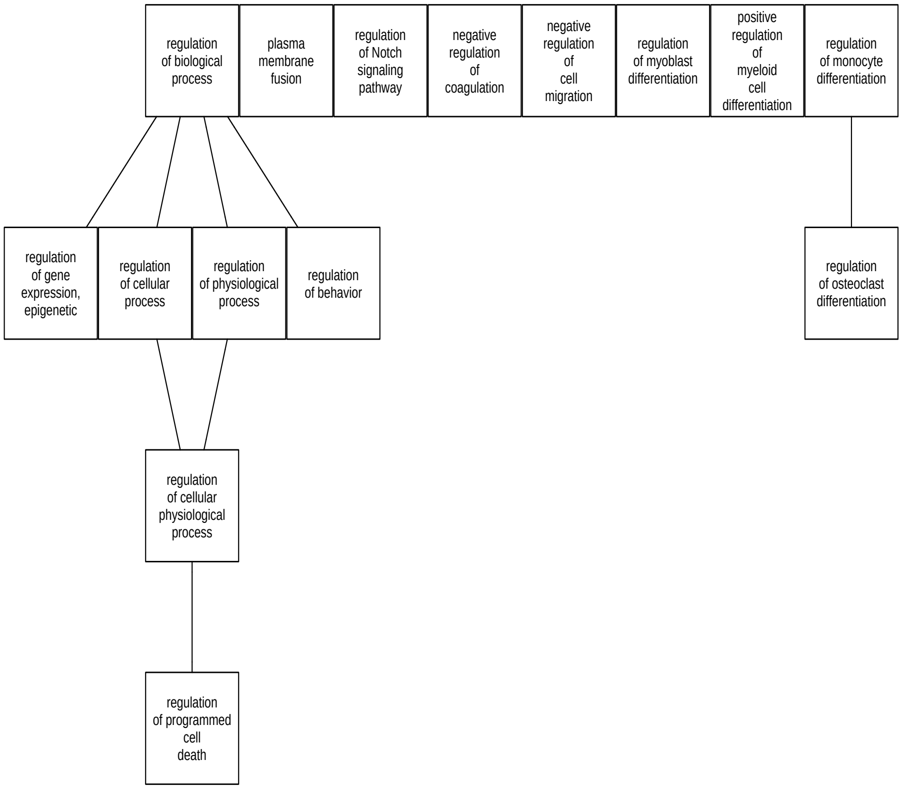

regulation of biological process
plasma membrane fusion
regulation of Notch signaling pathway
negative regulation of coagulation
negative regulation of cell migration
regulation of myoblast differentiation
positive regulation of myeloid cell differentiation
regulation of monocyte differentiation
regulation of gene expression, epigenetic
regulation of cellular process
regulation of physiological process
regulation of behavior
regulation of osteoclast differentiation
regulation of cellular physiological process
regulation of programmed cell death

Figure 0.5 The figure depicts those subsets of the GO which strongly distinguish the quiescent HSC lists from the other cohorts. Only the BiologicalProcess ontology is depicted

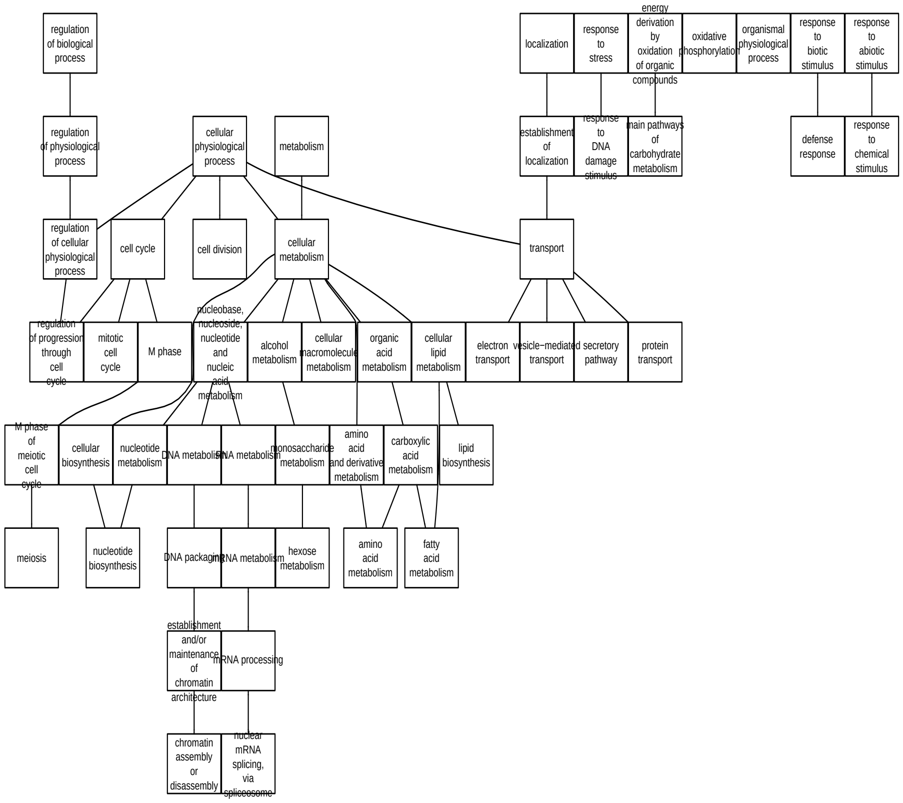

regulation of biological process
localization
response to stress
energy derivation by oxidation of organic compounds
oxidative phosphorylation
organismal physiological process
response to biotic stimulus
response to abiotic stimulus
regulation of physiological process
cellular physiological process
metabolism
establishment of localization
response to DNA damage stimulus
main pathways of carbohydrate metabolism
defense response
response to chemical stimulus
regulation of cellular physiological process
cell cycle
cell division
cellular metabolism
transport
regulation of progression through cell cycle
mitotic cell cycle
M phase
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
alcohol metabolism
cellular macromolecule metabolism
organic acid metabolism
cellular lipid metabolism
electron transport
vesicle−mediated transport
secretory pathway
protein transport
M phase of meiotic cell cycle
cellular biosynthesis
nucleotide metabolism
DNA metabolism
RNA metabolism
monosaccharide metabolism
amino acid and derivative metabolism
carboxylic acid metabolism
lipid biosynthesis
meiosis
nucleotide biosynthesis
DNA packaging
mRNA metabolism
hexose metabolism
amino acid metabolism
fatty acid metabolism
establishment and/or maintenance of chromatin architecture
mRNA processing
chromatin assembly or disassembly
nuclear mRNA splicing, via spliceosome

Figure 0.6 The figure depicts those subsets of the GO which strongly distinguish the proliferating lists from the other cohorts. Only the BiologicalProcess ontology is depicted.

0.6 Conclusions

This chapter considers GO based meta analysis of genome scale experiments. The GO is useful in this setting because it permits the joint analysis of lists which differ markedly in size, in experimental platform, or in species origin. The GO is a large controlled vocabulary for describing genes. A variety of statistics were described for analyzing cross tabulation of counts against the GO –both for single lists as well as for collections of lists. Example analysis using data from five stem cell studies shows the power of GO based analysis. Three distinct groups of gene lists are revealed, and these three lists are biologically meaningful. Deeper analysis using the GO reveals subtle but interesting differences between the lists. GO based data analysis is an interesting computational arena, and more work on GO based analysis remains to be done.

0.7 References

F. Al-Shahrour, F. Diaz-Uriarte, R., and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes.

Nathan Whitehouse Andrew Young, Joon Cho, and Chad Shaw. Ontologytraverser: an r package for go analysis.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, J. Michael Cherry Heather Butler, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Mateseand Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. Nature Genetics, 25:25–29, 2000.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, J. Michael Cherry Heather Butler, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Creating the gene ontology resource: Design and implementation.

Tim Berners-Lee, James, and Lassila Hendler. The semantic web. Scientific American, 2001.

G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki. David: Database for annotation, visualization, and integrated discovery.

T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing.

T.R. Gruber. A translational approach to portable ontologies. Knowledge Acquisition, 5:199–220, 1993.

T.R. Gruber. Formal ontology in information systems. IOS Press, Washington, D.C., 1998.

Natalia B. Ivanova, John T. Dimos, Christoph Schaniel, Jason A. Hackney, Kateri A. Moore, and Ihor R. Lemischka. A stem cell molecular signature. Science, 298:601–604, 2002.

Cyrus Mehta, Nitin Patel, and Robert Gray. Computing an exact confidence interval for the common odds ratio in several 2x2 contingency tables. Journal of the American Statistical Association, 80:969–973, 1985.

Rebecca J Morris, Yaping Liu, Lee Marles, Zaixin Yang, Carol Trempus, Shulan Li, Jamie S Lin, Janet A Sawicki, and George Cotsarelis. Capturing and profiling adult hair follicle stem cells. Nature Biotechnology, 22(4):411–417, 2004.

Miguel Ramalho-Santos, Soonsang Yoon, Yumi Matsuzaki, Richard C. Mulligan, and Douglas A. Melton. "stemness": Transcriptional profiling of embryonic and adult stem cells. Science, 298:597–600, 2002.

Tudorita Tumbar, Geraldine Guasch, Valentina Greco, Cedric Blanpain, William E. Lowry, Michael Rendl, and Elaine Fuchs. Defining the epithelial stem cell niche in skin. Science, 303:359–363, 2004.

Teresa A. Venezia, Akil A. Merchant, Carlos A. Ramos, Nathan L. Whitehouse, Andrew S. Young, Chad A. Shaw, and Margaret A. Goodell. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. PLoS Biology, 2(10):e301, 2004.