BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 11 2008, pages 1386–1393
doi:10.1093/bioinformatics/btn178

Databases and ontologies

Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists

Michael Hackenberg* and Rune Matthiesen

Bioinformatics Group, CIC bioGUNE, CIBER-HEPAD, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain

Received on December 21, 2007; revised on February 27, 2008; accepted on April 14, 2008

Advance Access publication April 23, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The ontological analysis of the gene lists obtained from DNA microarray experiments constitutes an important step in understanding the underlying biology of the analyzed system. Over the last years, many other high-throughput techniques emerged, covering now basically all 'omics' fields. However, for some of these techniques the generally used functional ontologies might not be sufficient to describe the biological system represented by the derived gene lists. For a more complete and correct interpretation of these experiments, it is important to extend substantially the number of annotations, adapting the ontological analysis to the new emerging techniques.

Results: We developed Annotation-Modules, which offers an improvement over the current tools in two critical aspects. First, the underlying annotation database implements features from many different fields like gene regulation and expression, sequence properties, evolution and conservation, genomic localization and functional categories—resulting in about 60 different annotation features. Second, it examines not only single annotations but also all the combinations, which is important to gain insight into the interplay of different mechanisms in the analyzed biological system.

Availability: http://web.bioinformatics.cicbiogune.es/AM/AnnotationModules.php

Contact: mlhackenberg@gmail.com

1 INTRODUCTION

In recent years we have witnessed a tremendous increase of high-throughput technologies which now exist in almost all fields of molecular biology. The first and still most widely used technology is the well-established DNA microarrays technique which allows monitoring the expression of thousands of genes simultaneously (Schena *et al*., 1995). However, in the last years many new technologies have been developed. One example of such novel methodologies is ChIP-on-chip design which can be used for detection of promoters, enhancers, etc. (Horak and Snyder, 2002; Wyrick and Young, 2002). Mass spectrometry is now often employed in the detection of protein abundances, protein interactions or post-translational modifications (Mann and Jensen, 2003; Selbach and Mann, 2006). Several other new technologies permit the detection of epigenetic modifications on DNA and histones (Eads *et al*., 2000; Estecio *et al*., 2007; Jenuwein and Allis, 2001; Weber *et al*., 2005) Although the technical aspects and the purposes of these techniques are quite different, the output of these experiments often consists of—or can be summarized by means of—a gene/protein list. Well-established examples are genes which are differentially expressed under different conditions (cancer, cell cycle, external stress, etc.) or co-regulated genes obtained by DNA microarray experiments. As these genes are potentially important for the analyzed biological system, the next step consists of translating the gene lists into biological knowledge from a system biology point of view.

Functional annotations like Gene Onotology (GO; Ashburner *et al*., 2000) or KEGG pathways (Ogata *et al*., 1999) have been widely used for this purpose over the last years. The GO provides a structured hierarchy of functional categories in form of a direct acyclic graph (DAG). Millions of genes and proteins are annotated as belonging to one or several of these categories (GO terms). The gene ontology constitutes a huge knowledge resource for the analysis of biological processes, cell components and molecular functions. However, sound statistical tests are needed since errors are frequently encountered in such databases.

Once we are equipped with this knowledge, a biologically meaningful question to ask is which of the functional categories are enriched or depleted among the input genes compared to a set of reference genes (Draghici *et al*., 2003). Such an analysis in general consists of three steps. First, the annotation items are assigned to the analyzed gene list and the corresponding reference set and for each item the number of associated genes is determined. Second, a statistical test is performed to calculate the *P*-value for each item. Third, and in general the last step is the correction of the *P*-values for multiple testing. The outcome is typically a list of single annotations with their corresponding (corrected) *P*-values and associated genes. Usually a reasonable biological interpretation of the cellular mechanisms of the underlying experimental/biological system can be achieved on the basis of such ranked list of functional categories.

Pioneered by Onto-Express (Khatri *et al*., 2002), many different programs, tools or web services have been developed over the last years addressing this issue, e.g. GeneMerge

*To whom correspondence should be addressed.

(Castillo-Davis and Hartl, 2003), FatiGO (Al-Shahrour *et al*., 2004), GoMiner (Zeeberg *et al*., 2003), DAVID (Dennis *et al*., 2003) or recently g:Profiler (Reimand *et al*., 2007). These were reviewed in (Khatri and Draghici, 2005), among many others. Several critical aspects of enrichment/depletion analysis have been addressed, like the introduction of multiple testing, the incorporation of different sources of functional annotations (KEGG, interPro), the performance and portability of the programs by web services, the presentation of the results and visualization capabilities (Khatri and Draghici, 2005). Moreover, recently a new algorithm called GENECODIS has been proposed which takes into account the potential relationships among single annotations by analyzing their various combinations (Carmona-Saez *et al*., 2007). In summary, it can be said that since the appearance of the first tools to carry out enrichment/depletion analysis of functional annotations, the main emphasis have been on the improvement of methodological and technical aspects. Not that much effort has been put into the incorporation of more biologically relevant annotations, apart from functional categories like GO and KEGG, a deficiency also pointed out by Al-Shahrour *et al*. (2007).

However, an expansion of the annotations beyond functional categories may be of crucial importance for the correct interpretation of gene lists obtained directly from genomics experiments (like ChIP-on-Chip) or derived from the analysis of the epigenetic state of gene promoters. Such lists might be composed of genes which interact with or are regulated by a given transcription factor (derived from ChIP-on-Chip). It might be interesting to see if these genes share certain promoter properties like the co-existence of other transcription factor binding sites (TFBS), the presence of genomic elements like CpG islands, or if a common epigenetic signature exists among these genes. Another example which emphasizes the need to incorporate more annotations can be found in the field of cancer research. It is now known that very often aberrant methylation of the promoter region or loss of microRNA regulation is involved in the formation of cancer (Greger *et al*., 1989; Gregory and Shiekhattar, 2005; Herman *et al*., 1994; Merlo *et al*., 1995; Saito *et al*., 2006). This suggests at least two possible courses of action leading to improvements. First, the incorporation of more relevant biologically annotations like regulation by microRNAs or epigenetic signatures and second—the exploration of all combinations between these annotations (this might uncover biologically interesting interplays of different mechanisms).

Given the lack of a tool with such characteristics, we developed Annotation-Modules, a web-based program which is aimed at filling this gap. First, we constructed the annotation database which expands considerably the range of annotations used so far by incorporating biological concepts from many different fields like gene regulation (TFBS, microRNA), conservation and evolution (conserved elements, taxonomic depth), population dynamics (SNPs), genomic localization and sequence properties. On top of the database a php/Java interface implements the methods and carries out the enrichment/depletion analysis. The user can customize the set of reference genes which is crucial for a correct calculation of the *P*-values (Khatri and Draghici, 2005), and it is possible to upload pre-annotated gene lists. These pre-annotated anonymous labels can then be combined with any of the features in our database. Finally, the algorithm calculates all combinations between the single annotations up to a given size (number of annotations in a combination set). This is especially important in our case where features from many different fields are annotated and analyzed simultaneously. Just these combinations can uncover the interplay between different biological mechanisms. We show the usefulness of this new tool by applying it to the well-studied CpG island genes. It is known that these genes have drastically different functional categories as compared to genes without CpG islands (Saxonov *et al*., 2006). We expanded this analysis by showing that they also have highly significant items related to the gene age, post-translational modifications, post-transcriptional regulation by microRNA and TFBS.

2 THE ALGORITHM

In general, the existing algorithms consider only single annotations and ignore interesting biology which might be inferred from the combinations of annotations. However, recently an algorithm has been published which also calculates statistically significant concurrencies of functional annotations in a gene list (Carmona-Saez *et al*., 2007). The web tool we present here is mainly based on this method to extract all combinations of annotations (Carmona-Saez *et al*., 2006). Furthermore, and apart from the extensive annotation database which underlies the algorithm, we implemented two features which we consider important. First, we allow the user to supply a pre-annotated gene list. These pre-annotated features are treated as anonymous items and can then either be analyzed on their own, or combined with any of the features which are present in our annotation database. Second, we allow the client to upload a user defined set of reference genes. This is important for the correct assessment of *P*-values if none of the standard reference sets models adequately the background probabilities; please see (Khatri and Draghici, 2005) for a more detailed discussion of this important issue.

2.1 Assigning and detecting concurrent annotations

The underlying annotation database (see Section 3) holds all features as pre-calculated and pre-assigned labels. In the first step, the algorithm reads all annotation labels chosen by the user and assigns them to the genes in the reference set and the supplied gene list. The second step finds all combinations of annotations. The number of theoretic combinations is given by:

$$N_{n,k} = \sum_{i=1}^{i=k} \binom{n}{i}$$

n being the number of different features and k the number of items per combination (the size of the combination set). This equation shows that if the number of different items or annotations is substantially increased, the algorithm will become computationally unfeasible. For example, if we

assume 1000 different items (which can be easily reached by utilizing the extensive number of annotations in our database) and a maximal combination size of $k=3$, this would lead to ~167 million different combinations. Therefore, it is mandatory to introduce some approximations in order to limit the number of combinations to an analyzable size. For example (Carmona-Saez *et al*., 2007), used a support threshold x, which reduces the number of combinations to those which have assigned at least x genes. We observed, however, that this threshold is not sufficient when a large number of annotations together with high k is analyzed. We introduce here an approach which is based on two concepts or assumptions: (1) a combination between an enriched and a depleted set of annotations is less likely to be statistically significant and (2) a maximum number of combinations which are processed on each level k.

Briefly, the modified algorithm performs the following steps:

(1) Calculates the *P*-values (see Section 2.2) for all single annotations, generates one set of depleted and one of enriched single annotations, initializes the sets of enriched and depleted combinations of annotations and stores the significant annotations.

(2) Combines in the following order as long as the number of combinations does not exceed the maximum number of combinations: (a) enriched single annotations versus enriched combinations of annotations, (b) depleted single annotations versus depleted combinations of annotations, (c) depleted single annotations versus enriched combinations of annotations and (d) enriched single annotations versus depleted combinations of annotations.

(3) Calculates the *P*-values of all resulting combinations and saves the significant ones.

(4) Generates the new sets for enriched and depleted combinations of annotations corresponding to the current level k.

(5) Repeats steps 2–4 until the threshold for k is reached.

(6) Applies the multiple testing (see Section 2.3) separately for each k.

2.2 The statistical analysis

The aim of the statistical test is to detect whether the genes in a subset are enriched or depleted for a given combination of annotations. In this sense, the fixed parameters are: the number of genes that have a given combination of annotations assigned, the number of genes that do not have them and the size of the subset (e.g. the number of genes in the gene list). The random variable which needs to be tested is the number of genes in the subset which have the given combination of annotations assigned. Rivals and coworkers (2007) showed that there is just one exact null distribution which is the hypergeometric distribution. We applied therefore an exact, two tailed hypergeometric test to calculate the *P*-values for each of the combinations applying the doubling approach (Rivals *et al*., 2007; Yates, 1984). Equation (1) shows the hypergeometric distribution.

$$P(x=i)=\frac{\binom{n_p}{i}\binom{n_n}{N-i}}{\binom{n_p+n_n}{N}} \quad (1)$$

n_p being the number of assigned genes, n_n the number of un-assigned genes, N the number of genes in the gene list and i the number of assigned genes in the gene list. Finally, the *P*-value for depletion can be calculated as two times the cumulative density function (CDF_i) at point i while the *P*-value for enrichment is given by $2*(1-CDF_i)$.

2.3 Correction for multiple testing

When several (combinations of) annotations are tested at the same time, the correction for multiple testing is of crucial importance, as reported before (Castillo-Davis and Hartl, 2003). Khatri and Draghici (2005) pointed out that the false discovery rate (FDR) is probably the best choice if several annotations are likely to be related. Given that we would expect a high number of different annotations and some of them may be related, we adjust all *P*-values by the FDR method (Benjamini *et al*., 2001).

3 THE ANNOTATION DATABASE

The annotation database currently stores information for three species, human (hg18), mouse (mm8) and rat (rn4). For each species it holds ~60 different features with nearly 18 000 different feature values. The feature values are assigned to the gene/protein tables in a pre-computed manner. As we mentioned in the introduction, apart from the widely used functional annotations of the GO ontology and Swiss-Prot keywords, the annotation database also holds features from the gene/protein sequence, evolution and conservation, as well as annotations from the gene regulation processes like TFBS or post-transcriptional regulation by microRNA. A classification of the features can be seen in Table 1.

3.1 Gene data and mapping

Some of the gene features need to be calculated or determined in a genomic context, like the presence of certain TFBS or the co-localization with CpG islands. In such cases a gene table holding information on the location of a given gene in the genome must be used internally. Right now, three different gene tables can be chosen for such analyses: RefSeq genes from NCBI, Ensembl genes from European institute of bioinformatics (EBI) and University of California Santa Cruz (UCSC)/known genes from UCSC. We downloaded all three gene tables from the UCSC table browser. If the user wishes to use annotations from a genomic context, one of these three gene lists has to be chosen. The provided IDs (ID will refer in a generic way to the labeling of a gene, transcript or protein) get automatically mapped to the IDs of the selected gene table. Currently, the annotation database allows the mapping between 12 different types of commonly used IDs (like Vega

Table 1. An overview of the annotation features implemented in our database

Annotation Field	Annotated entities
Regulation and expression	Transcription factor binding sites (TFBS), CpG islands, microRNA target sites, the expression breadth (housekeeping versus tissues specific)
Evolution and Conservation	Taxonomic depth (last common ancestor taxonomic level in the gene cluster to which the gene belongs), co-localization with phylogenetically conserved elements (PhastCons)
Functional annotations and network properties	GO-terms, Swiss-Prot keywords, post-translational modifications, disease association
Population genetics	Association with SNPs
Sequence properties (mRNA and protein)	GC-content, GC3, GC3s, mRNA length, codon usage (e.g. Nc: effective number of codons), protein properties
Miscellaneous	Distribution over chromosomes and isochores, spatial organization of gene (member of gene cluster etc.), co-localization with transposons, compositional features of the of promoter region (GC-AT classification, GC-skew, etc.)

genes, RefSeq IDs, Ensembl gene–protein-transcript IDs, IPI protein IDs, Gene Symbol, UniGene, Affymetrix IDs). The mapping between different IDs is a challenging problem (Draghici *et al*., 2006). This issue becomes especially demanding when species specific databases are involved (like WormBase, SGD, etc.). To map both the annotations between different databases and the input IDs, we used publically available mapping tables from EBI and UCSC Table Browser and cross-referenced the information to obtain all the mappings (for detailed descriptions please see http://web.bioinformatics.cicbiogune.es/AM/doc.php#Mapping).

3.2 Co-localization of genomic elements with the genes

The presence of certain genomic elements near or within genes has a clear biological meaning. Prominent examples are TFBS in the promoter region which are key factors in the regulation of gene expression or CpG islands which overlap the transcription start site (TSS) of most house-keeping genes. However, the presence of highly conserved elements—PhastCons (Siepel *et al*., 2005), SNPs or transponsable elements may uncover interesting facts and is a source of biological knowledge. We consider several different gene regions and define a genomic element as present if it overlaps with at least one base pair of the region under consideration. In this way we annotated the 'intrinsic', unambiguous regions like exons, introns, 5′UTR (untranslated regions) and 3′UTR. The definition of the promoter regions is more complicated and no consistent definition of those exists in the literature. The 'real' borders of the promoter regions may vary widely between different types of genes and will also depend on the genomic element whose co-localization with the genes is going to be established. We defined eight 'arbitrary' regions, out of which six are definitions of the promoter region and two define regions at the 3′ end of the gene (see http://web.bioinformatics.cicbiogune.es/AM/doc.php).

3.3 Gene regulation and expression

The annotation database assigns several features which are related to the regulation of gene expression and to the expression breadth of the genes (number of tissues in which the gene is expressed).

3.3.1 Detection and assignment of TFBS To detect putative binding sites of transcription factors in the promoter regions of the genes, we used the publically available position frequency matrixes (PFM) from TransFac (Matys *et al*., 2003, 2006). A well known problem is the high number of false positives which are obtained from a mere computational prediction. However, it has been reported that the incorporation of conservation may considerably improve the predictions, lowering the false positive rate (Levy and Hannenhalli, 2002). Therefore, to detect TFBS we used the multiple sequence alignments from UCSC genome browser which are built on 17 vertebrate genomes. We accept a predicted TFBS if the following conditions holds: (1) it is predicted in all species in the analysis, (2) the predicted TFBS is located at the same position in the alignment in all species and (3) the score exceeds a given threshold in all species. These conditions become more stringent the more species are included in the detection. We assembled two different prediction sets: one where we included human, mouse and rat, and a second where the conservation must exist between human, mouse, rat and dog.

It has been reported that the position of the TFBS respective to the TSS is important (Lim *et al*., 2004; Vardhanabhuti *et al*., 2007). We take this fact into account by binning the promoter region in different ways, assigning the TFBS in a function of membership to a given bin. In this way we generate four different annotation sets, dividing the promoter region (from TSS − 1500 bp to TSS + 500 bp) into 1, 2, 4 and 10 bins. Note that the more bins considered the higher is the resolution of the position. However, more bins will introduce more noise.

3.3.2 CpG islands CpG islands associate with around three quarters of all known TSS (Bajic *et al*., 2006). At least in humans, they are very important regulatory regions, involved in both the normal and disease-related regulation of gene expression (Antequera, 2003; Laird, 2005). Many different algorithms exist for the prediction of CpG islands. We incorporated the CpG islands predicted by the *CpGcluster* algorithm (Hackenberg *et al*., 2006) as they can be calculated

easily for each species applying the same thresholds and might have some advantages over other prediction algorithms by not being so sensitive to spurious transposable elements. A priori, the predicted CpG islands do not incorporate epigenetic or functional aspects, although the user can limit the analysis to those CpG islands which overlap with conserved elements; this increases the chance that these CpG islands are functional. Recently, a new method have been published which assigns a 'CpG island strength' based on the epigenetic states, histone modifications, and chromatin accessibility (Bock *et al*., 2007). We incorporated this prediction for human CpG islands, which also lets the user test against different predicted epigenetic states.

3.3.3 microRNA Over the last couple of years small non-coding RNA molecules have created a lot of interest as it became clear that the human genome is pervasively transcribed. Some members of this group, the microRNAs, are now recognized to be key players in many important biological functions, pathways and play important roles in animal evolution (Niwa and Slack, 2007). It is estimated that at least one third of all genes are subjected to post-transcriptional regulation by microRNA. Furthermore, many cases are known in which microRNAs are involved in the formation of cancer. Given that one of the sources of gene lists are the genes differentially expressed under pathologic conditions (like in cancer), we incorporated predictions of microRNA target sites which may shed new light on the underlying biology of gene lists derived from cancer assays.

In fact, we incorporated two different predictions. First, the predictions from the PicTar algorithm (Krek *et al*., 2005) which we downloaded from the UCSC table browser, both for 4-way (incorporates the conservation between four species) and 5-way (incorporates the conservation between five species). Second, we included the predictions from the miRBase (Griffiths-Jones *et al*., 2006) which are based on the miRanda algorithm (John *et al*., 2004).

3.3.4 Expression breadth We calculated the expression breadth as the percentage of tissues in which a gene is expressed. The expression values were derived from the human, mouse and rat gene atlas (Su *et al*., 2004) which we downloaded from the UCSC table browser. We averaged the expression values of different probes of one gene and considered a gene as expressed if the expression value was higher than 200 units. The expression breadth is a continuous distribution (between 0% and 100%) and therefore, in order to assign an annotation label, the distribution must be binned (see Section 3.8).

3.4 Functional annotations

Probably the most widely used set of functional annotations is the GO (Ashburner *et al*., 2000). We downloaded the gene association files and ontologies from EBI (ftp://ftp.ebi.ac.uk/pub/databases/GO/) and processed them as described before (Al-Shahrour *et al*., 2004). If a gene is annotated to a given level then we annotate it automatically to all parent levels as well. Just one level is analyzed at a time, but for all three organizing principles (molecular function, biological process and cellular component). More sophisticated methods like nested inclusive analysis (NIA) (Al-Shahrour *et al*., 2006), are not possible to implement due to the high computational burden when combining with other annotations. Note furthermore, that each gene to GO category association has assigned an evidence code like 'inferred from expression pattern' (IEP) or 'inferred from Electronic Annotation' (IEA). Right now, we include all evidence codes but remove the 'obsolete' categories.

Furthermore, we included several annotations which we extracted from the Swiss-Prot/UniProt KnowledgeBase (Bairoch *et al*., 2005). Beside the commonly used keywords, we also assigned some annotations from the feature table tag, like post-transcriptional modifications (MOD_RES) at two different evidence levels (all and just experimentally verified) or trans-membrane proteins (TRANS_MEM). Finally, we used also the comment tag from UniProt to assign the disease relatedness.

3.5 Evolution and conservation

In the current version of the database we take into consideration two types of annotations related to conservation and evolution. First, we determined for each gene a taxonomic depth which allows estimation of the age or time of the gene's appearance. We define the taxonomic depth as the last common taxonomic level of the genes which belong to the same homologous gene cluster. The gene clusters have been extracted from the HomoloGene database at NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene). As a second feature we analyzed the overlap of highly conserved genomic regions, PhastCons (Siepel *et al*., 2005) with some of the gene regions defined in Section 3.2. PhastCons are known to be associated with 3′ UTRs of regulatory genes and also show statistical evidence of enrichment for secondary RNA structure.

3.6 Sequence properties

Several sequence properties are known to be related to function. Apart from properties like the G + C contents (GC3s, GC3) we calculated the effective number of codons, Nc (Wright, 1990). This quantity, which is based on the codon homozygosities, might reveal constraints on the evolution of codon usage. The synonymous codon usage may be caused by various forms of natural selection, to optimize the efficiency and accuracy of translation or maintain structural features of the mRNA or DNA. This value can vary between 21 (very biased codon usage) and 61 (random usage).

3.7 Genomic localization

The genomic localization of a gene is believed to be related to some very interesting properties. GC-rich genomic regions are likely to be endowed with several specific features like high-transcription levels, an open chromatin structure and a very high density of genes, short introns and associated CpG islands (Bernardi, 2001). We used the IsoFinder algorithm to predict the isochores in the different genomes (Oliver *et al*., 2004) and assigned each gene to a physical isochore. Finally, we annotated the class name of its host isochore to each of

Table 2. Significant 'age' related terms of CpG island genes

Module	is	*P*-value	#Pos	#Neg	#Gene
Eutheria	depleted	4.38E − 164	1926	15 257	673
Eukaryota	enriched	1.40E − 163	3836	13 347	3134
Euarchontoglires	depleted	1.22E − 66	727	16 456	239
Hominidae	depleted	9.10E − 46	479	16 704	154
Metazoa	enriched	7.98E − 45	3140	14 043	2338
Homo	depleted	3.42E − 04	95	17 088	43
Euteleostomi	depleted	1.80E − 02	6980	10 203	4378

#Pos is the number of assigned genes and #Neg the number of genes which lack this term (both for the reference set). #Gene is the number of genes in the list which have assigned the annotation module (item set).

the genes using a classification with six isochore families (Hackenberg *et al*., 2005).

3.8 Handling continuous distributions

Some of the features we mentioned so far are not discrete or binary but continuous distributions, like the expression breadth or the coding usage Nc which may vary between 21 (highly biases codon usage) and 61 (random codon usage). In order to keep the user interface easy to manage, we pre-computed some classification/binning schemata and annotated the genes with discrete labels. We introduced three classes based on the gene frequencies. The label 'low' is assigned to a gene if it is among the *X*% of genes with the lowest values (for example the 10% of genes with lowest expression values). Consequently, we assign 'High' if the gene is among the (100-*X*) % of genes with the highest values. The rest of the genes get assigned label 'intermediate'. In the database we applied two different binning widths *X*, which can be 10 or 20%.

4 A WORKING EXAMPLE

To show the usefulness of this tool we tested it on a well-studied dataset which comprises all CpG island genes of the human genome. We define CpG island genes as those which have a CpG island overlapping its transcription start site. The methylation states of CpG islands seem to play important roles in epigenetic regulation of gene expression (Shen *et al*., 2007) and in the epigenetic formation of many cancer types, being also involved in the immortalization of the cells (Kulaeva *et al*., 2003; Neumeister *et al*., 2002).

Furthermore, it is known that they are associated with the 5′ region of almost all housekeeping genes, while they are much less common in tissue-specific genes (Antequera, 2003). Given these important functions and the enigmatic association with housekeeping genes, the CpG island genes have been widely analyzed in the past, mainly using GO ontologies. We analyzed several features which had not been taken into account before, like the age of the genes (taxonomic depth), post-translational modifications, microRNA binding sites and a combinatorial study of TFBS.

Several interesting new findings were obtained from these studies. First, (Table 2), there is a very strong difference between old and young genes. The two oldest taxonomic classes (those which contain genes which are present in all 'Eukaryota' and all 'Metazoa') are strongly enriched among CpG island genes, while a very marked difference appears at the rise of placental mammals ('Eutheria') which are strongly depleted among CpG island genes. In plants and lower eukaryotes/metazoa no CpG islands exist, so it can be assumed that at some point the oldest genes in the mammal genomes acquired the CpG islands.

Table 3. The significant terms related to post-translational modifications

Module	is	*P*-value	#Pos	#Neg	#Gene
Phosphoserine	enriched	8.60E − 77	3216	11 769	2494
No-MOD_RES	depleted	1.57E − 69	10 599	4386	6320
Phosphoserine, Phosphothreonine	enriched	6.34E − 34	1027	13 958	830
Phosphothreonine	enriched	8.04E − 34	1302	13 683	1027
Phosphotyrosine	enriched	1.65E − 13	754	14 231	575
Phosphotyrosine, Phosphoserine	enriched	9.50E − 11	384	14 601	304
Pyrrolidone carboxylic acid	depleted	3.44E − 08	58	14 927	16

The most significantly enriched terms are those related to phosphorylation.

Table 4. The significant terms related to the putative post-transcriptional regulation by microRNAs

Module	is	*P*-value	#Pos	#Neg	#Gene
No-Target	depleted	3.52E − 135	16 347	8575	9384
hsa-miR-124a	enriched	1.06E − 24	822	24 100	651
hsa-miR-19a	enriched	2.14E − 19	698	24 224	548
hsa-miR-19b	enriched	8.52E − 19	692	24 230	542
hsa-miR-93	enriched	1.93E − 17	712	24 210	552
hsa-miR-372	enriched	2.90E − 15	748	24 174	570
hsa-miR-9	enriched	8.22E − 15	808	24 114	610

Many post-transcriptional modifications are important regulatory mechanisms. Table 3 shows that the phosphorylations (phosphoserine, phosphothreonine and phosphotyrosine) are modifications which are highly over-represented among the products of CpG island genes and that the genes with no known post-translational modifications are highly under-represented.

Another post-transcriptional regulation mechanism is carried out by small non-coding RNA molecules (microRNAs) by either degradation of mRNA or inhibition of translation (Lee *et al*., 1993). It can be seen that CpG island genes seem to be heavily regulated by microRNAs (Table 4).

Given that the majority of CpG island genes are thought to be active in all cells of an organism, the promoters of these genes would be expected to contain many ubiquitous

Table 5. Significant terms related to TFBS

Module	is	*P*-value	#Pos	#Neg	#genes
V$SP1_Q6	enriched	7.88E − 117	2393	11 527	1974
V$AP2GAMMA_01	enriched	1.13E − 89	5009	8911	3686
V$STAT5A_04	depleted	4.74E − 87	7682	6238	4262
V$SP1_01	enriched	4.40E − 80	3218	10 702	2464
V$STAT4_01, V$SRY_01	depleted	8.10E − 58	3006	10 914	1505
V$STAT4_01	depleted	1.52E − 57	6383	7537	3550
V$AP2_Q6	enriched	8.62E − 56	1392	12 528	1131
V$GC_01	enriched	1.41E − 46	2102	11 818	1604
V$GATA6_01	depleted	1.10E − 41	1485	12 435	689

The transcription factors were detected in the promoter region from TSS − 1500 bp to TSS + 500 bp.

transcription factor target sites. Table 5 shows that the most enriched transcription factors are from the SP1 family, which bind to GC-rich motifs that occur frequently within CpG islands.

Probably more interesting is the high enrichment of binding sites of the AP2 gamma (activating enhancer binding protein 2 gamma) transcription factor which plays a role in the development of the eyes, face, body wall, limbs and neural tube (Werling and Schorle, 2002). It was assumed that CpG island genes are active during early embryonic development (Antequera, 2003) and this heavy enrichment of AP-2 gamma binding sites might deliver evidence in favor of this. Finally, the analysis reveals also some markedly depleted binding sites like those of some members of the STAT protein family, STAT4 and STAT5 which are transcription activators.

5 CONCLUSIONS

A new web tool for the detection of significant enrichment and depletion of combinations of annotations is presented. The tool accepts 12 different input IDs, allows free selection of the reference genes and the upload of pre-annotated gene lists. Currently, it holds ∼60 different annotation features from functional annotations, regulation of gene expression, conservation/evolution and sequence properties, which extends by far the number of available annotations compared to current tools. Furthermore, it not only analyses single annotations but also combinations of different annotations. This combinatorial analysis may be important to discover the interplay between different biological mechanisms in the analyzed biological system.

ACKNOWLEDGEMENTS

The authors would like to thank Ewa Gubb for her help in the preparation of the manuscript and Gorka Lasso for his help in the layout of the tool.

Funding: Support for M.H. and R.M. was provided from The Department of Industry, Tourism and Trade of the Government of the Autonomous Community of the Basque Country (Etortek Research Programs 2005/2006) and from the Innovation Technology Department of the Bizkaia County.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al.* (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Al-Shahrour,F. *et al.* (2007) FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.

Al-Shahrour,F. *et al.* (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**, W472–W476.

Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci.*, **60**, 1647–1658.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Bajic,V.B. *et al.* (2006) Mice and men: their promoter properties. *PLoS Genet.*, **2**, e54.

Benjamini,Y. *et al.* (2001) Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.*, **125**, 279–284.

Bernardi,G. (2001) Misunderstandings about isochores. Part 1. *Gene*, **276**, 3–13.

Bock,C. *et al.* (2007) CpG island mapping by epigenome prediction. *PLoS Comput. Biol.*, **3**, e110.

Carmona-Saez,P. *et al.* (2006) Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, **7**, 54.

Carmona-Saez,P. *et al.* (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.

Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.

Dennis,G.Jr. *et al.* (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Draghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Draghici,S. *et al.* (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.

Eads,C.A. *et al.* (2000) MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res.*, **28**, E32.

Estecio,M.R. *et al.* (2007) High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome Res.*, **17**, 1529–1536.

Greger,V. *et al.* (1989) Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum. Genet.*, **83**, 155–158.

Gregory,R.I. and Shiekhattar,R. (2005) MicroRNA biogenesis and cancer. *Cancer Res.*, **65**, 3509–3512.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Hackenberg,M. *et al.* (2005) The biased distribution of Alus in human isochores might be driven by recombination. *J. Mol. Evol.*, **60**, 365–377.

Hackenberg,M. *et al.* (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.

Herman,J.G. *et al.* (1994) Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl Acad. Sci. USA*, **91**, 9700–9704.

Horak,C.E. and Snyder,M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.*, **350**, 469–483.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

John,B. *et al.* (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Khatri,P. *et al.* (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.

Krek,A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Kulaeva,O.I. *et al.* (2003) Epigenetic silencing of multiple interferon pathway genes after cellular immortalization. *Oncogene*, **22**, 4118–4127.

Laird,P.W. (2005) Cancer epigenetics. *Human Mol. Genet.*, **14**, R65–R76.
Lee,R.C. *et al.* (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.
Lim,C.Y. *et al.* (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.*, **18**, 1606–1617.
Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.
Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
Merlo,A. *et al.* (1995) 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.*, **1**, 686–692.
Neumeister,P. *et al.* (2002) Senescence and epigenetic dysregulation in cancer. *Int. J. Biochem. Cell Biol.*, **34**, 1475–1490.
Niwa,R. and Slack,F.J. (2007) The evolution of animal microRNA function. *Curr. Opin. Genet. Dev.*, **17**, 145–150.
Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
Oliver,J.L. *et al.* (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, **32**, W287–W292.
Reimand,J. *et al.* (2007) g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
Saito,Y. *et al.* (2006) Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells. *Cancer cell*, **9**, 435–443.
Saxonov,S. *et al.* (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
Selbach,M. and Mann,M. (2006) Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nat. Methods*, **3**, 981–983.
Shen,L. *et al.* (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.*, **3**, 2023–2036.
Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Vardhanabhuti,S. *et al.* (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.
Weber,M. *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.
Werling,U. and Schorle,H. (2002) Transcription factor gene AP-2 gamma essential for early murine development. *Mol. Cell. Biol.*, **22**, 3149–3156.
Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.*, **12**, 130–136.
Yates,F. (1984) Test of significance for 2x2 contingency tables. *J. Royal Stat. Soc. Ser. A.*, **147**, 426–463.
Zeeberg,B.R. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.