Artificial Intelligence in Medicine (2008) xxx, xxx–xxx

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Formulating and testing hypotheses in functional genomics

Joaquin Dopazo [a,b,c,*]

[a] Department of Bioinformatics, and Functional Genomics Node (INB), Valencia E-46013, Spain
[b] Centro de Investigación Príncipe Felipe, Valencia E-46013, Spain
[c] CIBER de Enfermedades Raras (CIBERER), Valencia, Spain

Received 8 February 2008; received in revised form 4 August 2008; accepted 4 August 2008

KEYWORDS
Functional profiling;
Functional enrichment;
Gene-set enrichment;
Gene ontology;
Functional genomics;
Transcriptomics

Summary

Objective: The ultimate goal of any genome-scale experiment is to provide a functional interpretation of the results, relating the available genomic information to the hypotheses that originated the experiment.
Methods and results: Initially, this interpretation has been made on a pre-selection of relevant genes, based on the experimental values, followed by the study of the enrichment in some functional properties. Nevertheless, functional enrichment methods, demonstrated to have a flaw: the first step of gene selection was too stringent given that the cooperation among genes was ignored. The assumption that modules of genes related by relevant biological properties (functionality, co-regulation, chromosomal location, etc.) are the real actors of the cell biology lead to the development of new procedures, inspired in systems biology criteria, generically known as gene-set methods. These methods have been successfully used to analyze transcriptomic and large-scale genotyping experiments as well as to test other different genome-scale hypothesis in other fields such as phylogenomics.

1. Introduction

A sentence attributed to Sydney Brenner, Nobel laureate in Medicine in 2002, states: "Progress in science depends on new techniques, new discoveries and new ideas, probably in that order." This sentence summarizes to perfection the impact of high-throughput technologies in biomedicine and other areas such as toxicology, agronomy, nutrition, etc. The deluge of data available from microarrays, proteomics, metabolomics, large-scale genotyping, etc., opened up new avenues to interrogate living systems at an unprecedented level of detail. However, the mere existence of this possibility does not mean that the questions will be addressed in a way that guarantees the proper answers. The use of data

* Correspondence address: Department of Bioinformatics, and Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, Valencia E-46013, Spain. Tel.: +34 963289680; fax: +34 963289701.
E-mail address: jdopazo@cipf.es.

0933-3657/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.08.003

coming from different ''omics'' technologies to try to better understand the biology of the cell configures the field of functional genomics. Nevertheless, formulating and testing hypotheses in functional genomics requires, in addition to technology, of new ways of addressing biological questions.

Microarrays can be considered the most paradigmatic among the methodologies used in functional genomics. Since the first proposals for the analysis of whole-genome gene expression based on microarrays in the late nineties [1] this technology has matured through a series of periods in which different interests were dominating. In retrospect, we can distinguish an initial period in which most microarray publications were dealing with issues such as reproducibility and sensitivity. Seminal papers dating from the late nineties were mainly proof-of-principle experiments [2,3]. As an anecdotic note it is worth mentioning that the methodological approaches used in such papers were mainly related to clustering. Paradoxically this has caused a subsequent confusion with respect to the choice of the appropriate methodology for a proper data analysis, as noted by some authors [4]. After this period, it was soon obvious that functional genomics experiments should be carefully analysed because many apparent associations happened merely by chance [5] and sensitivity became a main concern. In this context, different methods for the adjustment of *p*-values, which are considered standard today, started to be extensively used for differential gene expression analysis [6–8]. The consolidation of microarrays came from practical applications such as its use as predictors of clinical outcomes [9–11].

Although microarray experiments can be used to address a large variety of biological problems, scientific literature on this subject concentrates in three main types of objectives: ''class comparison,'' ''class prediction,'' and ''class discovery'' [12]. The first two objectives usually involve the application of tests to find genes with a significant differential expression, or the use of distinct procedures to predict class membership on the basis of the values observed for a number of ''key'' genes. Clustering methods belong to the last category, also known as unsupervised methods, because no previous information about the class structure of the data set is used in the study.

The functional interpretation of the experiments is typically made on the genes selected as relevant by one of these procedures. To this end, predefined modules of genes related among them by any interesting biological property (common function, regulation, chromosomal location, etc.) are used. Functional enrichment methods [13,14] are used to find if one or more of these gene modules is significantly over-represented among the relevant genes selected in the experiment. Over-representation of a given gene module means that genes with a particular property have been activated or deactivated in the experiment.

Obviously, the way in which the relevant genes are defined is implicitly conditioning the functional interpretation of the whole experiment. Paradoxically, many of the biological properties used to define gene modules (function, regulation, etc.) implies the existence of a high level of co-expression among them [15–17], while most of the tests used to select relevant genes assume independence in the behaviours of the genes imposing thus an artificial threshold with a unfavourable effect in the results [18]. Essentially, there is a basic inconsistency in the way functional hypothesis are tested by the functional enrichment approach.

Gene-set-based analysis came up as a response to the necessity of having a framework for testing hypothesis in agreement with the functional properties of the genes [13,19,20]. Pioneered by the gene-set enrichment analysis (GSEA) [21], a collection of methods aiming to test the behaviour of gene modules in opposition to gene enrichment methods that relied on previous gene-by-gene selection procedures has been published [13,20].

Finally, the ideas of using gene modules are being applied not only at the level of the final step of functional interpretation, but as constitutive parts of the testing procedures for supervised and unsupervised gene expression data analysis. This review describes the different approaches used for the functional interpretation of functional genomics experiments and the future trends in this field.

2. Definition of gene modules: sources of information

Any functional analysis relies on the definition of gene teams related by biological properties of interest. Here, some of the most commonly used sources of biological information are described.

Probably the most widely used source of definition of functional modules is the gene ontology (GO) catalogue [22]. GO represents the biological knowledge as a tree (more precisely as a directed acyclic graph (DAG) in which a node can have more than one parent) where functional terms near the root of the tree make reference to more general concepts while deeper functional terms near the leaves of the tree make reference to more specific concepts. If a gene is annotated to a given level then it is automatically considered to be annotated at all the upper levels (all the parent levels) up to the root. Since genes are

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

annotated at different levels of the GO hierarchy, it is common to use this abstraction to choose a predefined level in the hierarchy instead of using directly the original levels of annotation of the genes [23,24], which increases the power of the enrichment tests [23,25–27].

The Kioto encyclopedia of genes and genomes (KEGG) pathways database [28] or the Biocarta pathways (http://www.biocarta.com) are two extensively used sources of functional information. There are also databases that contain functional motifs mapped to proteins such as the Interpro database [29] and many other.

Modules do not necessarily need to be configured upon functional basis. Thus, transcriptional modules can be defined as groups of genes under the same regulatory control. Databases that collect regulatory motifs can be found. Among the most popular are: CisRed [30] and Transfact, that contains predictions of transcription factor binding sites [31]. Also negative regulation mediated by microRNAs has recently gained relevance. The miRBase [32] contains putative gene targets of such microRNAs. Genes sharing one or more of these regulatory motifs can be considered a putative regulatory module.

Other ways of defining modules of different nature include the use of information obtained using text-mining procedures [33], chromosomal location [34,35], protein–protein interactions, etc.

3. Functional enrichment methods

As previously commented, the final aim of a microarray experiment is to find a functional explanation at molecular level for any given macroscopic observation (e.g. what biological processes differentiate a healthy control from a diseased case, etc.). In the classical approach, known as functional enrichment analysis, the functional interpretation of microarray data is performed in two steps: in a first step genes of interest are selected using different procedures. In a subsequent step, the selected genes of interest are compared to a background (usually the rest of the genes) in order to find enrichment in any gene module. This comparison to the background is essential because an apparently high proportion of a given functional module could easily be nothing but a reflection of a high proportion of this particular module in the whole genome but not a proper enrichment. Actually, both enrichments and depletions of gene modules are potentially of interest. Therefore, unless there is a specific reason not to consider enrichment or depletion, two-sided tests are appropriate [36]. This comparison between the selected genes and the background can be carried out by means of the application of different tests, such as the hypergeometric, Fisher's exact test χ^2 and binomial, which are considered to give similar results [36]. Since many tests are conducted in order to check all the gene modules, adjustment for multiple testing, such as the false discovery rate [6] of others, must be used.

Another important aspect particular to gene modules defined using GO annotations is the way in which the structure of the ontology is taken into account. Many programs test each GO module independently, which do not respect dependencies between the GO terms. This constitutes a major drawback given that the true-path rule (each term in GO shares all the annotations of all of its descendants) is ignored in this case. Other programs partly circumvent the problem by selecting a particular level of the DAG and analyse the gene annotations at this level [37,38], use a "slim" ontology, that is a reduced set of terms with more informative content [39] or even try to find the optimal and more informative level for each case [40]. There are also more sophisticated approaches that try to decorrelate the GO graph structure by processing the GO DAG from most specific to least specific terms [41,42]. Other approaches to statistical analysis of GO term overrepresentation examine each term in the context of its parent terms (parent–child approach) in the DAG context [43].

Table 1 presents an exhaustive (although probably not complete) list of tools for functional profiling that implement tests for functional enrichment. Here the number of Scholar Google citations has been used as an approximate popularity index, given that it is reflecting the number of academic documents (mostly papers) citing a particular paper. Following this criterion, the most popular tools having over 200 citations are EASE [44], DAVID [45], GOMiner [46], Babelomics/FatiGO [26,27,37], MAPPFinder [47], GOStats [38] and Ontotools [48].

3.1. Problems with testing functional hypothesis in a gene-based framework

A drawback in the two-step functional enrichment analysis comes from the fact that the gene selection process applied in the first step does not take into account that these genes are acting cooperatively in the cell and that consequently their behaviour must be coupled to some extent. In this selection process, under the unrealistic simplification of independence among gene behaviours, stringent thresholds to reduce the false positives ratio in the results are

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

Table 1 Functional enrichment data analysis tools

Tool	Application type or URL for web servers	References	Citations[a]
EASE	Windows application	[44]	603
DAVID	http://www.DAVID.niaid.nih.gov	[45]	504
GOMiner	http://discover.nci.nih.gov/gominer/	[46,79]	408
Babelomics	http://www.babelomics.org	[26,27,37,40,66]	402
MAPPFinder	http://www.GenMAPP.org	[47]	379
FatiGO	http://www.fatigo.org	[37]	341
GOStat	http://gostat.wehi.edu.au/	[38]	249
Ontotools	http://vortex.cs.wayne.edu/ontoexpress/	[24,48,80–82]	223
GOTM	http://genereg.ornl.gov/gotm/	[83]	164
GO:: TermFinder	Perl script	[84]	152
FunSpec	http://funspec.med.utoronto.ca webcite	[85]	100
GeneMerge	http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html	[86]	96
FuncAssociate	http://llama.med.harvard.edu/Software.html	[87]	91
BINGO	Cytoscape plugin	[88]	75
GOToolBox	http://gin.univ-mrs.fr/GOToolBox	[39]	74
GFINDer	http://www.medinfopoli.polimi.it/GFINDer/	[89,90]	49
WebGestalt	http://bioinfo.vanderbilt.edu/webgestalt/	[91]	46
GOSurfer	R package	[92]	45
CLENCH	Perl script	[93]	26
Pathway Explorer	https://pathwayexplorer.genome.tugraz.at/	[94]	25
Ontology Traverser	R package	[95]	24
THEA	Java standalone	[96]	11
WebBayGO	http://blasto.iq.usp.br/~tkoide/BayGO/	[97]	10
GOStat	R package	[42]	10
eGOn	http://www.genetools.microarray.ntnu.no/egon/index.php	[98]	7
WholePathwayScope	Windows	[99]	6
FIVA	Java standalone	[100]	4
GENECODIS	http://genecodis.dacya.ucm.es/	[101]	3
SeqExpress	R package	[102]	3
G:Profiler	http://biit.cs.ut.ee/gprofiler/	[103]	2
PathExpress	http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress	[104]	–

[a] Citations are taken from Scholar Google (as of January 2008). Scholar Google is taken as an indirect estimation of the citation in papers but gives an idea on the impact in the scientific community.

usually imposed. As previously discussed, the biological properties used to define gene modules (function, regulation, etc.) sought in the functional enrichment test conducted in the second step entails dependences [15–17] which are ignored and mostly lost in below such threshold (see Fig. 1). So, there is a paradoxical incongruence in the way functional hypothesis are tested by the functional enrichment approach.

Indirectly, this fact lies in the core of one of the most common problems with molecular signatures or predictors: their instability. Variable selection with microarray data can lead to many solutions that are equally good from the point of view of prediction rates, but that share few common genes [49]. This multiplicity problem has been emphasized by several authors [50] and recent examples are shown in [49,51].

Another example that highlights the uncertainty derived from focusing on genes comes from the comparisons of platforms. While intra-platform reproducibility is high there are still some concerns about the cross-platform coherence of results [52]. Paradoxically, despite the fact that gene-by-gene results are not always the same, the comparative analysis of biological terms emerging from the different platforms are clearly consistent [53].

3.2. The biology behind

All these observations clearly show that attempts to link macroscopic observations (such as physiological responses or phenotypes) to genes, as their causative functional elements, involve enormous amounts of uncertainty. The most plausible explanation for this fact does not rely in the accuracy of

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

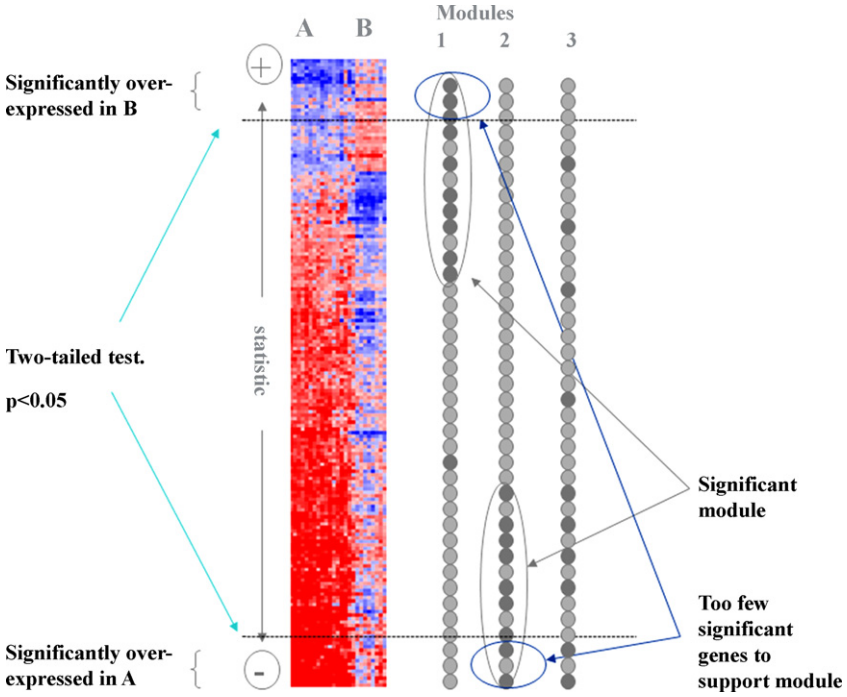

Figure 1 GSA strategy of analysis. Genes are arranged according to differential expression between the experimental classes A and B. The right side of the picture represents the location of the genes corresponding to three hypothetical modules across the arranged list. Module 1 is predominantly composed of genes with high expression in class B (red values corresponding to highest expression), but scarcely represented among genes highly expressed on class A. The behaviour of module 2 is exactly the opposite. Finally, module 3 is completely unrelated with the experiment because genes belonging to this module are expressed all across the list, without ant trend towards any of the extremes.

the testing methodology but in the fact that most probably the ultimate functional "bricks" of the cell are not the genes but more complex entities that we can represent by modules, composed of gene teams, which act cooperatively to carry out functions. Actually, an increasing corpus of evidence reveals that genes do not operate alone within the cell, but in an intricate network of interactions that we only recently start to envisage [54—56]. It is a well recognized fact that genes with similar overall expression often share similar functions [15—17] and, in fact, this causal relationship has been used to predict gene function from patterns of co-expression [15,57]. Therefore, the above observations are consistent with the hypothesis of modularly behaving gene programmes, where sets of genes are activated in a coordinated way to carry out functions. In this scenario, a different type of inference can be made based on testing hypothesis centred on modules of genes related by biological properties, instead of testing one gene at a time. From a clinical point of view, the validity of the traditional reductionist vision, in which one or a few key genes would be the causative factors of diseases [58] must be thoroughly revised. Then, it is imperative taking into consideration the functional dimension in the interpretation of genome-scale experiments. In this new scenario, the deregulation of modules of functionally related genes would be the real causative factor behind the disease phenotype [59].

4. Testing systems-based hypothesis with gene-set analysis (GSA) methods

The interpretation of a genome-scale experiment using the two-step functional enrichment approach is far from being optimal given that the thresholds imposed in the first step assuming independence precludes the detection of many gene modules. Methods directly inspired in systems biology focus on collective properties of the genes more than on individual gene expression values. Modules of genes related by common functionality, regulation or other interesting biological properties will simultaneously fulfil their roles in the cell and, consequently, they are expected to display a coordinated expression.

In its simplest formulation GSA method uses a rank of values derived from the experiment analysed. Mootha et al. [21] ranked the genes according to their differential expression when two predefined classes (diabetic versus normal controls) were comparing by means of any appropriate statistical test [60]. The position of the genes (that cooperatively

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

act to define modules) within this ranked list is related to its participation in the trait studied in the experiment. Consequently, each module that is a causative agent of the differences between the classes compared will be found in the extremes of the ranked list with highest probability. Fig. 1 summarizes this strategy. Thus, instead of testing differential activities of genes, which implicitly assumes independent behaviour (an aspect often ignored by the researchers applying the test), and later searching for enrichment in gene modules among the selected genes, GSA directly tests for gene modules significantly cumulated in the extremes of a ranked list of genes. In this way, artificial previous thresholds, which inadvertently change the meaning of our hypothesis testing schema, is avoided.

There are different methods that have been proposed for this purpose such as the GSEA [21,61] or the SAFE [62] methods, which use a non-parametrical version of a Kolmogorov–Smirnov test. Other strategies proposed are the direct analysis of functional terms weighted with experimental data [63] or model-based methods [64]. With similar accuracy although conceptually simpler and quicker methods have also been proposed, for instance the parametrical counterpart of the GSEA, the PAGE [65] or the segmentation test, Fatiscan [66].

There are two major strategies used in GSA to perform module tests: either proving that a module has a significantly altered expression compared to other groups or all remaining genes (competitive strategy) or proving that expression in a module is altered between different conditions of interest (self-contained strategy) [19,20]. In spite of some criticisms [67], the competitive strategy is more often used.

Beyond other technical or statistical considerations, the approximate level of acceptance of different GSA methods among the scientific community is reported in Table 2. More than the 75% of the Scholar Google citations are monopolized by two tools: GSEA and Babelomics.

4.1. Gene-set-based supervised and unsupervised analysis

In addition to be used as the final step of functional interpretation, gene modules can conceptually be used in problems of class comparison or class prediction (supervised) as well a in class discovery (unsupervised) problems.

Table 2 Tools available for functional profiling by gene-set analysis

Tool	Application type or URL for web servers	References	Test[a]	Citations[b]
GSEA	http://www.broad.mit.edu/gsea/	[21,61]	GS, C	1013
Babelomics (FatiGO + FatiScan)	http://www.babelomics.org	[26,27,37,40,66]	FE/GS, C	402
FuncAssociate	http://llama.med.harvard.edu/Software.html	[87]	FE/GS, C	91
Global test	R package	[64]	GS, SC	89
PAGE	Python script	[65]	GS, C	42
ErmineJ	Java	[105]	GS, C	35
FatiScan	http://www.babelomics.org	[66]	GS, C	34
GO-mapper	Windows, Perl script	[63]	GS, C	33
SAFE	R package	[62]	GS, C	27
GOAL	http://microarrays.unife.it	[106]	GS, C	25
Catmap	Perl script.	[107]	GS, C	19
PLAGE	http://dulci.biostat.duke.edu/pathways/	[108]	GS, SC	18
GODist	Mathlab program	[109]	GS, SC	17
t-profiler	http://www.t-profiler.org/	[110]	GS, C	12
JProGO	http://www.jprogo.de/	[111]	GS, C	7
ADGO	http://array.kobic.re.kr/ADGO	[112]	GS, C	3
GeneTrail	http://genetrail.bioinf.uni-sb.de/	[113]	GS, C	3
ASSESS	Java	[114]	GS, C	2
DEA	R package	[115]	GS, C	1
GlobalANCOVA	R package	[67]	GS, SC	1
GAZER	http://integromics.kobic.re.kr/GAzer/index.faces	[116]	GS, C	–
SAM-GS	Windows excel add-in	[117]	GS, SC	–

[a] Type of test: GS: gene set; C: competitive, SC: self-contained; FE: functional enrichment.
[b] Citations are taken from Scholar Google (as of January 2008). Scholar Google is taken as an indirect estimation of the citation in papers but gives an idea on the impact in the scientific community.

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

As described above, existing gene selection or class prediction methods treat all the genes equally, ignoring biological knowledge of gene functions. New methods attempt to exploit such prior information using functional modules based on GO or KEGG annotations. Thus, a recent proposal is based on the GO hierarchical structure where child nodes have more specific function definitions while its parent node has a more general one. The method works by initially building a separate classifier for each node using a conventional method (e.g. shrunken centroids, support vector machines, etc.), then propagating their classification results by a weighted sum to their parent nodes, where the weights are related to the performance of the classifiers [68]. An alternative proposal uses a modified boosting method called non-parametric pathway-based regression [69] were genes are first partitioned into several groups or pathways then in boosting, only pathway-specific new classifiers were obtained. A similar approach uses random forests to rank biological pathways in regression and classification [70]. Finally this concept has been generalized as a penalized approach that can be applied to any classifier [71].

In these lines, recent proposals make use of biological information [30,72] or phenotypic information [73] as a constitutive part of clustering algorithms.

4.2. Gene-set analysis in genotyping

Another field in which a gene-set-based approach could be very useful is genotyping. Association and linkage studies with chips with increasingly density result in a frustrating effect of decrease in the power of the tests, given to the strict corrections that must be applied to the tests. Most genetic disorders have a complex inheritance and can be considered the combined result of variants in many genes, each contributing only weak effects to the disease. Given that in any disorder, most of the disease genes will be involved in only a few different molecular pathways, the knowledge of the relationships (functional, regulatory, interactions, etc.) between the genes can help in the assessment of possible candidates (which may reside in different loci) with a joint basis for the disease etiology. The use of different gene module definitions (GO, KEGG, protein interactions and co-expression) in an integrated network was recently applied to interrelate positional candidate genes from different disease loci and then to test 96 heritable disorders for which the Online Mendelian Inheritance in Man database [74]. This gene-set-based strategy resulted in a 2.8-fold increase over random selection.

4.3. Gene-set analysis in evolution

Many other hypotheses can be tested on gene module basis beyond functional genomics. Evolution is a paradigmatic field in which gene-based analysis has given results noticeably below the expectations. In the last years, several papers have been published that attempted to elucidate the intricacies of human evolution by means of comparing rate differences and positive selection in human genes against their homologues in other fully sequenced species [75—78]. Nevertheless, beyond some conjectures, no statistical significant conclusions about the functionality of the genes under different types of selective pressures could be derived from these studies. One possible explanation for the failure in finding a functional interpretation to the human evolution comes, most probably, from the fact that these studies followed an inefficient functional enrichment, two-step approach.

Mutations occur on single genes but natural selection acts on phenotypes by operating on whole subcellular systems. Mutations in genes either remain finally fixed or disappear because of their beneficial or disadvantageous effect, respectively. This effect can only be understood in the context of the module (e.g. pathways, gene ontology terms, etc.) in which such genes are involved. If a list of genes arranged by some parameter that accounts for their evolutionary rates is examined, it is expectable that gene modules favoured or disfavoured by selection will tend to appear towards the extremes. Recently, in a study carried out under this viewpoint the ratio of non-synonymous to synonymous mutations, a widely accepted measure of selection, was used to rank genes [25]. A GSA method, the FatiScan [66], was applied to the ranked list of genes and the following GO terms significantly cumulated at the extreme of the distribution corresponding to the highest selective values: sensory perception of smell (GO:0007608), sensory perception of chemical stimulus (GO:0007606) and G-protein coupled receptor protein signalling pathway (GO:0007186). Previous publications pointed out to these GO categories as positively selected but it could not be properly demonstrated [76,77]. This application shows how the concept of GSA can be extended to other domains beyond functional genomics.

5. Conclusions

The functional interpretation of the experiments is typically made on the genes selected as relevant by applying tests that assume independence. Predefined modules of genes related among them by any

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

interesting biological property (common function, regulation, chromosomal location, interaction of the corresponding proteins, etc.) are used for this purpose. Functional enrichment methods [13,14] are used to find if one or more of these gene modules are significantly over-represented among the relevant genes selected in the experiment. Such over-representation would indicate that genes with a particular property have been activated or deactivated in the experiment. Paradoxically, the tests used to select such relevant genes implicitly assume independence in their behaviours, while the biological properties used to define gene modules (function, regulation, etc.) implies the existence of a high level of cooperation among them [15–17]. This fact imposes an artificial threshold that results in a unfavourable effect in the efficiency of this approach [18]. Essentially, there is a basic incongruence in the way functional hypothesis are tested by these means.

GSA methods provide an elegant and efficient alternative that offers a framework for testing hypothesis in agreement with the functional properties of the genes [13,19,20]. Pioneered by the GSEA [21], different methods that intend to test the activities of gene modules instead of testing genes isolates from their context have been proposed [13,19,20]. Such methods have been successfully applied to the analysis of transcriptomes [21,25] but also in large-scale genotyping [74] or in genome-scale evolutionary studies [25]. In a near future, GSA methods will probably become integral parts of gene selection prediction methods.

Acknowledgements

This work is supported by grants from projects BIO2005-01078 from the Spanish Ministry of Education and Science and National Institute of Bioinformatics (www.inab.org), a platform of Genoma España. The CIBER de Enfermedades Raras (CIBERER) is an initiative of the ISCIII.

References

[1] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–8.
[2] Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999;96:9212–7.
[3] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
[4] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
[5] Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 2003;19:551–60.
[6] Benjamini Y, Yekutieli D. The control of false discovery rate in multiple testing under dependency. Ann Stat 2001;29:1165–88.
[7] Storey JD, Tibshirani R. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 2003;100:9440–5.
[8] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368–75.
[9] Quackenbush J. Microarray analysis and tumor classification. N Engl J Med 2006;354:2463–72.
[10] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[11] Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. Br J Cancer 2003;89:1599–604.
[12] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
[13] Dopazo J. Functional interpretation of microarray experiments. Omics 2006;10:398–410.
[14] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21:3587–95.
[15] Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res 2002;12:1703–15.
[16] Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res 2004;14:1085–94.
[17] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.
[18] Pan KH, Lih CJ, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. Proc Natl Acad Sci USA 2005;102:8961–5.
[19] Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007;23:980–7.
[20] Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform 2008;9:189–97.
[21] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–73.
[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
[23] Al-Shahrour F, Dopazo J. Ontologies and functional genomics. In: Azuaje F, Dopazo J, editors. Data analysis and visualization in genomics and proteomics. Chichester, West Sussex, UK: Wiley; 2005. p. 99–112.
[24] Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. Recent additions and improvements to the Onto-Tools. Nucleic Acids Res 2005;33:W762–5.
[25] Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, Minguez P, Montaner D, et al. From genes to functional classes in the

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

study of biological systems. BMC Bioinformatics 2007;8:114.
[26] Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, et al. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res 2006;34:W472—6.
[27] Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Nucleic Acids Res 2005;33:W460—4.
[28] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277—80.
[29] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33:D201—5.
[30] Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, et al. cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res 2006;34:D68—73.
[31] Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 2000;28:316—9.
[32] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006;34:D140—144.
[33] Minguez P, Al-Shahrour F, Montaner D, Dopazo J. Functional profiling of microarray experiments using text-mining derived bioentities. Bioinformatics 2007;23:3098—9.
[34] Conde L, Montaner D, Burguet-Castell J, Tarraga J, Al-Shahrour F, Dopazo J. Functional profiling and gene expression analysis of chromosomal copy number alterations. Bioinformation 2007;1:432—5.
[35] Conde L, Montaner D, Burguet-Castell J, Tarraga J, Medina I, Al-Shahrour F, et al. ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling. Nucleic Acids Res 2007;35:W81—5.
[36] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007;23:401—7.
[37] Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20:578—80.
[38] Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004;20:1464—5.
[39] Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol 2004;5:R101.
[40] Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, et al. FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res 2007;35:W91—96.
[41] Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006;22:1600—7.
[42] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics 2007;23:257—8.
[43] Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics 2007;23:3024—31.
[44] Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4:R70.
[45] Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:P3.
[46] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003;4:R28.
[47] Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 2003;4:R7.
[48] Khatri P, Bhavsar P, Bawa G, Draghici S. Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. Nucleic Acids Res 2004;32:W449—56.
[49] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171—8.
[50] Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19:1484—91.
[51] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006;103:5923—8.
[52] Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M. Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 2003;19:570—7.
[53] Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, et al. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2005;2:351—6.
[54] Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006;124:47—59.
[55] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein—protein interaction network. Nature 2005;437:1173—8.
[56] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein—protein interaction network: a resource for annotating the proteome. Cell 2005;122:957—68.
[57] van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. Trends Genet 2003;19:238—42.
[58] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37(Suppl):S11—7.
[59] Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nat Biotechnol 2004;22:1249—52.
[60] Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 2003;4:210.
[61] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545—50.
[62] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21:1943—9.
[63] Smid M, Dorssers LC. GO-Mapper: functional analysis of gene expression data using the expression level as a score

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

to evaluate Gene Ontology terms. Bioinformatics 2004;20:2618–25.
[64] Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004;20:93–9.
[65] Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 2005;6:144.
[66] Al-Shahrour F, Diaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 2005;21:2988–93.
[67] Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics 2008;24:78–85.
[68] Lottaz C, Spang R. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 2005;21:1971–8.
[69] Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics 2007;8:265–84.
[70] Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics 2006;22:2028–36.
[71] Tai F, Pan W. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics 2007;23:1775–82.
[72] Pan W. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics 2006;22:795–801.
[73] Jia Z, Xu S. Clustering expressed genes on the basis of their association with a quantitative phenotype. Genet Res 2005;86:193–207.
[74] Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 2006;78:1011–25.
[75] Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, et al. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 2006;34:W621–5.
[76] Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, et al. Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. Science 2003;302:1960–3.
[77] Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 2005;3:e170.
[78] The-chimpanzee-sequencing-and-analysis-consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005;437:69–87.
[79] Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, et al. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC Bioinformatics 2005;6:168.
[80] Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Res 2003;31:3775–81.
[81] Khatri P, Desai V, Tarca AL, Sellamuthu S, Wildman DE, Romero R, et al. New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate. Nucleic Acids Res 2006;34:W626–31.
[82] Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, et al. Onto-Tools: new additions and improvements in 2006. Nucleic Acids Res 2007;35:W206–11.
[83] Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004;5:16.
[84] Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004;20:3710–5.
[85] Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics 2002;3:35.
[86] Castillo-Davis CI, Hartl DL. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 2003;19:891–2.
[87] Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. Bioinformatics 2003;19:2502–4.
[88] Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21:3448–9.
[89] Masseroli M, Galati O, Pinciroli F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res 2005;33:W717–23.
[90] Masseroli M, Martucci D, Pinciroli F. GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res 2004;32:W293–300.
[91] Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33:W741–8.
[92] Zhong S, Storch KF, Lipan O, Kao MC, Weitz CJ, Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. Appl Bioinformatics 2004;3:261–4.
[93] Shah NH, Fedoroff NV. CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. Bioinformatics 2004;20:1196–7.
[94] Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res 2005;33:W633–7.
[95] Young A, Whitehouse N, Cho J, Shaw C. OntologyTraverser: an R package for GO analysis. Bioinformatics 2005;21:275–6.
[96] Pasquier C, Girardot F, Jevardat de Fombelle K, Christen R. THEA: ontology-driven analysis of microarray data. Bioinformatics 2004;20:2636–43.
[97] Vencio RZ, Koide T, Gomes SL, Pereira CA. BayGO: Bayesian analysis of ontology term enrichment in microarray data. BMC Bioinformatics 2006;7:86.
[98] Beisvag V, Junge FK, Bergum H, Jolsum L, Lydersen S, Gunther CC, et al. GeneTools–application for functional annotation and statistical hypothesis testing. BMC Bioinformatics 2006;7:470.
[99] Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM. WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. BMC Bioinformatics 2006;7:30.
[100] Blom EJ, Bosman DW, van Hijum SA, Breitling R, Tijsma L, Silvis R, et al. FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. Bioinformatics 2007;23:1161–3.

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003

[101] Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 2007;8:R3.
[102] Boyle J. SeqExpress: desktop analysis and visualization tool for gene expression experiments. Bioinformatics 2004;20:1649—50.
[103] Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 2007;35:W193—200.
[104] Goffard N, Weiller G. PathExpress: a web-based tool to identify relevant pathways in gene expression data. Nucleic Acids Res 2007;35:W176—81.
[105] Lee HK, Braynen W, Keshav K, Pavlidis P. ErmineJ: tool for functional analysis of gene expression data sets. BMC Bioinformatics 2005;6:269.
[106] Volinia S, Evangelisti R, Francioso F, Arcelli D, Carella M, Gasparini P. GOAL: automated Gene Ontology analysis of expression profiles. Nucleic Acids Res 2004;32:W492—9.
[107] Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinformatics 2004;5:193.
[108] Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 2005;6:225.
[109] Ben-Shaul Y, Bergman H, Soreq H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics 2005;21:1129—37.
[110] Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. T-profiler: scoring the activity of predefined groups of genes using gene expression data. Nucleic Acids Res 2005;33:W592—5.
[111] Scheer M, Klawonn F, Munch R, Grote A, Hiller K, Choi C, et al. JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. Nucleic Acids Res 2006;34:W510—5.
[112] Nam D, Kim SB, Kim SK, Yang S, Kim SY, Chu IS. ADGO: analysis of differentially expressed gene sets using composite GO annotation. Bioinformatics 2006;22:2249—53.
[113] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res 2007;35:W186—92.
[114] Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, Febbo PG, et al. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. Bioinformatics 2006;22:e108—16.
[115] Liu J, Hughes-Oliver JM, Menius Jr JA. Domain-enhanced analysis of microarray data using GO annotations. Bioinformatics 2007;23:1225—34.
[116] Kim SB, Yang S, Kim SK, Kim SC, Woo HG, Volsky DJ, et al. GAzer: gene set analyzer. Bioinformatics 2007;23:1697—9.
[117] Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 2007;8:242.

Please cite this article in press as: Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.08.003