Prosecutor: parameter-free inference of gene function for prokaryotes using DNA microarray data, genomic context and multiple gene annotation sources

Evert Jan Blom1, Rainer Breitling2, Klaas Jan Hofstede1, Jos BTM Roerdink3, Sacha AFT van Hijum1,4 and Oscar P Kuipers*1

Address: 1Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, the Netherlands, 2Groningen Bioinformatics Centre, University of Groningen Kerklaan 30, 9751 NN, Haren, the Netherlands, 3Institute for Mathematics and Computing Science, University of Groningen, Nijenborgh 9, 9747 AG, Groningen, the Netherlands and *Current address: NIZO Food Research, Kernhemseweg 2, 6718 ZB, Ede the Netherlands

Email: Evert Jan Blom - e.j.blom@rug.nl; Rainer Breitling - R.Breitling@rug.nl; Klaas Jan Hofstede - klaasjanhofstede@home.nl; Jos BTM Roerdink - j.b.t.m.roerdink@rug.nl; Sacha AFT van Hijum - sacha.vanhijum@nizo.nl; Oscar P Kuipers* - o.p.kuipers@rug.nl

* Corresponding author

Abstract

Background: Despite a plethora of functional genomic efforts, the function of many genes in sequenced genomes remains unknown. The increasing amount of microarray data for many species allows employing the guilt-by-association principle to predict function on a large scale: genes exhibiting similar expression patterns are more likely to participate in shared biological processes.

Results: We developed Prosecutor, an application that enables researchers to rapidly infer gene function based on available gene expression data and functional annotations. Our parameter-free functional prediction method uses a sensitive algorithm to achieve a high association rate of linking genes with unknown function to annotated genes. Furthermore, Prosecutor utilizes additional biological information such as genomic context and known regulatory mechanisms that are specific for prokaryotes. We analyzed publicly available transcriptome data sets and used literature sources to validate putative functions suggested by Prosecutor. We supply the complete results of our analysis for 11 prokaryotic organisms on a dedicated website.

Conclusion: The Prosecutor software and supplementary datasets available at http://www.prosecutor.nl allow researchers working on any of the analyzed organisms to quickly identify the putative functions of their genes of interest. A de novo analysis allows new organisms to be studied.

Background

One of the central challenges in computational biology is the prediction of gene function [1]. The inference of gene function typically starts with DNA sequence analysis based on ortholog information [2-5]. Although this method has proven to be successful in many cases, considerable numbers of genes (20–50%) in current genome annotations still are of unknown function. Complementary approaches are therefore required to characterize the function of these genes.
Since the start of the DNA microarray era, the "guilt-by-association" (GBA) methodology has been used to infer gene function [6-9]. This concept is based on the assumption that genes involved in similar cellular functions are likely to display correlated expression behavior [10-12]. In addition, this correlated behavior might identify common regulatory mechanisms.

Ultimately, to understand the function of a new gene, one should exploit all available experimental data sources (e.g., transcriptomics, proteomics, protein-protein interactions and metabolomics) [13,14] or even by the joint efforts of many scientists in a community annotation [15]. Previous work on gene function prediction has mainly been focused on higher organisms using multiple high-throughput data sources [16-18]. On the other hand, genome organizational principles that are unique for prokaryotes supply valuable additional information about gene function.

However, it is expected that the GBA method is particularly powerful for prokaryotes, due to their tight coupling of transcription and translation [19]. In addition, for many prokaryotes, the available gene expression datasets greatly outnumber other experimental data sources.

To improve the analysis of the predictions, Prosecutor provides additional information for each annotated gene, most notably in its genomic context, which is particularly useful for operons. The occurrence of adjacent divergent co-expressed genes is also highlighted since these are expected to be co-regulated [20]. Finally, putative new members of transcriptional modules are examined for the presence of the same regulatory motif that is already known for the module.

Our Prosecutor software imposes no constraints on the biological annotations used; it generates hypotheses based on large variety of annotation sources e.g., Gene Ontology, metabolic pathways, UniProt keywords, etc. This is in contrast to most other methods [11,12,16-18,21-24] which, with few exceptions [8,10], are focused on coupling genes to Gene Ontology sources only.

We discuss some of the functional assignments obtained by Prosecutor, as well as a number of mining capabilities provided by the software. We find that the increasing variety of experimental conditions used in DNA microarray experiments has greatly improved the ability to identify the function of unknown genes using GBA principles.

**Results and discussion**

**Prosecutor software**

Prosecutor is a standalone application developed in Java and shares its functional database structure with the FIVA software [25]. It features an iterative implementation of the GBA method which is based on iterative Group Analysis algorithm (iGA) [26]. Several characteristics of the software analysis modules are described below.

**The Iterative Guilt-By-Association (iGBA) method**

The iGBA method requires DNA microarray datasets and functional categories from annotation sources to infer putative gene functions. The rationale for our approach is the GBA principle, i.e., genes that are functionally involved in, or linked to, the same function will in general show higher expression correlations than genes that are not functionally related. The prediction algorithm of Prosecutor calculates the significance of association for all pairs of genes and functional categories. For n genes, expression profiles from DNA microarrays (Fig. 1A) are used to create an n × n correlation matrix M (Fig. 1B). Each row j of this matrix represents the (Pearson or Spearman) expression correlation between gene gj and all other genes. To annotate each gene gj, we sort all other genes by their correlation with gene gj and subject the resulting sorted gene list to iGA (Fig. 1C). This results in a list of functional categories that are over-represented among the genes that are highly correlated with gene gj, with associated p-values. The iGA algorithm works iteratively and therefore does not require a fixed cutoff of the sorted correlation list, no minimum correlation has to be defined. Instead, iGA determines the appropriate cutoff that yields the lowest p-value for each individual analysis of a gene to a functional category. As a consequence, the function assignment by iGA is very sensitive [26] compared to methods which use a predefined correlation cut-off.

**Performance of functional categories**

**Receiver Operating Characteristic curves**

The performance on well-annotated genes was assessed to evaluate the sensitivity of the iGBA method. This evaluation has to be specific for each functional category, because for some of them we expect that all members show close correlation, while others are so general that their members will not correlate and iGBA is expected to fail. The category specific evaluation of expression coherence is done as follows: Our iGBA algorithm yields a p-value for every pair of gene-functional category pair (Fig. 1C). This p-value is indicative of the confidence of the assignment of a gene to a functional category. For each category we sort the gene list by p-values and examine the positions of the p-values of its known members in this sorted list. We are then able to calculate an "expression coherence value" for each functional category by plotting the true and false positive rates on Receiver Operating Characteristic (ROC) curves (Fig. 1D) [27]. The corresponding Area Under the ROC Curve (AUC) is a quantitative measure of the expression coherence of the genes of a functional category. A functional category in which all
Flowchart of Prosecutor

Figure 1. Flowchart of the functional prediction process in Prosecutor. First, the expression profiles from DNA microarrays (1A) are used to create a correlation matrix (1B). For every gene, the correlations with the remaining genes are retrieved from the correlation matrix and sorted (1B2). The sorted gene list is used to perform an iterative Group Analysis for every functional category (1B3). The resulting p-value is indicative for the prediction of a gene as a member of a functional category (1C). At this step, the regular iGBA process ends. However, to also assess the reliability of each prediction, the following steps are added. The complete list of p-values for every functional category is sorted (1C4), after which the positions of the members of the functional category are determined (1C5). These positions are used to create ROC curves (1D; see Results section for more information concerning ROC curves). The corresponding Area Under the ROC Curve (AUC) is then used as a measure of expression coherence value of a functional category.
known members show strong co-expression will have an AUC close to 1.0, whereas a randomly predicting functional category (i.e., a category that does not show co-expression of its members) would yield AUC values around 0.5. Using the AUC measure, we are now able to select the most promising functional categories for further analysis.

**Parameter free approach**

Various methods have been developed that specifically employ data from microarrays studies [21-24]. Some of these methods are designed for temporal gene expression profiles [23,24] or calculate a functional enrichment for each dataset [22]. Other approaches require preprocessing of the annotation data, e.g., generating a set of validated and highly unlikely associations (see [28] for more information) used for training of the prediction model [21]. Our Prosecutor application improves on previous methods by providing a parameter free approach for the inferential gene function. No trusted set of functional associations between proteins is required since Prosecutor treats every functional category individually, thereby circumventing preselection toward particular processes.

**Additional layers of information**

The strength of Prosecutor comes also from its additional prokaryote-specific layers of information combined with a convenient visualization of the functional predictions. This prioritizing of the results allows for the rapid identification of the most promising function predictions.

**Genomic context analysis**

The function predictions generated by Prosecutor are provided for individual genes. Genes co-transcribed to a polycistronic messenger RNA are known as operons whose members typically share biological function. Predictions for genes of which other member(s) of the same operon were already linked to the predicted function are highlighted in the visualization of the results. The same procedure is applied to divergent genes which share the same upstream region (Fig. 2B). This layer of information that is based on the genomic context of genes provides additional, and in some case cases vital, information concerning putative function predictions.

**Regulatory mechanism analysis**

Transcriptional modules represent genes that are regulated by a common regulator. The regulatory mechanisms underlying the co-expression of members of a transcriptional module are used as additional evidence to prioritize the Prosecutor results. For some organisms, functional annotations based on curated knowledge of transcriptional modules are available [29,30]. Motif instances from all members of a transcriptional module are used to create a position specific scoring matrix. This matrix is used to search for additional hits in the upstream and coding regions from the first gene of the operon as well as the gene of interest (in case of residing in an operon). Using this approach, we are able to predict putative new targets for transcriptional modules that exhibit significant co-expression with known members of the transcriptional module and a putative regulatory motif in their upstream regions (Fig. 2C).

**Graph visualization**

Functional predictions are represented by Prosecutor as graphs using the Prefuse toolkit [31] to visualize the gene redundancy and overlap between the functional categories of different functional predictions. This method allows to visually determine the uniqueness of each of the function predictions. A force-directed layout from the Prefuse visualization framework is used to position the different nodes (genes) in the network (Fig. 2D).

**Performance compared to random microarray data**

The performance of different annotation sources (e.g., Gene Ontology terms) was investigated by comparing AUC results for real and random data using a two-sample Kolmogorov-Smirnov test. This method was used to compare the distribution of AUC values of our algorithm based on 305 microarrays from *E. coli* (Fig. 3A) as compared to results for which the genes were randomized (the link between expression and annotation is expected to be lost) (Fig. 3B). The null hypothesis that the true data do not significantly deviate from the random distribution is rejected with a *p*-value of 2e-16. The real data yield significantly higher AUC values than expected by chance. This confirms that the coexpression enrichment of many functional categories is predictive of gene function. Additional analysis of the AUC distribution across the annotation sources shows that the transcription module annotation source contains a large number of high scoring functional categories (i.e., categories exceeding an AUC value of 0.9). Moreover, we found that applying a Pearson correlation measure for calculating the correlation matrix outperforms Spearman correlations, generating 16% more functional categories with an AUC value of 0.8 or higher (data not shown).

**Prosecutor test-cases**

Most genome annotations deposited to GenBank are rarely if ever updated [32]. As research progresses, knowledge of many previously uncharacterized genes improves. This annotation gap enables us to analyze results obtained by Prosecutor by manual literature mining of genes for which no function was available in the original genome annotation. For this validation, only functional categories exhibiting strong predictive properties, with AUC values higher than 0.7, were taken into account.
Figure 2  (see legend on next page)
context analysis is also used to detect known regulatory sequences for transcriptional modules (2C). Lastly, graph visualization

genes that share the same upstream region (in this example transcription modules.

large number of putative and validated members for tran-

and the additional analysis of the results using known reg-

tion are marked "putative" in Table 2. Due to the

members which could not be verified using literature informa-

remainder of the putative transcriptional module mem-

literature, but are not yet catalogued in RegulonDB. The

found that some of the newly identified putative tran-

were supplemented with data obtained from the position

mation from RegulonDB [30]. The results of Prosecutor

affiliations, 2) a sorted list of functional categories

tional website [35]. On this dedicated web-site func-

The first analysis deals with results obtained from Prose-

cut for all tested organisms and was based on data from
dual-dye microarrays. Prosecutor predicted a large number of gene functions for previously unannotated genes which could be validated using literature informa-

The complete results of this analysis is available on the supplemental website. Analysis of the results for the model organisms E. coli and B. subtilis was facilitated by the large diversity of microarray perturba-
tion studies available. A detailed analysis for B. subtilis revealed that for 25% of the best 160 predictions suffi-
cient literature data was available to positively confirm the predictions (data not shown).

The second analysis dealt with the detection of putative new members of existing transcriptional modules in E. coli (Table 2). We used gene expression data from 305 Affymetrix genechips [33] combined with functional annotations based on curated regulatory network inform-

ation from RegulonDB [30]. The results of Prosecutor were supplemented with data obtained from the position

specific scoring matrices. These matrices were based on

aligned motif sequences of the known DNA binding sites from the members of every transcriptional module. We found that some of the newly identified putative trans-

criptional module members had been confirmed in the literature, but are not yet catalogued in RegulonDB. The

remainder of the putative transcriptional module mem-

bers which could not be verified using literature informa-

tion are marked "putative" in Table 2. Due to the

exceptional predictive performance (almost 60% of the transcriptional modules shows an AUC value above 0.9) and the additional analysis of the results using known reg-

ulatory mechanisms, we were able to reliably predict a large number of putative and validated members for tran-

scription modules.

Various layers of information are supplied for the iGBA results (2A) from Prosecutor. Predicted functional assignments for genes whose operon members are already linked to the predicted function are indicated in the results (2B). In addition, this protocol is also followed for divergent genes that share the same upstream region (in this example pps and ydiA). The operon information that is used for the genomic context analysis is also used to detect known regulatory sequences for transcriptional modules (2C). Lastly, graph visualization is used to visualize the gene redundancy of the different functional assignments of Prosecutor (2D). Nodes in the graph repres-

ent functional categories and genes. Arrows represent membership of gene nodes to a functional category node as well as the putative functional prediction of the studied gene. The members of individual categories are placed in colored aggregates. In addition to the aggregates, a colored square is placed in each gene member of a category. The squares are colored using the
colors of their matching aggregates. Members of different categories can easily be distinguished using the colored squares. An example of a functional prediction found by Prosecutor for ydiE from E. coli is shown. The expression of this gene was corre-
lated with members of various functional categories involved in the uptake of iron. In addition to the functional association with the transcriptional module Fur, the upstream region of ydiE also contains a putative Fur DNA binding site.

The genome annotation of S. cerevisiae is available in Gen-

bank as well as EMBL format, allowing our Prosecutor software to perform an iGBA analysis. For this third anal-

ysis we used two annotation sources (metabolic pathways and Gene Ontology). The gene expression data was obtained from the Stanford microarray database [34]. The distribution of AUC values of our algorithm (Fig. 4A) is compared to results for which the genes were randomized (Fig. 4B). The results based on the real data yield more large AUC values than expected by chance. The categories with high AUC values will presumably allow our iGBA method to assign reliable functional predictions. This demonstrates that Prosecutor, while being specifically optimized for prokaryotes, will also be a useful tool for the general biologist community.

The complete results of the annotation efforts from our software for twelve organisms are available on the supplement-

al website [35]. On this dedicated web-site functional couplings can be mined in three ways: 1) through a

list of the best functional couplings for each functional category; this allows "browsing" through the most prom-

ising associations, 2) a sorted list of functional categories and their predictive power (AUC); in case that one is inter-

ested in the genes that are associated with a specific func-

tional category, and 3) a sorted list of genes; allows to identify to which functional categories a gene of interest is associated. All data sources used for analysis are available, allowing researchers studying any of the analyzed organ-

isms to perform a functional analysis for their expression dataset and/or functional categories.

Prosecutor uses DNA microarray data combined with functional annotations to infer putative gene functions. We show that multiple annotation sources are informa-

Figure 2 (see previous page)
Schematic overview of the additional information provided by Prosecutor. Various layers of information are sup-
plied for the iGBA results (2A) from Prosecutor. Predicted functional assignments for genes whose operon members are already linked to the predicted function are indicated in the results (2B). In addition, this protocol is also followed for divergent genes that share the same upstream region (in this example pps and ydiA). The operon information that is used for the genomic context analysis is also used to detect known regulatory sequences for transcriptional modules (2C). Lastly, graph visualization is used to visualize the gene redundancy of the different functional assignments of Prosecutor (2D). Nodes in the graph represent functional categories and genes. Arrows represent membership of gene nodes to a functional category node as well as the putative functional prediction of the studied gene. The members of individual categories are placed in colored aggregates. In addition to the aggregates, a colored square is placed in each gene member of a category. The squares are colored using the colors of their matching aggregates. Members of different categories can easily be distinguished using the colored squares. An example of a functional prediction found by Prosecutor for ydiE from E. coli is shown. The expression of this gene was correlated with members of various functional categories involved in the uptake of iron. In addition to the functional association with the transcriptional module Fur, the upstream region of ydiE also contains a putative Fur DNA binding site.
Figure 3 (see legend on next page)
tive and non-redundant and allow maximizing the use of all available DNA microarray data. For B. subtilis, we were able to confirm 40 out of the 160 best functional predictions generated by Prosecutor, using published literature. We therefore believe that the other functional assignments based on our analysis are also likely to be informative and reliable. Combined with regulatory motif information for the species B. subtilis and E. coli, Prosecutor allows the identification of new transcriptional module members. Prosecutor can thus serve as a generic tool for a genome-wide (re)annotation of gene functions in prokaryotes. The results of such a re-annotation effort, for 11 widely studied bacterial species, is supplied as a community resource at the associated website [35].

**Figure 3 (see previous page)**

**Prediction ability of four annotation sources.** Histograms of ROC areas (Area Under the Curve) for four annotation sources for E. coli based on 305 microarrays (3A) compared to randomized results (3B). The real data reveal a large amount of categories with AUC values larger than 0.8, which are almost absent in randomized results. These categories are the most promising candidates for which the iGBA approach will enable confident gene assignments functional predictions. Analysis of the AUC distribution across the annotation sources shows that the "transcription module" annotation source is the most informative, i.e., contains the largest amount of categories exceeding an AUC value of 0.9 (3A). This is intuitively very convincing as shared transcriptional regulation is the basis of coexpression. In addition to ROC areas for all GO terms, we have also analyzed the distribution of ROC areas for the GO annotation source using the "gold standard" [28]. This proposed "gold standard" (GS) consists of a specific trusted set of biological processes that maps proteins to well-defined functional classes to evaluate predictions. The authors supply a set of biological processes that is based on selection by a panel of biology experts. We have included AUC results for the GO annotation for E. coli using the GS. Analysis of the AUC distributions shows that the distribution of relative occurrences of the GS analysis and the analysis using a fixed member cutoff is comparable.

| Organism             | Gene      | Functional category                        | Rank | AUC  | Reference  |
|----------------------|-----------|--------------------------------------------|------|------|------------|
| Campylobacter jejuni | GJ0391c   | Pathway flagellar assembly                 | 2    | 0.76 | [41,42]    |
|                      | GJ1242    | GO:0003774 motor activity                   | 3    | 0.75 | [42]       |
|                      | GJ1316c   | GO:0019861 agellum                          | 7    | 0.72 |            |
| Escherichia coli     | yncE      | GO:0015343 siderophore-iron transmembrane transporter activity | 7    | 0.96 | [44]       |
|                      | ybiX      | UP:Enterobactin biosynthesis               | 1    | 0.99 | [44]       |
|                      | cho       | GO:0009432 SOS response                     | 17   | 0.92 | [45]       |
|                      | ybeD      | GO:0051082 unfolded protein binding         | 1    | 0.78 | [46]       |
|                      | ulaC      | GO:0019852 L-ascorbic acid metabolic process | 1    | 0.76 | [47]       |
|                      | yciW      | GO:0006534 cysteine metabolic process       | 10   | 0.82 | [48]       |
| Bacillus subtilis    | yphG      | transcriptional module SigM                 | 18   | 0.81 | [49]       |
|                      | ykuO      | transcriptional module Fur                  | 8    | 0.75 | [50]       |
|                      | ywif      | GO:0006935 chemotaxis                       | 37   | 0.90 | [51]       |
|                      | yxlf      | Pathway Flagellar assembly                  | 7    | 0.89 | [52]       |
|                      | yfHE      | transcriptional module GerE                | 52   | 0.81 | [53]       |
| Streptomyces coelicolor | SCBAC28G1.05 | PW:Biosynthesis of type II polyketide back- bone | 1    | 0.99 | [54]       |
|                      | SCBAC28G1.07 | PW:Biosynthesis of type II polyketide products | 8    | 0.76 | [54]       |
| Vibrio cholera       | VC1688    | GO:0006826 iron ion transport               | 2    | 0.87 | [55]       |
|                      | VCA0216   | GO:0019290 siderophore biosynthetic process  | 8    | 0.98 | [55]       |
|                      | VC1267    | GO:0019290 siderophore biosynthetic process  | 6    | 0.98 | [55]       |

Functional predictions identified by Prosecutor for several organisms that are confirmed using literature information. Gene: the gene for which a validated functional prediction with a functional category (column three) was found. Rank: the position of a gene in the prioritized list based on p-values. These p-values describe the functional prediction significance for every individual gene with a specific functional category. AUC: the expression coherence value for a functional category with respect to its own members. Notice the examples of genes for which a particular annotation is assigned rank 1. This means that this gene is more close associated with this functional category then any of the original known members of the category, indicating a very high confidence in the prediction.
Methods

Implementation & Availability

Prosecutor was programmed as a multithreaded standalone application in Java using the Eclipse framework http://www.eclipse.org/ as a Rich Client Platform (source code is available upon request). Prosecutor runs on all Java-supporting operating systems (MS Windows, Linux and Mac OS). The Prosecutor was developed from a bacterial perspective and therefore supports the two major prokaryotic genome annotation formats (Genbank and

Table 2: Extending transcriptional modules of E. coli

| transcriptional module       | gene | prosecutor rank | motif rank | motif sequence in the intergenic region of either the gene or its operon | literature reference |
|------------------------------|------|----------------|------------|------------------------------------------------------------------------|---------------------|
| **ArgR** Amino acid biosynthesis: Arginine **AUC 0.92** | artJ | 21             | 8          | TGCATAACATTTCG                                                       | [56]                |
|                              | aroP | 58             | 39         | TGGTTTTTTATCTCA                                                      | [57]                |
|                              | artL | 131            | 50         | TGCATAATTTCCT                                                       | [56]                |
|                              | hisL | 16             | 4          | TGAATAAACATTC                                                       | putative            |
|                              | pyrL | 32             | 61         | TGAACCTTTTTATCT                                                      | putative            |
|                              | metH | 36             | 76         | TGAATTTTTTTATATA                                                    | putative            |
|                              | ydcS | 43             | 63         | TGAATAAATTTTCCT                                                    | putative            |
|                              | sfpA | 132            | 21         | TGCATTTTTTTATTC                                                    | putative            |
|                              | hisG | 141            | 8          | TGAATAAACATTC                                                       | putative            |
|                              | hisJ | 144            | 27         | TGAACCTTTTTATTC                                                      | putative            |
|                              | hisC | 145            | 13         | TGAATAAACATTC                                                       | putative            |
|                              | hisA | 147            | 14         | TGAATAAACATTC                                                       | putative            |
|                              | potF | 162            | 46         | TGCATAAAAATTC                                                       | putative            |

| **CysB** Amino acid biosynthesis: Cysteine **AUC 0.91** | sbp  | 12             | 0          | CGCAAGTTATAGCAATCTTTTTATTTTTTAT                                    | [48,58]             |
|                                                          | nlPA | 36             | 17         | CAGACCTTTATTTCCACTTTTTATTCTT                                       | [48]                |
|                                                          | mmuP | 28             | 40         | AAGCGGTTAAACAAAACTTTCTTGG                                          | putative            |

| **Fur** iron regulatory gene **AUC 0.84** | yncD | 74             | 32         | GGAAGTAGTTATACATTTTAT                                               | [44]                |
|                                            | yboN | 37             | 5          | GAAATGTAAATTTGTTGATT                                               | putative            |
|                                            | floE | 101            | 29         | GGCAATTAAATACATTT                                              | putative            |

| **LexA** major regulator of DNA repair **AUC 0.87** | yebG | 0              | 21         | CTGTATAAAAATCACAG                                                   | [59,60]             |
|                                                     | dinI | 2              | 6          | CTGTATAATTAACACC                                                       | [61,62]             |
|                                                     | dinB | 6              | 51         | CTGTATACTTTACCA                                                       | [63]                |
|                                                     | dinD | 19             | 0          | CTGTATAATTTACCA                                                       | [45]                |
|                                                     | yjW  | 39             | 20         | CTGTATATAATACCA                                                       | [45]                |
|                                                     | ybE  | 120            | 31         | CTGTATAAAAACCCCA                                                        | [45]                |
|                                                     | sbmC | 125            | 4          | CTGTATATAAAACAG                                                        | [64]                |

| **MetJ** Amino acid biosynthesis: Methionine **AUC 0.88** | ybdH | 10             | 6          | AGACGTTTAGATGTCT                                                   | [65]                |
|                                                        | ybdL | 106            | 0          | AGACATTAAACGGTCT                                                   | [65]                |
|                                                        | ycbK | 198            | 17         | AGTACCTTTGAGGTCT                                                   | [65]                |
|                                                        | mmuP | 14             | 15         | GGATGTGGTATGC                                                       | putative            |

Transcriptional module member predictions identified by Prosecutor applied to well-known transcriptional modules for E. coli. Gene: the gene for which a validated functional prediction with a transcriptional module (column three) was found. Rank: represents the position of a gene in the sorted list with p-values. These p-values describe the functional prediction significance for every individual gene with a specific functional category. More significant p-values are matched with lower ranks. AUC: describes the association efficiency for a transcriptional module with respect to its own members. In addition to the rank information provided by Prosecutor, supplemental motif information is provided. This data is obtained by applying a position specific scoring matrix (PSSM) to the upstream sequences of all genes in the genome. The PSSM is derived from aggregating all known consensus target sequences (DNA regulatory binding sites). The additional motif information allows users to concentrate on genes that exhibit coexpression with a transcriptional module as well as possessing a predicted consensus sequence. This additional evidence contributes to the confidence in assigning a gene to a particular transcriptional module. Motif rank: based on the results for a PSSM when matched to every upstream sequence in the genome. For example, based on the PSSM of the regulator LexA, the upstream region of gene dinD contains the best ranking motif (rank 1).
Figure 4
Prediction ability of two annotation sources for yeast. Histograms of ROC areas (Area Under the Curve) for two annotation sources (Gene Ontology and metabolic pathways) for S. cerevisiae based on 1079 datasets from Stanford microarray database (4A) compared to randomized results (4B). The real data reveal a large number of categories with AUC values larger than 0.8, which are almost absent in randomized results. These categories are the most promising candidates for which the iGBA approach will enable confident gene assignments of functional predictions.
EMBL). A simplified tabulated genome annotation format can also be used, enabling organisms for which no Genbank or EMBL file is available to be studied.

Data sources
The basic requirements of an analysis consist of a genome annotation (i.e., Genbank or EMBL) and a collection of microarray data. Currently, six different annotation systems are implemented: (i) transcriptional modules, (ii) gene ontologies (GO) [36], (iii) metabolic pathways from the KEGG database [37] (iv) UniProt keywords [38], (v) InterPro domains [39] and (vi) user-defined categories.

DNA microarray datasets
DNA microarray data used in this study consisted of dual dye arrays for 11 prokaryotic organisms and yeast from the KEGG expression database [40] and the Stanford microarray database [34]. For *E. coli*, an additional 305 Affymetrix expression arrays were obtained from the M3D Database [33].

Multiple testing correction
A typical problem in genome-wide statistical analysis is the occurrence of many false positives (i.e., a functional prediction that is mistakenly found significant due to multiple testing). The incidence of false positives is roughly proportional to the number of tests performed. Since a typical search in Prosecutor may consist of thousands of tests, the chance of obtaining false positive predictions is large. We have used a strict Bonferroni multiple testing correction method to correct the raw p-values from the iGA results to minimize this problem.

Authors' contributions
EJB conceived the study and programmed the software. EJB and RB devised the iGBA algorithm. KIH designed and programmed the analysis interface. EJB and SAFTH wrote the manuscript. JBTMR and OPK guided and coordinated the project. All authors read, corrected and approved the final manuscript.

Acknowledgements
This study was supported by a grant from the Netherlands Organisation for Scientific Research and industrial partners in the NWO-BMI project number 050.50.206 on Computational Genomics of Prokaryotes and by the Center IOP Genomics. Work of SVH was in part supported by the BMBF within the framework of the transnational SysMO initiative in the project BaCell-SysMO. This work was in part supported by an EU program in FW6: Bacell Health, European Union Grant LSHG-CT-2004-503468. We thank Grayson H. Kleine C.B. de Miranda for creating a multi-threaded implementation of the iGA algorithm.

References
1. Friedberg I: Automated protein function prediction—the genomic challenge. Brief Bioinform. 2006, 7:225-242.
2. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Díaz N, Dísz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goessmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweiger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Raskóvá DA, Rickert C, Rostocker J, Stevens R, Thiele I, Vassiava O, Ye Y, Zagnitko W, O, Vonstein V: The sysbio systems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005, 33:5691-5702.
3. Hug, M Snel B, Lathe W, Bork P: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000, 10:1204-1210.
4. Wu J, Hu Z, DeLisi C: Gene annotation and network inference by phylogenetic profiling. BMC Bioinformatics 2006, 7:80.
5. Wu H, Su Z, Mao F, Olman V, Xu Y: Prediction of functional modules based on comparative genome analysis and Gene Ontology application. Nucleic Acids Res 2005, 33:2822-2837.
6. Brent R: Functional genomics: learning to think about gene expression data. Curr Biol 1999, 9(R338-R341).
7. Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, Marti-Renom M, Oliva B: Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics 2008, 9:249.
8. Mons B, Ashburner M, Chichester C, van Mulligen E, Weebber M, den Dunnen JT, Musen M, Cockrell M, Hermjakob H, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Borner K, Meijssen G, Moeller E, Roos PJ, Mons A, van Ommen GJ, Wales J, Bairoch A: Calling on a million minds for community annotation in WikiProteins. Genome Biol 2008, 9(R8).
9. Pavlidis P, Weston J, Cai J, Noble WS: Learning gene functional classifications from multiple data types. J Comput Biol 2002, 9:401-411.
10. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA 2003, 100:8348-8353.
11. Lee I, Li Z, Marcotte EM: An improved, bias-reduced probabilistic functional gene network of baker’s yeast, *Saccharomyces cerevisiae*. PLoS ONE 2007, 2:e988.
12. Miller OL, Hankisho BM, Lomas CA: Visualization of bacterial genes in action. Science 1970, 169:392-395.
13. Korbel JO, Jensen LJ, von Mering C, Bork P: Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 2004, 22:919-927.
14. Hubs, MA, Hess DC, Myers CL, Hutenhoven C, Li K, Troyanskaya OG: Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 2007, 23(20):2692-2699.
23. Lagreid A, Hvidsten TR, Midelhart H, Komorowski J, Sandvik AK: Predicting gene ontology biological process from temporal gene expression patterns. Genome Res 2004, 14:965-979.

24. Zhang X, Zou S, Song J: Temporal-domain specific models for prediction of gene ontology biological processes using transcriptional profiles of aging in Drosophila melanogaster. BMC Bioinformatics 2008, 9:129.

25. Blouin EJ, Bosman OWJ, van Huijum SA, Breitling R, Tijmjen L, Silvis R, Roerdink JB, Kuipers OP: FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. Bioinformatics 2007, 23:1161-1163.

26. Breitling R, Amtmann A, Herzky P: Iterative Group Analysis (IGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics 2004, 5:34.

27. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L: The use of role modeling to compute characteristic values in biomedical informatics. J Biomed Inform 2005, 38:404-415.

28. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG: Kalmokoff M, Lanthier P, Tremblay TL, Foss M, Lau PC, Sanders G, Shumaker VS, Yoder H, Akey JM, Green ED, Kharchenko PV,LY, Shamey AA, Eichten AE, Large-scale mapping and validation of expression regulation in Escherichia coli. J Mol Biol 2005, 34:6036-6039.

29. Zhang X, Zou S, Song J, Blouin E: Dif-ferential expression of Escherichia coli genes in response to high cell density. J Bacteriol 2002, 184:3161-3163.

30. Ploeg JR van der, Eichhorn E, Le L, Lebron C, Fournier M, Alvarado R, Santoyo-Posada K, Smith KJ, Wullschleger SD, Tolkamih N, Hsu C, Bissantz K, Pastusek A, Gheysen G, Weng Y, Weiszmann C, Ongusaha P, Xu X, Xiao Z, Tomlinson B, Bossuyt X, Wu R, Shapleigh J, Dardalot C, Hahne M, Juvonen T, Hefter G, Ghatan M, Tosi M, Zoller M, Guiguerré B, Zhang G, Wang S, Bae Y, Zheng Q, Zhang J, Tang L, Su Z, Lai S, Yang C, Gu X, Chen X, Xu H, Zhao X, Wang H, Zhou Q, Zhao G, Liu L, Wang H, Tan Y, Wu Z, Zhang X, Sun Q: The ascorbate transporter of Escherichia coli. J Biol Chem 2008, 283:3986-3994.

31. Heer J, Card SK, Landay JA: ArgR-dependent repression of arginine and nitrate transport genes in Escherichia coli K-12. J Mol Biol 2007, 373:251-267.

32. Salzberg SL: JBrowse: a genome browser for visualization of large DNA datasets. Genome Res 2002, 12:377-385.

33. Blouin EJ, Bosman OWJ, van Huijum SA, Breitling R, Tijmjen L, Silvis R, Roerdink JB, Kuipers OP: FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. Bioinformatics 2007, 23:1161-1163.

34. McHugh JP, Rodríguez-Quinoñes F, Abdul-Tehrani H, Svinutenko DA, Poole RK, Cooper CE, Andrews SC: Global iron-dependent gene regulation in Escherichia coli. A new mechanism for iron homeostasis. J Biol Chem 2003, 278:29478-29486.

41. Kalmokoff M, Lanthier P, Tremblay TL, Foss M, Lau PC, Sanders G, Shumaker VS, Yoder H, Akey JM, Green ED, Kharchenko PV,LY, Shamey AA, Eichten AE, Large-scale mapping and validation of expression regulation in Escherichia coli. J Mol Biol 2005, 34:6036-6039.

42. Guerry P, Ewing CP. Schirm M, Lorenzo M, Kelly J, Pattarini D, Majam J, Thibault P, Logan S: Changes in flagellin glycosylation affect Campylobacter autoagglutination and virulence. Mol Microbiol 2006, 60:299-311.

43. Faith JJ, Driscoll ME, Fuseara VO, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS: Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. Nucleic Acids Res 2007, 35:D102-D109.

44. Ohno M, Araki M, Goto S, Kato Y, Motoki T, Yamanishi Y: Identification of additional genes belonging to the LexA regulon in Escherichia coli. Mol Microbiol 2000, 35:1560-1572.

45. Lesley SA, Graziano J, Cho CY, Knuth MW, Klock HE: Gene expression response to misfolded protein as a screen for soluble recombinant protein. Protein Eng 2002, 15:153-160.

46. Zhang Z, Abuolwafa M, Smith MH, Saier MH: The ascorbate transporter of Escherichia coli. J Bacteriol 2003, 185:2243-2250.

47. Gyaneshwar P, Pally O, McAuliffe J, Popham DL, Jordan MI, Kustu S: Sulfur and nitrogen limitation in Escherichia coli K-12: specific homeostatic responses. J Bacteriol 2005, 187:1074-1090.

48. Moore CM, Nakano MM, Wang T, Ye RW, Helmann JD: Response of Bacillus subtilis to nitric oxide and the nitrosating agent sodium nitroprusside. J Bacteriol 2004, 186:6003-6014.

49. Eichenberger P, Fujita M, Jensen ST, Conlon EM, Rudner DZ, Wang ST, Ferguson C, Haga K, Sato T, Liu J, Losick R: The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. Proc Natl Acad Sci USA 2004, 101:2328.

50. Merrell DS, Thompson LJ, Kim CC, Mitchell H, Tompkins LS, Lee A, Falkow S: Growth phase-dependent response of Helicobacter pylori to iron starvation. Infect Immun 2003, 71:6510-6525.

51. Amati G, Bisicchia P, Galzetti A: DegU-P represses expression of the motility fla-che operon in Bacillus subtilis. J Bacteriol 2004, 186:6003-6014.

52. Kalmokoff M, Lanthier P, Tremblay TL, Foss M, Lau PC, Sanders G, Shumaker VS, Yoder H, Akey JM, Green ED, Kharchenko PV,LY, Shamey AA, Eichten AE, Large-scale mapping and validation of expression regulation in Escherichia coli. J Mol Biol 2005, 34:6036-6039.