OPEN ACCESS Freely available online

PLoS one

Functional Annotation of Hierarchical Modularity

Kanchana Padmanabhan[1,2], Kuangyu Wang[3], Nagiza F. Samatova[1,2]*

1 Department of Computer Science, North Carolina State University, Raleigh, North Carolina, United States of America, **2** Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **3** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America

Abstract

In biological networks of molecular interactions in a cell, network motifs that are biologically relevant are also functionally coherent, or form functional modules. These functionally coherent modules combine in a hierarchical manner into larger, less cohesive subsystems, thus revealing one of the essential design principles of system-level cellular organization and function–hierarchical modularity. Arguably, hierarchical modularity has not been explicitly taken into consideration by most, if not all, functional annotation systems. As a result, the existing methods would often fail to assign a statistically significant functional coherence score to biologically relevant molecular machines. We developed a methodology for hierarchical functional annotation. Given the hierarchical taxonomy of functional concepts (e.g., Gene Ontology) and the association of individual genes or proteins with these concepts (e.g., GO terms), our method will assign a Hierarchical Modularity Score (HMS) to each node in the hierarchy of functional modules; the HMS score and its $p-$value measure functional coherence of each module in the hierarchy. While existing methods annotate each module with a set of "enriched" functional terms in a bag of genes, our complementary method provides the hierarchical functional annotation of the modules and their hierarchically organized components. A hierarchical organization of functional modules often comes as a bi-product of cluster analysis of gene expression data or protein interaction data. Otherwise, our method will automatically build such a hierarchy by directly incorporating the functional taxonomy information into the hierarchy search process and by allowing multi-functional genes to be part of more than one component in the hierarchy. In addition, its underlying HMS scoring metric ensures that functional specificity of the terms across different levels of the hierarchical taxonomy is properly treated. We have evaluated our method using *Saccharomyces cerevisiae* data from KEGG and MIPS databases and several other computationally derived and curated datasets. The code and additional supplemental files can be obtained from http://code.google.com/p/functional-annotation-of-hierarchical-modularity/ (Accessed 2012 March 13).

Citation: Padmanabhan K, Wang K, Samatova NF (2012) Functional Annotation of Hierarchical Modularity. PLoS ONE 7(4): e33744. doi:10.1371/journal.pone.0033744

Editor: Yong He, Beijing Normal University, Beijing, China

Received October 12, 2011; **Accepted** February 16, 2012; **Published** April 4, 2012

Funding: This work was supported in part by the U.S. Department of Energy (DOE), Office of Science, the Office of Advanced Scientific Computing Research (ASCR) and the Office of Biological and Environmental Research (BER) and the U.S. National Science Foundation (Expeditions in Computing). Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. DOE under contract no. DEAC05-00OR22725. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Samatova@csc.ncsu.edu

Introduction

Network motifs are recurring, statistically significant patterns of node interactions that act as building blocks of complex networks [1]. In biological networks of molecular interactions in a cell, such as protein-protein interaction (PPI) networks or gene transcriptional regulatory networks (TRN), network motifs that are biologically relevant are also *functionally coherent*, or form *functional modules* [2], such as a ribosomal module synthesizing proteins or a signal transduction system governing bacterial chemotaxis. These functionally homogenous modules combine in a hierarchical manner into larger, less cohesive subsystems, thus revealing one of the essential design principles of system-level cellular organization and function–*hierachical modularity* [3,4].

Hierarchical modularity manifests itself at various levels of cellular organization. At the metabolism level, for example, hierarchical modularity within *Escherichia coli* closely overlaps with known metabolic functions, such as pyrimidine metabolism [3].

At the regulation level, for instance, in the *E. coli* TRN network, network motifs without global regulators, such as feed forward loops and bi-fan motifs, form the multi-layered hierarchical structure without feedback regulation [5]. Analysis of the hydrogen-producing *Rhodopseudomonas palustris* transcriptome [6] also suggests the interplay between functionally coherent modules related to electron transport (*fixX*, *fixC*, *fixB*, *fixA*, *ferN*, *fer1*), co-factor synthesis (*nifB*, *nifV*, *nifQ*, *nifN*, *nifE*, *nifX*), assembly or stability (*nifW*, *nifS2*, *nifU*), and regulation (*nifA*). Likewise, the CD4+ T-cell modules involved in human immune protection and regulation are made up of polarizing cues, lineage-specifying transcription factors, homing receptors, and effector molecules [7].

At the protein-protein interaction level, the discovered functional modules in the *Saccharomyces cerevisiae* PPI network consist of sub-components in the form of protein complexes and other macro-molecular assemblies [8]. For instance, the DNA replication, chromosome segregation, and chromatin assembly module consists of several submodules including DNA repair, DNA replication, chromosome segregation, origin recognition complex, anaphase promoting complex, spindle pole body, and chromatin assembly [9].

Thus, these examples provide a strong support not only for the network modularity principle introduced by Hartwell *et al.* [2] but

Table 1. Statistical significance of protein pairs' functional coherence in *Saccharomyces cerevisiae*.

Protein pair				*p*-value (pair/module/module size)				Ref.
ID	Description	ID	Description	HMS	[20,21]	[25]	[26]	[27,28]	
SNU13	RNA binding protein	*DIB1*	17-kDa component of the U4/U6aU5 tri-snRNP	0.0/0.0/2	0.23	0.01/0.01/2	0.1/0.1/2	0.42/0.42/2	[56]
HAP1	Zinc finger transcription factor involved in the complex regulation of gene expression in response to levels of heme and oxygen	*RPM2*	Protein subunit of mitochondrial RNase P	0.0/0.01/3	0.214	0.1/0.337/3	0.02/1.0/3	0.51/1.0*/3	[57]
SRB2	Subunit of the RNA polymerase II mediator complex	*RPB9*	RNA polymerase II subunit B12.6	0.0/0.01/58	0.48	0.12/1.0*/58	0.333/1.0/58	0.98/1.0*/58	[55]
NSR1	Nucleolar protein that binds nuclear localization sequences	*DBP2*	Essential ATP-dependent RNA helicase of the DEAD-box protein family	0.0/0.0/2	0.44	0.1/0.1/2	0.13/0.13/2	0.74/0.74/2	[58]

*assigned a *p*-value of 1 because the tool was unable to find a score for the entire module.

doi:10.1371/journal.pone.0033744.t001

also for the hierarchical modularity as a generic principle of system-level cellular organization and function [3].

Arguably, hierarchical modularity has not been explicitly taken into consideration by most, if not all, functional annotation systems [10,11]. Instead, a functional module is traditionally viewed as a "*bag of genes*," and methods that assess its functional coherence, or provide functional annotation, analyze this bag in its entirety. As a result, the existing methods would often fail to assign a statistically significant functional coherence score to biologically relevant molecular machines (see Table 1).

To address this gap, we developed a methodology for hierarchical functional annotation of biological network motifs. Given the hierarchical taxonomy of functional concepts (e.g., Gene Ontology) and the association of individual genes or proteins with these concepts (e.g., GO terms), our method will assign a *Hierarchical Modularity Score* (**HMS**) to each node in the hierarchy of functional modules; the HMS score and its $p-$value measure functional coherence of each module in the hierarchy. While existing methods annotate each module with a set of "enriched" functional terms in a bag of genes, our complementary method provides the hierarchical functional annotation of the modules and their components that are hierarchically organized.

A hierarchical organization of functional modules often comes as a bi-product of cluster analysis of gene expression data or protein interaction data. Otherwise, our method will automatically build such a hierarchy by directly incorporating the functional taxonomy information into the hierarchy search process and by allowing multi-functional genes to be part of more than one component in the hierarchy. In addition, its underlying HMS scoring metric ensures that functional specificity of the terms across different levels of the hierarchical taxonomy is properly treated.

We have evaluated our method using *Saccharomyces cerevisiae* data from KEGG [12–14] and MIPS [15] and several other computationally derived and curated datasets [8,16–19]. We compared our method with several biological significance analysis methods [20–28]. The hierarchical modularity built by our method from a set of genes in various KEGG pathways produces biologically relevant modules, namely, at various levels of the hierarchy, the corresponding modules match quite well with the manually-curated hierarchy of pathways in KEGG. We have obtained similar results for the protein complexes in the MIPS database. We provide literature evidence for several functional modules that have been identified by HMS as signicant both at the protein pairs and at the module levels but have been missed by some existing methods.

Results

Benchmark data and tools

To evaluate the performance of our method, we first need to define (1) the model organism; (2) the benchmark data of known functional annotations for this organism; (3) the hierarchical taxonomy of functional terms, and (4) the state-of-the-art methods that are most suitable for our comparative analysis.

Saccharomyces cerevisiae is our model organism. The reason is that its genome annotation is mostly complete and manually curated by human experts [22]. Apart from annotation quality, the availability of functional module datasets, both manually curated and experimentally generated, for *S. cerevisiae* is advantageous for our method validation purposes.

For benchmark data, we plan to use both metabolic pathways from KEGG database [12–14] and protein complexes in MIPS

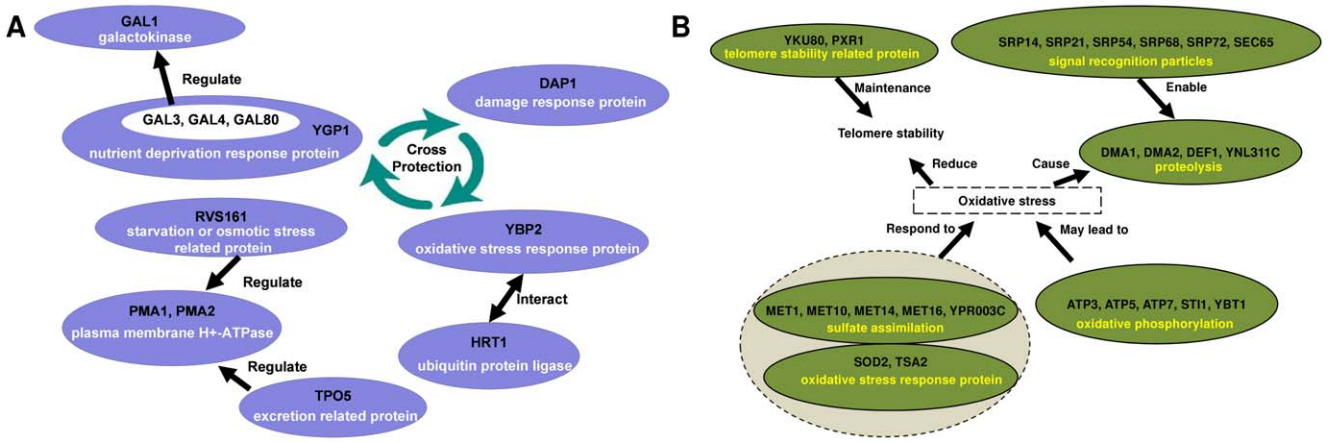


Figure 1. Functionally coherent modules from the Chen and Yuan [8] study. (A) Module ID M1 and (B) Module ID M3.
doi:10.1371/journal.pone.0033744.g001

database [15] including experimental protein-protein interaction data and protein complexes derived from this data [8,16–19].

For the hierarchical taxonomy of functional terms, we will rely on the commonly-used functional annotation taxonomy provided by the Gene Ontology consortium [10]. As such, we will limit ourselves to the existing methods that are also based on GO ontology. Namely, we will compare our method with the ones by Pandey *et al.* [20,21], Chen *et al.* [22], and GS^2 [23] methods. The former makes use of the lowest common ancestor principle to score functional coherence for a protein pair; it is based on Jiang and Conrath's scoring method [29], which is a normalized version of the scoring method in Resnik *et al.* [30]. It has been shown that Jiang and Conrath's method is the best measure to capture *semantic relatedness* [31]. The method by Chen *et al.* is based on a widely used *cosine* similarity measure to assess functional coherence for a protein pair and the authors provide a Matlab implementation for the same. The GS^2 [23] uses the overlap similarity measure, and the authors provide a Python implementation for the same. Additionally, we perform comparisons with methods described in [23–28]. These methods [24–28] have web-based implementations. The *p*-value for our method is calculated using the Monte Carlo procedure [32] and is discussed in detail in the *Methods* section.

We conducted three major types of performance evaluation: (1) at the level of functional coherency for protein pairs; (2) at the level of functional coherency for protein functional modules (with two or more proteins in each); and (3) functional annotation of reconstructed hierarchical functional modules. Both large-scale comparative analysis and small-scale literature mining based validation are performed.

Functional coherency of protein functional modules

Detailed biological analysis of modules from Chen and Yuan [8]. Two functional modules, M1 (Figure 1.A) and M3 (Figure 1.B), with the same ID's as in [8], have been reported as insignificant by several existing functional enrichment analysis methods (see Table 2). We used the web-based implementations for the functional enrichment analysis methods. However, the modules were identified as significant by our HMS method. In the following paragraphs, we provide biological evidence for the subtrees in the functional hierarchy of the two modules.

In module M1 (see Figure 1.A), *GAL1* is the galactose structural gene and *GAL3*, *GAL4*, and *GAL80* are transcriptional regulators involved in activation of the *GAL* genes in response to galactose; they form a sub-module in the hierarchy. The pair-wise functional associations between these genes are well-documented. Transcription of the galactose pathway genes in *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Kluyveromyces lactis* (*K. lactis*) is induced by galactose through the activities of the regulatory proteins, *GAL4*, *GAL80*, and *GAL3* (*S. cerevisiae*) or *GAL1* (*K. lactis*) [33,34]. *GAL4* binds to its binding sites in both the absence and the presence of galactose [35]; it has the capacity to activate transcription, while *GAL80* inhibits *GAL4* in the absence of galactose [36]. At the presence of galactose, *GAL3* (*GAL1* in *K. lactis*) binds to *GAL80* that alleviates the inhibition effect of *GAL80* upon *GAL4* [37].

PMA1 and *PMA2* form another sub-module that encodes plasma membrane H+-ATPase (PM-H+-ATPase), an enzyme with

Table 2. Functional modules evaluated using existing enrichment analysis tools in comparison with HMS.

	p-value			
Module ID	**[25]**	**[27,28]**	**[26]**	**HMS**
M12	1.02E-12	3.4E-19	5.73E-01	0.00
M94	1.05E-07	4.0E-7	6.03E-04	0.00
M3	1.0	0.1	1.0	0.02
M1	1.0	0.18	1.0	0.04

The first two rows show two homogeneous functional modules and the next two rows of the table show heterogeneous functional modules that have coherent submodules. Functional modules have been obtained from Chen and Yuan [8] of the *Saccharomyces cerevisiae* PPI network.
doi:10.1371/journal.pone.0033744.t002

critical physiological roles both in the absence or presence of environmental stress. *PMA2*, showing 89% identity to *PMA1* at the amino acid sequence level, encodes an H+-ATPase that is functionally interchangeable with the one encoded by *PMA1* [38].

The third sub-module involves *DAP1*, the damage response protein, and *YGP1* induced by nutrient deprivation-associated growth arrest. *DAP1* is required for growth in the presence of the methylating agent methyl methanesulfonate (MMS). *DAP1* is required for cell cycle progression following damage [39], while *YGP1* is induced after exposing cells to nutrient limitation [39]. It has already been demonstrated that exposure to one kind of stress can activate protective mechanisms against other different stresses, a phenomenon known as cross-protection [40]. Since *DAP1* and *YGP1* act both in the process of stress response, cross-protection might associate these two genes together.

The same relationship based on cross-protection can be observed in another sub-module that consists of *YBP2* that plays the role in resistance to oxidative stress and *HRT1* that is involved in stress response. The transcription factor *YBP2* and its homologue play central roles in the determination of resistance to oxidative stress [41], while *HRT1* forms ubiquitin ligase complex with other scaffold proteins [42]. The critical stress response factor *Nrf2* has been shown to be repressed by the ubiquitin-proteasome system under normal, unstressed conditions, with *Nrf2* exploiting ubiquitin ligase complexes [43].

The next module is made up of *PMA1* and *TPO5* that are involved in excretion of putrescine and spermidine. *TPO5* functions as a suppressor of cell growth by excreting polyamines [44]. *PMA1* is a polytopic membrane protein, whose essential physiological function is to pump protons out of the cell. Both the excretion of putrescine by *TPO5* and the delivery of *PMA1* to cell surface rely on secretory pathway. Furthermore, small portions of *TPO5* are co-localized with *PMA1* in plasma membrane, which indicates possible interactions between these two proteins [45].

PMA1 also forms a sub-module together with *RVS161* that regulates polarization of the actin cytoskeleton. *RVS161* regulates secretory vesicle trafficking [46] as well as cell polarity [47], actin cytoskeleton polarization [48], and endocytosis [49]. It is already known that the efficient delivery of *PMA1* to cell surface relies on secretory pathway [45]. Thus, *RVS161* has a regulatory effect upon *PMA1*.

Genes in this module have coherent functions, namely more than half of the proteins in this module are related to stress response, five out of 20 total have regulatory roles in cell cycle, four out of 20 total are evolved in endocytosis. Stress conditions are likely to cause cell cycle arrest, as well as endocytosis induction.

For module M3 (see Figure 1.B), the vast majority of the genes in the module enjoy oxidative stress response as the common theme. *SOD2* protects cells against oxygen toxicity and *TSA2*, responsible for the removal of reactive oxygen, directly protects cells against oxidative stress, while *PXR1* plays the role in negative regulation of telomerase, and *YKU80*, a subunit of the telomeric Ku complex, contributes to the maintenance of telomere stability, since oxidative stress is likely to induce telomere attrition [50].

Meanwhile, proteolysis could also be the result of oxidative stress: *YNL311C* is part of an ubiquitin protease complex, *DEF1* enables ubiquitination, *DMA1* is involved in ubiquitin ligation, and *DMA2* is involved in ubiquitination [51]. Since proteolysis involves many protein transportation processes, the signal recognition particles are essential to enable transportation: *SRP14*, *SRP21*, *SRP54*, *SRP68*, *SRP72*, and *SEC65* are all part of the signal recognition particle (SRP) subunit, and appear in module M3.

Furthermore, Wu *et al.* [52] showed that repression of sulfate assimilation is an adaptive response of yeast to the oxidative stress of zinc deficiency, while we notice that *MET1*, *MET10*, *MET14*, *MET16*, and *YPR003C* are basic proteins or protein subunits that are required for sulfate assimilation. Finally, oxidative phosphorylation produces ATP by utilizing electron transport trains. As a result, the inhibition of electron transport chain will lead to oxidative stress [53]. That is probably why *ATP3*, *ATP5*, and *ATP7* are all part of the enzyme complex required for ATP synthesis. Also, *STI1*, ATPase inhibitor activity, and *YBT1*, ATPase activity, coupled to transmembrane movement of substances, are part of the module.

Table 3. Percentage of significant (p-value ≤ 0.05) and highly significant (p-value ≤ 0.001) functionally coherent modules from Chen and Yuan [8] and CYS2008 [19].

Dataset	Method	Significant	Highly significant
CYS2008 protein complex database [19]	HMS	96.7%	82.9%
	GS^2 [23]	79.5%	40.1%
Chen and Yuan [8]	HMS	63.5%	46.4%
	GS^2 [23]	42.6%	29.8%

doi:10.1371/journal.pone.0033744.t003

Large-scale analysis of protein functional modules. Protein functional modules predicted by Chen *et al.* using their betweenness-based network partitioning algorithm [8] and protein complexes from CYS2008 database [19] are analyzed as modules for their functional coherency. Table 3 summarizes the results obtained by our HMS scoring method and GS^2 [23] method for both *significant* ($p-\text{value} \leq 0.05$) and *highly significant* ($p-\text{value} \leq 0.001$) cut-offs. HMS predicted 96.7% of the CYS complexes and 63.5% of the modules from Chen and Yuan study to be significant. GS^2 predicted 79.5% of the CYS complexes and 42.6% of the modules from Chen and Yuan [8] to be significant. The results can be found in Supplement S1.

HMS comparison with protein-set semantic similarity scoring metric. Protein pairs from the same protein complexes in [15–18] or the same metabolic pathways in KEGG [12–14] are assessed for functional coherency using HMS scoring method, the cosine similarity metric [22], the Jaccard similarity metric [22], and the GS^2 [23] method. We filter our results as being *significant* ($p-\text{value} \leq 0.05$, Figure 2.A) and *highly significant* ($p-\text{value} \leq 0.001$, Figure 2.B).

For MIPS-curated data [15], HMS, cosine, and Jaccard methods predicted nearly 100% of the protein functional modules as being functionally coherent, while GS^2 only predicted 84% to be significant. For Ho *et al.* data [17], HMS, on average, provided 30% higher predictions than other methods. For Krogan *et al.* data [16], HMS performed 20% better than the other methods, on average. For KEGG data, our HMS method, on average, performed 8% better than the other methods. For Gavin *et al.* data [18], HMS performed about 15% better, on average. Additionally, Chagoyen *et al.* [22] mentioned some complexes and pathways that in spite of being functionally related were predicted incoherent. We list some of those modules in Table 4 and show that our HMS method is able to predict them as functionally related. The results can be found in Supplement S1.

Functional coherency of protein pairs

HMS comparison with pair-wise semantic similarity metrics. We also calculated HMS score for 150 functionally-

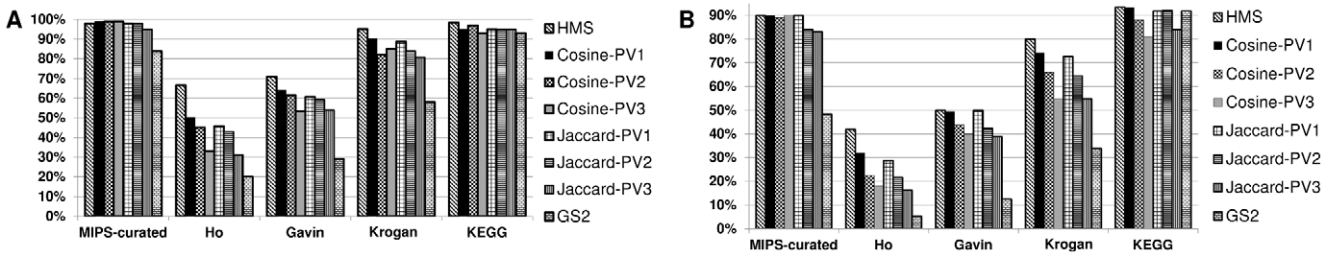


Figure 2. Functional coherence analysis of protein complexes and pathways. Functional coherence analysis of protein complexes from MIPS-curated [15], Ho [17], Gavin [18], and Krogan [16] as well as metabolic pathways from KEGG. Comparison between our HMS scoring, cosine similarity with different p-value methods from [22], Jaccard similarity with different $p-$value methods from [22] and GS^2 [23] methods. (A) Significant Modules ($p-\text{value} \leq 0.05$ and (B) Highly Significant Modules ($p-\text{value} \leq 0.001$).
doi:10.1371/journal.pone.0033744.g002

associated protein pairs and compared these scores with the HMS scores for an equal number of non-functional protein pairs in *S. cerevisiae*. The former were obtained from STRING [54] with a strong functional association score of 999 out of 999. The latter were sampled from those pairs that were not scored in STRING (i.e., there is no evidence for their functional association). We also performed a similar analysis using four other pair-wise protein similarity scores, Pandey *et al.* [20,21] metric, GS^2 [23] metric, overlap score [24], and cosine similarity [22,24]. The results of the analysis are summarized in Table 5. For all methods, the mean score for the functionally-associated pairs is significantly different from the mean score for the non-functional pairs, but HMS has the lowest $p-$value. Additionally, we calculated the percentage of the total number of pairs whose score is lower than the maximum score of the non-functional pairs but greater than the minimum score of the functionally-associated pairs. We found that except for HMS and Pandey *et al.* [20,21], all the other methods have an overlap. This is one of the reasons why we selected Pandey *et al.* [20,21] method for comparison in the next section. The results can be found in Supplement S1.

We analyzed some functionally-associated protein pairs from STRING that were classified as functionally coherent and thus biologically relevant ($p-\text{value} \leq 0.05$) by our method, yet were assessed as incoherent by Pandey's *et al.* We found literature support for biological relevance of these protein pairs. The results are summarized in Table 1. *RPB9* and *SRB2* proteins are part of *RNA polymerase II holoenzyme* in *S. cerevisiae* [55]. *SNU13* and *DIB1* proteins have been shown to be associated with the U4/U6U5 pre-mRNA splicing small nuclear ribonucleoprotein (snRNP) complex [56]. *HAP1* and *RPM2* are related by the fact that *RPM2* is required for repression of the heme activator protein *HAP2* in the absense of heme [57]. When *NSR1* was used as a bait in the protein-fragment complementation assay (PCA), the experiment pulled out *DBP2* as one of its prey proteins [58].

Inferred hierarchy of functional modules

To assess the quality of the hierarchy of functional modules derived from a given "bag of genes" using our HMS scoring metric and the hierarchical modularity inference methodology described in the *Methods* section, we assess the consistency between the predicted hierarchy and the hierarchy of known functional concepts in KEGG and MIPS databases. Remind that HMS, by default, uses GO ontology as its hierarchical taxonomy of functional terms.

Consistency analysis for KEGG metabolic pathways. Note that each metabolic pathway is a functional module. We consider the genes from several metabolic pathways as one "bag of genes" to build the hierarchy of functional modules. If the constructed hierarchy of functional modularity is biologically relevant, then the genes in each pathway should form a subtree in the hierarchy and not be "contaminated" by the genes from the other pathways. We set the fuzziness to null before running the algorithm in order to be able to use standard clustering validation metrics like the Heidke Score [59], Gerrity Score [60], and Peirce Score [61].

Since KEGG is organized into a three level hierarchy, the pathways at the lower levels of the hierarchy are functionally more coherent. Hence, they should be harder to separate into different subtrees. This hierarchical specificity of the KEGG knowledgebase provides us with an opportunity to check both the specificity and the sensitivity of our hierarchical modularity inference method.

We build contingency tables to provide a mathematically and statistically sound way for assessing the performance at large-scale. To construct a contingency table, the inferred hierarchy is first cut

Table 4. HMS results for some KEGG metabolic pathways and MIPS protein complexes [22] classified as insignificant by Chagoyen *et al.* [22].

		Chagoyen *et al.* [22]			HMS	
Pathway or Complex Name	**Size**	**pv1**	**pv2**	**pv3**	S_{HMS}	**p-value**
DNA helicases	2	4.12E-01	5.36E-01	5.36E-01	0.21	0.0
Mitochondrial processing complexes	4	1.05E-01	1.46E-01	2.73E-01	0.35	0.0
Tryptophan metabolism	16	7.67E-02	4.12E-01	5.08E-01	0.34	0.0
Lipoic acid metabolism	3	4.09E-01	4.69E-01	4.69E-01	0.50	0.0
Limonene and pinene degradation	6	2.59E-01	4.13E-02	1.66E-01	0.30	0.0

doi:10.1371/journal.pone.0033744.t004

Table 5. Comparison of pair-wise semantic similarity metrics using functionally-associated and non-functional protein pairs.

Method	Ref.	Functionally-associated Pairs			Non-functional Pairs				
		Mean	Std	Median	Mean	Std	Median	Overlap (%)	p-value
HMS		0.56	0.06	0.53	0.02	0.03	0.0	0	3.89E-61
GS^2	[23]	0.78	0.18	0.79	0.38	0.13	0.36	63.2	8.42E-74
Pandey	[20,21]	0.71	0.23	0.73	0.07	0.02	0.06	0	6.53E-29
Overlap Score	[24]	0.91	0.13	1.00	0.41	0.33	0.25	54.8	8.59E-40
Cosine similarity	[22,24]	0.78	0.17	0.80	0.20	0.08	0.20	6	3.05E-94

doi:10.1371/journal.pone.0033744.t005

at the level that produces s subtrees that are then compared with s pathways used as input to the algorithm. In the ideal scenario, all the genes in a given pathway (or row in the contingency table) will end up in the corresponding subtree (or the column in the contingency table) and vice versa; or the contingency table will form a diagonal matrix with the number of pathway genes along the diagonal and zero's on the off-diagonal elements of the table. By completing such a contingency table, we could then utilize various skill metrics, such as Heidke Score [59], Gerrity Score [60], and Peirce Score [61], to measure the goodness of the predicted hierarchical modularity.

We also performed all the experiments by replacing the S_{HMS} scoring metric with the one proposed by Pandey *et al.* [20,21] and compiled the results in Table 6. We found that at "Level 1" in the KEGG hierarchy, both methods had a perfect score of 1.0 for all three metrics, but as we moved down the hierarchy, we found that our method performed consistantly better than Pandey's *et al.* [20,21]. At "Level 2," we found that our method performed 6%, 7%, and 8% better in terms of the Heidke score, the Pierce score, the Gerrity score, respectively. At "Level 3," which is probably the hardest of the three in terms of pathways seperability, we performed about 13%, 6%, and 2% better for the same skill metrics. The results can be found in Supplement S2.

Consistency analysis for MIPS protein complexes. Protein complexes are functionally coherent modules, and hence experiments similar to the ones preformed using KEGG pathways can be designed. The results can be found in Table 7. We compared the mean score reported for our method and the one proposed by Pandey *et al.* We found that at "Level 1" in the MIPS hierarchy, our method performed 12% better than Pandey's *et al.* for both the Heidke and Pierce scores and 13% better for the Gerrity score. At "Level 2," our method performed approximately 26% better in terms of the Pierce Score and 35% and 26% better in terms of the Heidke and Gerrity scores, respectively. The results can be found in Supplement S2.

Table 6. Skill metrics for *Saccharomyces cerevisiae* KEGG experiments.

	KEGG	Heidke Score	Pierce Score	Gerrity Score
HMS	Level 1	1 ± 0	1 ± 0	1 ± 0
[20,21]		1 ± 0	1 ± 0	1 ± 0
HMS	Level 2	0.97 ± 0.04	0.98 ± 0.05	0.98 ± 0.05
[20,21]		0.91 ± 0.1	0.91 ± 0.12	0.90 ± 0.12
HMS	Level 3	0.90 ± 0.03	0.90 ± 0.05	0.90 ± 0.06
[20,21]		0.77 ± 0.14	0.84 ± 0.10	0.88 ± 0.07

doi:10.1371/journal.pone.0033744.t006

Effect of fuzziness

To evaluate the effect of incorporating fuzziness into the reconstruction of hierarchical modularity, we selected several KEGG pathways with common genes and then reconstructed the hierarchy with the fuzziness parameter $\mu=0.90$. For each pathway, we identified the corresponding cluster with the maximum gene overlap (at least 75%). We analyzed multi-pathway genes in terms of their membership in the corresponding clusters. Table 8 summarizes the results of the analysis for multi-pathway genes. Except for *UGA1* gene, which was missed in the cluster corresponding to the *Valine, leucine and isoleucine degradation* pathway, all the other genes were properly identified in their corresponding clusters.

Choosing ω_λ value

Our ω_λ selection strategy aims to optimize the method performance on a validation set of protein complexes, that are essentially known functional modules (Figure 3). This prior knowledge is derived from manually curated set of complexes from MPact-MIPS [15] database. Starting with the most conservative value of 1 for the ω_λ value, for each ω_λ value, we calculate the accuracy of identifying known protein complexes from the validation set as being statistically significant. We pick a value that is lenient enough to classify most of the known functional modules (manually curated protein complexes) as significant, while being stringent enough to avoid predicting random modules from getting high S_{HMS} scores. Thus, we select the largest ω_λ (in this case ($\omega_\lambda=10$)) value that ensures that at least 95% of the validation protein complex set is predicted as being

Table 7. Skill metrics for *Saccharomyces cerevisiae* MIPS experiments.

	MPact-MIPS	Heidke Score	Pierce Score	Gerrity Score
HMS	Level 1	1 ± 0	1 ± 0	1 ± 0
[20,21]		0.88 ± 0.25	0.88 ± 0.25	0.87 ± 0.26
HMS	Level 2	0.89 ± 0.14	0.90 ± 0.13	0.90 ± 0.13
[20,21]		0.54 ± 0.37	0.64 ± 0.27	0.64 ± 0.27

doi:10.1371/journal.pone.0033744.t007

Table 8. Consistency of multi-pathway genes across clusters that enrich the corressponding pathways.

	Genes								
Pathways	*ALD4*	*ALD5*	*ALD6*	*ERG10*	*ERG13*	*SHM1*	*SHM2*	*UGA1*	*POX1*
Propanoate metabolism	1/1	1/1	1/1	1/1	0/0	0/0	0/0	1/1	0/0
Valine, leucine and isoleucine degradation	1/1	1/1	1/1	1/1	1/1	0/0	0/0	1/0	0/0
Cyanoamino acid metabolism	0/0	0/0	0/0	0/0	0/0	1/1	1/1	0/0	0/0
Methane metabolism	0/0	0/0	0/0	0/0	0/0	1/1	1/1	0/0	0/0
beta-Alanine degradation	1/1	1/1	1/1	0/0	0/0	0/0	0/0	1/1	0/0
Synthesis and degradation of ketone bodies	0/0	0/0	0/0	1/1	1/1	0/0	0/0	0/0	0/0
Lysine degradation	1/1	1/1	1/1	1/1	0/0	0/0	0/0	0/0	0/0
Biosynthesis of unsaturated fatty acids	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1
Fatty acid metabolism	1/1	1/1	1/1	0/0	0/0	0/0	0/0	0/0	1/1

doi:10.1371/journal.pone.0033744.t008

statistically significant. The significance of a protein complex score is calculated using the Monte Carlo method discussed in the *Methods* section (using a $p-$value threshold of 0.05).

Discussion

Functional coherency analysis and *functional enrichment analysis* are two important concepts in genome annotation. Functional coherency analysis assesses if a set of genes or proteins are biologically relevant. Functional enrichment analysis determines if the distribution of a functional term in the set of genes is significantly different from the distribution of the same functional term in a background set of genes. Thus, functional coherency analysis scores a *functional module*, whereas functional enrichment analysis scores a *functional term*.

With functional enrichment analysis, it is sometimes difficult to conclude whether the set of genes is coherent. For example, for a set of 14 genes, one Gene Ontology (GO)-based functional enrichment analysis tool could infer that 9 out 14 genes are enriched with *GO term A* with a $p-$value of 0.001 and 11 out of 14 genes are enriched with *GO term B* with $p-$value of 0.434. Such inference creates ambiguity in deciding whether the set of genes is coherent. Therefore, it is functional coherence analysis that often determines whether the given functional module is biologically significant.

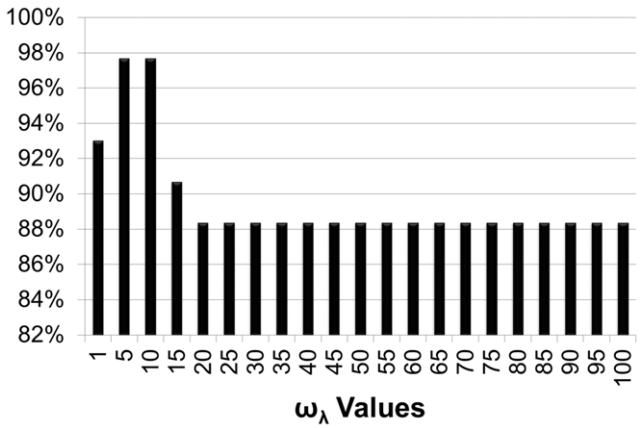


Figure 3. Effect of different values of ω_λ on the $S_{HMS}()$ score.
doi:10.1371/journal.pone.0033744.g003

Most existing functional enrichment analysis methods [25,26,62–70] assume that proteins in the same functional module perform the same function, or they are *functionally coherent*. Hence, for functionally coherent modules, all the proteins are annotated with the same functional term.

One of the main limitations of the existing functional enrichment analysis methods is that they require the use of a *background set of annotated genes*. As discussed by Shah and Fedoroff [63], the background could severely affect the assigned $p-$value, because this background information is directly incorporated into the scoring mechanism. Thus, functional enrichment scores that require a background set must be interpreted with caution. Unlike these functional enrichment analysis methods, we analyze functional coherency of proteins in the functional module without incorporating a *prior* knowledge about a background set into the scoring metric.

Both functional coherency analysis and functional enrichment analysis often rely on functional annotation taxonomies, such as the Gene Ontology (GO) [10] or FunCat [11] that are *hierarchical* by nature. Hence, any protein or gene associated with the child node is also associated with the parent node in the taxonomy. As discussed by Khatri and Drăghic [64], some tools [65–67] only utilize direct annotations, or functional terms associated with the child node. Yet, other tools [71–73] use functional terms associated only with the nodes at the user-specified hierarchical level; the more specific functional terms associated with the nodes below this level are replaced with a more generic parent's term at the user-defined level. Likewise, some methods take into consideration only the parent's term but not all its ancestors' [74]. And other tools use both [25,26,62,67,75–77]. Unlike these methods, we take all the levels of the hierarchy (a node and all its ancestors) into consideration while assessing a module for its functional coherence.

One of the drawbacks of these hierarchical taxonomy-based tools is their inability to differentiate between the functional terms that directly annotate the gene and those that annotate its ancestors; basically, they assign the same weight to both. Some methods [68,69] make this differentiation but in a statistically non-sound manner [69]. Unlike these methods, we utilize the hierarchical taxonomy of functional terms in its entirety, by discriminating between direct annotations and those associated with gene's ancestors.

Existing functional coherency analysis methods including the ones by Pandey *et al.* [20,21] and Chagoyen *et al.* [22] assess

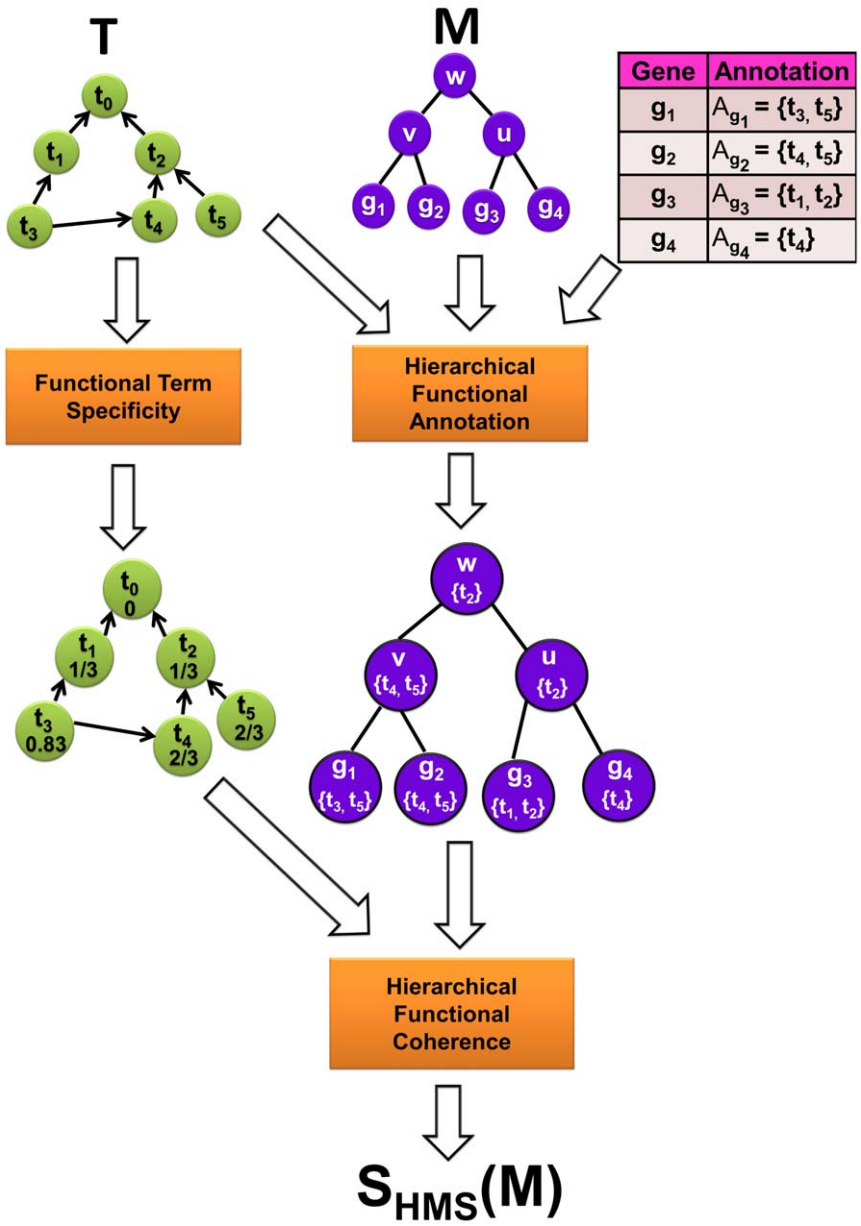


Figure 4. Functional annotation and coherence of hierarchical modules. The figure shows the overview of the methodology to assess functional coherence and assign annotation to hierarchical functional modules.
doi:10.1371/journal.pone.0033744.g004

functional coherence of a pair or of a set of genes using *GO term* annotations. The methods by Pandey *et al.* [20,21] are heavily based on Resnik's [30] information-theoretic score and its extension by Lin *et al.* [78]. This scoring incorporates the functional term's distribution for the background set directly into the scoring. The method by Chagoyen *et al.* [22] utilizes the cosine similarity to assess functional similarity between a pair of proteins. The number of functional terms that the protein is annotated with affects the scoring; multi-functional proteins will probably have more 1's in their term-vector. The method weighs the term's specificity but the weighting scheme is still based on the distribution of the term in the background set, and thus it has its drawbacks. In addition, these methods do not annotate the functional module. Unlike these methods, we do not measure functional specificity based on any background set. We calculate the specificity based on the position of a node in the hierarchical functional taxonomy. We also provide functional annotation for the module.

An important requirement for a good functional coherence scoring method is its ability to distinguish the number of levels traversed in the taxonomy to identify the common ancestor for a pair of proteins. This requirement is currently not incorporated into any of the existing methods. Also, the existing techniques assess functional coherency of the module in its entirety and do not take into consideration any structural information of the module. In contrast, our method addresses those and some other limitations.

The methods discussed so far directly rely on functional annotation taxonomies. In contrast, there are methods [79–81] that suggest mining biomedical literature and infer functional

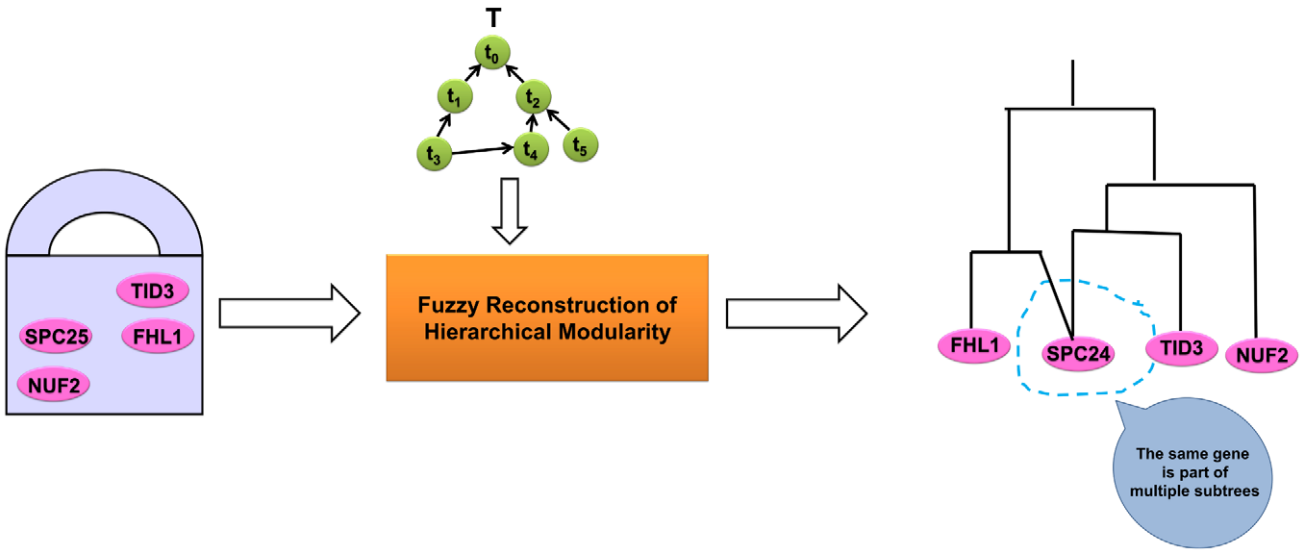


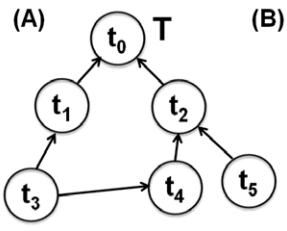
Figure 5. Overview of fuzzy reconstruction of hierarchical modularity.
doi:10.1371/journal.pone.0033744.g005

similarity of proteins in the set based on the semantic similarity of biological concepts, or topics, covered by various literature sources that reference these proteins. In this regard, these methods are complementary to the aforementioned ones. They are particularly suitable for gene sets with missing annotations (e.g., no GO terms are assigned). They could also be used for validating and/or comparing against the GO-based inference methodologies. It is worthwhile observing that, while biomedical literature is abundant, the analysis quality is dependent on the quality of the literature used. Additio'nally, some organisms are more heavily studied than others, and hence protein sets may be evaluated as insignificant purely on the basis that the knowledge about that set is not yet available. This problem can be compared to the problem of incomplete annotations of certain genomes, and hence the disadvantage of using functional annotation taxonomies is also present here.

The number of functional modules output by a computational method is in the order of hundreds and all of them cannot be tested via experimenatation by a biologist. Hence, functional coherence and significance methods can help narrow down the search space by only selecting the most promising modules. Our method goes one step further in that, given a hierarchical module it provides a global functional view, i.e., the entire picture about all the functions within a module and suggests clues on how various submodules within the module could relate to each other. Additionally, it scores the module keeping in mind the existing hierarchical structure.

A well-known hierarchical modularity principle suggests that protein modules are hierarchically organized; multi–functional proteins further suggest that such modules could be overlapping. Moreover, hierarchical taxonomies of functional terms manifested by GO ontology or by KEGG knowledgebase further suggest a possible hierarchical functional organization of the consituent submodules of the target module. Hence, given a *bag of genes* as a functional module, our method recreates its putative hierarchical functional view, while taking into consideration the fact that some proteins could be multi–functional. This kind of functional hierarchy could help with understanding the functioning of the module at various levels of functional specificity. For example, the overall function of the target module could be *chromosome segregation*, but at lower level of the functional hierarchy, we could find a submodule responsible for *proper alignment and attachment of chromosomes* and another submodule responsible for *translating the force generated by microtubule depolarization into movement to facilitate chromosome segregation* [8].

Additionally, our functional coherence method scores each submodule and uses this information to score the overall functional coherency of the module. Building the functional hierarchy for *a bag of genes* in a target module could additionally provide a clearer picture about the core and peripheral proteins for the functioning of this module. Since the method allows *fuzziness*, the core protein's interaction with a peripheral protein (that may not interact with any other protein in the module) could thus be captured. For example, if the "bag of genes" contains *CHD1, RAD16, VPS1, NHP10, ISW2, NHP6B, ISW1*, and *RSC6*, then the core of this module could include *CHD1, VPS1, NHP10, ISW2, NHP6B*, and *ISW1* [82], while the peripheral protein *RAD16* [82] could be functionally associated only with *VPS1* and the peripheral protein *RSC6* [82] could be associated with *ISW1*. Such information could thus be explicitly captured in the hierarchy, because both *VPS1* and *ISW1* could be associated with the core and the peripheral proteins when "fuzziness" is allowed.



(B)

ID	Level Set	S_{LS} (L)
L_0	$\{t_0\}$	0
L_1	$\{t_1, t_2\}$	1/3
L_2	$\{t_3, t_4, t_5\}$	2/3
L_3	$\{t_3\}$	1

Figure 6. An illustration of a hierarchical taxonomy T over the set of functional annotation terms $A=\{t_0,t_1,t_2,t_3,t_4,t_5\}$. (A) A DAG view. (B) A level set view.
doi:10.1371/journal.pone.0033744.g006

Methods

Our method provides two main functionalities:

1. Given a hierarchical module and a hierarchical functional taxonomy, our method can assess the functional coherence of the module and provide a hierarchical functional annotation. The overview of this functionality is provided in Figure 4.
2. Given a module as a "bag of genes" and a hierarchical functional taxonomy, our method can build the functional hierarchy of the module, i.e, provide a global functional view of the module. The overview of this functionality is provided in Figure 5.

In the following subsections we discuss the technical details of the two functionalities.

Hierarchical taxonomy of functional terms (HTFA)

Let $A=\{t_0,t_1,\ldots,t_n\}$, $n\in\mathbb{N}$, be a set of functional annotation terms. A functional annotation term (e.g., LYASE ACTIVITY) describes a function that a gene or a protein can carry out in the cell. A gene g can be annotated with a subset $A_g\subseteq A$ of functional annotation terms. If $|A_g|>1$, then g is multi-functional. If $A_g=\varnothing$, then g is called a hypothetical or unannotated gene. A functional term t_i is more specific than a functional term t_j, if it is a subtype of t_j. For example, lyase activity is a subtype of catalytic activity. Moreover, the same term can be a subtype of multiple terms. To capture functional specificity of terms, we will next define a hierarchical taxonomy of functional terms (HTFA).

A hierarchical taxonomy T_{t_0} of functional terms A is a directed tree or a directed acyclic graph (DAG) with the set V of labeled nodes (see Figure 6.A), such that

1. The labeling function $l: V(T_{t_0})\rightarrow A$ is a *bijection*, i.e., every node $v\in V(T_{t_0})$ is labeled with only one term $t\in A$, and each term t is assigned to only one node v, and
2. Label t_0 is assigned to only one node that is called the root node.

Whenever the context is clear, T and T_{t_0} will be used interchangeably. Likewise, we will simply use t to refer to the node v with label t (i.e., $l(v)=t$).

Due to its hierarchical nature, T can be represented as a level set $\mathbf{L}(T)=\{L_0,L_1,\ldots,L_{D_T}\}$ (see Figure 6.B), where $L_0=\{t_0\}$ and level $L_d\subseteq V$ is a set of nodes visited at distance d from the root t_0 during the depth-first traversal of T, and $D_T\in\mathbb{N}$ is the tree depth. Note that if T is a DAG (e.g., the Gene Ontology [10]), then $L_i\cap L_j\neq\varnothing$ for some $i\neq j$. In other words, the node can occur at different levels in the taxonomy.

A pair of nodes t_i and t_j in T_{t_0} forms an *ancestor* relationship ($t_i\prec t_j$), if there is a simple directed path from t_i to t_j in T_{t_0}. An ancestor relationship between a pair of nodes t_i and t_j in T_{t_0} represents a *functional specificity* relationship, namely, a functional term of the child node is a subtype of the functional term of its parent, grandparent, grandgrandparent, and so on. This relationship is transitive, i.e, $t_i\prec t_j$ and $t_j\prec t_k$ imply $t_i\prec t_k$. Also, a child can have multiple parents, as in a DAG.

Given this fact, we next introduce the *functional term specificity score* $S_{FTS}(t)$ for a node $t\in T$ as follows:

$$S_{FTS}(t)=\frac{\sum_{t\in L}S_{LS}(L)}{\sum_{L\in\mathbf{L}}\delta(t,L)},\tag{1}$$

where $S_{LS}(L)$ (see Figure 6.B) is the level specificity score associated with level $L\in\mathbf{L}$ and defined as

$$S_{LS}(L_d)=\frac{d}{D_T},\tag{2}$$

and $\delta()$ is a term characteristic function that specifies whether the term t occurs at level L:

$$\delta(t,L)=\begin{cases}1, & \text{if } t\in L\\ 0, & \text{if } t\notin L\end{cases}\tag{3}$$

A distinct pair $(t_i,t_j)\in A\times A$ of functional terms is called *related*, if the corresponding nodes in T form an ancestor relationship, i.e., $t_i\prec t_j$. More generally, a set of terms $U\subseteq A$ is called an *unrelated* set, or an *unrelated term set* in T, if no distinct pair (t_i,t_j), s.t. $t_i,t_j\in U$, is related in T.

Let U be an unrelated functional term set in T. Then, as defined by Equation 4, the *ancestor functional term set* $\mathbf{U}$ of U is the set of all the functional terms $t\in T$ on any simple path from any node $\hat{t}\in U$ to the root node $t_0\in T$:

$$\begin{aligned}&\mathbf{U}=U\cup P\cup\{t_0\},\\ &P=\{\forall t,t\in T:\exists\hat{t}\in U:\hat{t}\prec t\prec t_0\}\end{aligned}\tag{4}$$

For example, consider an unrelated functional term set $A_u=\{t_4,t_6\}$ in Figure 6. According to Equation 4, its ancestor functional term set is $\mathbf{A}_u=A_u\cup\{t_0,t_2,t_3\}$.

Hierarchical gene module (HGM)

Given a set of genes $G=\{g_1,g_2,\ldots,g_m\}$, a hierarchical gene module (HGM) M is an undirected tree over the set G of leaf nodes. Given the hierarchical taxonomy T of functional terms A, let an unrelated term set A_g, $A_g\subset A$, denote the functional annotation of gene g.

Hierarchical functional annotation. Given an HTFA taxonomy T and an HGM module M with a functional annotation $A_g\subset A$ for each leaf node gene g, *hierarchical functional annotation* of M is the function $h: V(M)\rightarrow\wp(A)$ that maps each node v in $V(M)$ to the set A_v from the power set of A such that:

1. A_v is the set of the *most specific common functional terms* among v's children, and
2. A_v is an unrelated functional term set in T.

Next, we will formally define the first condition, i.e., the set of the most specific common functional terms among the child nodes of v. Let C_v be the set of child nodes of v. Note that if $C_v=\varnothing$, then v is a leaf node g and $A_v=A_g$. Otherwise, as defined by Equation

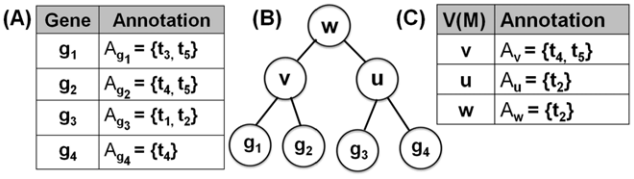


Figure 7. Hierarchical functional annotation of a gene module M for a gene set $G=\{g_1,g_2,g_3,g_4\}$ given the taxonomy T in Figure 6. (A) Functional annotation of genes in G by unrelated term sets in T. (B) A hierarchical gene module M. (C) The resulting annotation of the internal (non-leaf) nodes in $V(M)$.
doi:10.1371/journal.pone.0033744.g007

5, A_v is derived from the intersection of the ancestor functional term sets of v's children (see Equation 4) by maximizing the size of the unrelated functional term set U in the power set of this intersection:

$$A_v = \hat{U}, \tag{5}$$

$$\hat{U} = \underset{U \in \wp(\mathbf{A}_{C_v}) \,\&\&\, (U \text{ is unrelated})}{\operatorname{argmax}} |U|,$$

$$\mathbf{A}_{C_v} = \bigcap_{w \in C_v} \mathbf{A}_w$$

For a hierarchical functional module M in Figure 7 and a taxonomy T in Figure 6, consider $v \in V(M)$ as an example with $C_v = \{g_1, g_2\}$ and $\mathbf{A}_{C_v} = \{t_0,t_1,t_2,t_3,t_4,t_5\} \cap \{t_0,t_2,t_4,t_5\} = \{t_0,t_2,t_4,t_5\}$. Then the maximum size unrelated term set in the power set of this intersection defines the functional annotation set $A_v = \{t_4,t_5\}$ for the internal node v.

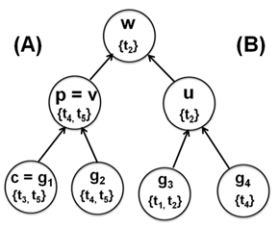
Functional coherence. Existing functional coherence analysis techniques analyze the input functional module in its entirety without considering its hierarchical structure. Additionally, most methods depend on a *reference set* by incorporating its annotation distribution and size into their scoring formula [20–22]. A reference set is a group of proteins that forms a superset of the functional module. Khatri *et al.* discuss the difficulties with selecting the right reference set [64].

Here, we introduce a method that accounts for the hierarchical structure of the module M when determining its functional coherence. Additionally, the scoring function does not directly rely on any reference set. More specifically, given the hierarchical gene module M and its functional annotation, the functional coherence score, called *hierarchical modularity score* (HMS), $S_{HMS}(M)$ of M is defined by Equation 6:

$$S_{HMS}(M) = \frac{1}{|V(M)|} \times \sum_{v \in V(M)} \left[\lambda(v) \times \frac{1}{|A_v|} \times \sum_{t \in A_v} S_{FTS}(t) \right], \tag{6}$$

where the functional term specificity $S_{FTS}(t)$ is defined by Equation 1, and the *penalization factor* $\lambda(v)$ is discussed in the following section.

Penalization factor (λ). Consider a hierarchical gene module M with its hierarchical functional annotation (see Figure 8.A), as described in *Hierarchical functional annotation* section. Let $p \in V(M)$ be a parent node with its children C_p.



(B)

c = g1 \ p = v	t4	t5
t3	1	∞
t5	∞	0
min	**1**	**0**
$\psi(p,c)$ = **0.5**		

Figure 8. Illustration of penalization factor calculation. (A) Hierarchical annotation of the functional module defined in Figure 7. (B) Dissimilarity score $\psi(p,c)$ for a parent $p = v$ and a child $c = g_1$.
doi:10.1371/journal.pone.0033744.g008

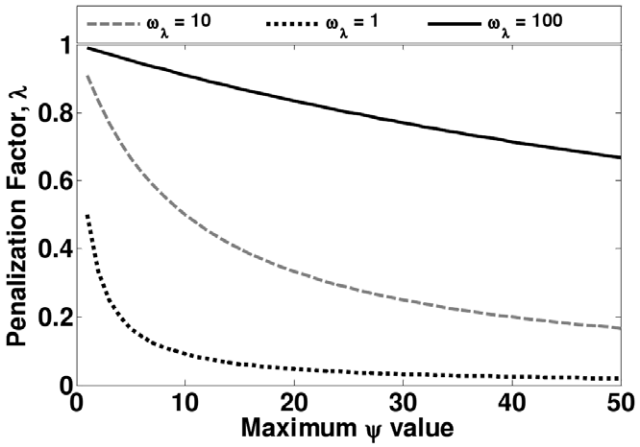


Figure 9. Comparison between the three penalization factor functions considered.
doi:10.1371/journal.pone.0033744.g009

Given the functional annotation term sets, A_p and A_c, for the parent p and its child $c \in C_p$, respectively, a dissimilarity score $\psi(p,c)$ between p and c is defined by Equation 7 (see Figure 8.B):

$$\psi(p,c) = \frac{1}{|A_p|} \times \sum_{t \in A_p} \min_{t' \in A_c} d(t,t'), \tag{7}$$

where the distance $d(t,t')$ is the length of the shortest simple path ($t' \prec t$) from node t' to t in T, or $d(t,t') = \infty$, if t' and t are not related.

Given Equation 7, the penalization factor $\lambda(p)$ for the parent node $p \in V(M)$ is then defined by Equation 8:

$$\lambda(p) = \left[1 + \frac{1}{\omega_\lambda} \max_{c \in C_p} \psi(p,c) \right]^{-1} \tag{8}$$

For the example in Figure 8, $\lambda(v) = 0.95$ for $\omega_\lambda = 10$, because the dissimilarity scores between v and its children g_1 and g_2 are 0.5 and 0, respectively. Also, Figure 9 depicts the behavior of $\lambda(p)$ for different values of ω_λ in Equation 8, as the maximum value of $\psi(p,c)$ varies from zero to its maximum possible value of the tree depth D_T in the taxonomy T. If ω_λ increases from one to 100, then node score's penalty decreases from 50% (even for immediate neighbors in $\psi()$) to 13% (for the largest taxonomy depth $D_T = 15$ in the Gene Ontology [10]). More information on choosing ω_λ values can be found i *Choosing ω_λ value* section.

Assessing statistical significance

To provide a robust assessment of statistical significance for $S_{HMS}(M)$, we measure an empirical $p-$value for $S_{HMS}(M)$ score assigned to each hierarchical module M using the Monte Carlo procedure described in [32]. Specifically, for hierarchical module M over a set of $|G|$ genes from organism O, we randomly sample N subsets of size $|G|$ from the entire genome of organism O, build the hierarchy, and compute the $S_{HMS}()$. Then, we estimate an empirical $p-$value for $S_{HMS}(M)$ as $p-\text{value} = R/N$, where N is the total number of random samples ($N \sim 1000$) and R is the number of the samples that produce a test statistics $S_{HMS}()$ greater than or equal to the $S_{HMS}(M)$.

Fuzzy reconstruction of hierarchical modularity

In *Hierarchical gene module* section, the hierarchical structure for a gene module M was provided as an input. Based on this structure and the hierarchical taxonomy of functional annotation terms (*Hierarchical taxonomy of functional terms* section), we provided means both for inferring M's hierarchical functional annotation (*Hierarchical functional annotation* section) and for estimating M's functional coherency via hierarchical modularity scoring (*Functional coherence* section).

In contrast, here we consider a somewhat inverse problem, namely, the reconstruction of a hierarchical structure of a functional module M defined by its gene set G. G is often referred as a "bag of genes." On the one hand, it seems that any hierarchical clustering method could be used to reconstruct the hierarchical functional modularity from such a "bag of genes." On the other hand, the presence of multi-functional genes suggests that the same gene could belong to multiple subtrees in the hierarchy–the property that is not often guaranteed by any hierarchical clustering method. Therefore, we will first need to introduce "fuzziness" into the process of building a functional hierarchy for G. For example, in Figure 5, a bag of genes containing *SPC24*, *TID3*, *NUF2*, and *FHL1* and a functional annotation taxonomy T are provided as input to the method. It is known that *SPC24*, *TID3*, and *NUF2* are functionally related because they are part of the Ndc80 protein complex but *SPC24* is also trnascriptionally regulated by *FHL1* [83] and so *SPC24* is part of multiple subtress and, hence, fuzziness is introduced.

Existing fuzzy clustering schemes typically introduce some fuzziness into some known clustering algorithm. C-means [84,85] is a typical example of this kind. Others are typically partitional by nature [27,28,86]. Agglomerative fuzzy clustering algorithms are not common, because agglomerative techniques are considered "hard clustering," i.e., it becomes difficult to move an element from an existing cluster to a new cluster. Ideally, any fuzziness in a clustering procedure should be introduced, while the hierarchy is being built and not as a post-processing step.

To meet these requirements, we propose a *taxonomy-based, agglomerative, fuzzy inference* (TAFI) of the hierarchical gene module M from a gene set G, provided each gene $g \in G$ is annotated with an unrelated functional term set $A_g \subset A$ in a hierarchical taxonomy T of functional annotation terms A (see *Hierarchical taxonomy of functional terms* section). The overview of this method is provided in Figure 5.

Similar to an agglomerative hierarchical clustering (AHC) process, TAFI starts with assigning each gene to its own cluster and proceeds building the hierarchy in an iterative, bottom-up manner, but it introduces fuzziness by allowing multiple cluster pairs to merge simultaneously at each iteration. The two user-defined parameters control this fuzziness process at each iteration: (a) the *merging factor* μ and (b) the *stopping criterion* τ_s. The former defines what cluster pairs get merged at a given iteration it. Namely, unlike traditional AHC that merge the pair of clusters with the maximum similarity $S_{max}(it)$, TAFI allows for clusters with suboptimal similarity to be merged as well. Suboptimality is defined by the percentage μ of $S_{max}(it)$. In addition, TAFI prevents the formation of unrelated clustering modules by stopping the bottom-up cluster merging process at iteration $\hat{it}$ as soon as $S_{max}(\hat{it})$ value falls below τ_s.

Note that Horng *et al.* [87] proposed to merge the cluster pairs whose similarity is greater than $S_{max}(i)-\Delta$ unlike TAFI 's way of restricting to $\frac{\mu}{100} \times S_{max}(i)$ similarity threshold. The reason behind our choice of multiplicative factor rather than additive/subtractive factor is the following. If $\Delta = 0.1$ and $S_{max}(i) = 0.5$, then any cluster pair with inter-cluster similarity greater than 0.4 would be merged. The value of 0.4 is 80% of 0.5. However, if $S_{max}(i) = 0.2$, then any cluster pair with inter-cluster similarity greater than 0.1 would be merged, but the value of 0.1 is only 50% of 0.2. The criterion becomes more stringent with a larger value of $S_{max}()$ and, conversely, it becomes more lenient, as $S_{max}()$ gets smaller. In contrast, our choice of the merging factor allows for resolving this inconsistency issue.

Also, observe that multiple merges at each iteration can sometimes result in the same subtree being formed repeatedly. This leads to redundancy. Thus, TAFI employs pruning, where a merge is allowed only if the merge results in a subtree that has not been already formed.

In addition, we need to make two important decisions in order to apply TAFI: (1) the inter-cluster similarity measure and (2) the linkage algorithm. For the inter-cluster similarity measure, we use Equation 6 that calculates the hierarchical modularity score $S_{HMS}(M)$ for a hypothetical module M that could be formed if the two clusters, or hierarchical tree modules M_1 and M_2, were merged by adding a new root node v_{new} and making the root nodes $v_1 \in V(M_1)$ and $v_2 \in V(M_2)$ to be the children of v_{new}.

It is worth noticing that $S_{HMS}(M)$ is a semi-metric, and this property has direct implications on our choice of the base clustering algorithm. Since semi-metrics do not adhere to the triangle inequality principle, we can resort to an average, single, complete, or centroid linkage algorithm as our base clustering technique. Therefore, the effective clustering techniques, such as Ward's method cannot be used in conjunction with semi-metrics [88].

Supporting Information

Supplement S1 Results discussed in *Functional coherency of protein functional modules* and *Functional coherency of protein pairs* sections.
(XLSX)

Supplement S2 Results discussed in *Consistency analysis for KEGG metabolic pathways* and *Consistency analysis for MIPS protein complexes* sections.
(RAR)

Author Contributions

Conceived and designed the experiments: KP NS. Performed the experiments: KP. Analyzed the data: KP KW. Wrote the paper: KP KW NS. Developed and implemented the methodology: KP. Provided the problem statement, supervised the developement of the computational methodology, and provided suggestions on methodology validation: NS.

References

1. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: Simple building blocks of complex networks. Science 298: 824–827.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: 47–52.
3. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabsi A (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551–1555.
4. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. PNAS 100: 12123–12128.
5. Ma H, Buer J, Zeng A (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics 5: 199.
6. Rey FE, Heiniger EK, Harwood CS (2007) Redirection of metabolism for biological hydrogen production. Appl Environ Microbiol 73: 1665–1671.
7. Sallusto F, Lanzavecchia A (2009) Heterogeneity of CD4+ memory T cells: Functional modules for tailored immunity. Eur J Immunol 39: 2076–2082.

8. Chen J, Yuan B (2006) Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22: 2283–2290.
9. Zhou H, Lipowsky R (2006) The yeast protein-protein interaction map is a highly modular network with a staircase community structure. Available: http://www.docstoc.com/docs/44073723/The-yeast-protein-protein-interaction-map-isa-highly-modular-network. Accessed 2012 Mar 15.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontologyy: tool for the unification of biology. Nat Genet 25: 25–29.
11. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539–5545.
12. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–D360.
13. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res 34: D354–D357.
14. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
15. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, et al. (2005) MPact: The MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436–D441.
16. Krogan NJ, Peng W, Cagney G, Robinson MD, Haw R, et al. (2004) High-definition macromolecular composition of yeast RNA-processing complexes. Mol Cell 13: 225–239.
17. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–183.
18. Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.
19. Pu S, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 37: 825–831.
20. Pandey J, Koyutrk M, Subramaniam S, Grama A (2008) Functional coherence in domain interaction networks. Bioinformatics 24: i28–i34.
21. Pandey J, Koyuturk M, Grama A (2010) Functional characterization and topological modularity of molecular interaction networks. BMC Bioinformatics 11: S35.
22. Chagoyen M, Carazo J, Pascual-Montano A (2008) Assessment of protein set coherence using functional annotations. BMC Bioinformatics 9: 444.
23. Ruths T, Ruths D, Nakhleh L (2009) GS^2: An e_ciently computable measure of GO-based similarity of gene sets. Bioinformatics 25: 1178–1184.
24. Mistry M, Pavlidis P (2008) Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 9: 327.
25. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::termfinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
26. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24: 1650–1651.
27. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
28. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
29. Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics. pp 19–33.
30. Resnik P (1999) Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 11: 95–130.
31. Budanitsky A (2001) Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: Proceedings of the North American Conference on Chinese Linguistics 2001 Workshop: on WordNet and other lexical resources: Applications, extensions, and customizations. pp 29–34.
32. North B, Curtis D, Sham P (2002) A note on the calculation of empirical p-values from Monte Carlo procedures. Am J Hum Genet 71: 439–441.
33. Johnston M (1987) A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces cerevisiae*. Microbiol Rev 51: 458–476.
34. Schaffrath R, Breunig KD (2000) Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. Fungal Genet Biol 30: 173–190.
35. Selleck SB, Majors JE (1987) *In vivo* DNA-binding properties of a yeast transcription activator protein. Mol Cell Biol 7: 3260–3267.
36. Pilauri V, Bewley M, Diep C, Hopper J (2005) Gal80 dimerization and the yeast gal gene switch. Genetics 169: 1903–1914.
37. Platt A, Reece RJ (1998) The yeast galactose genetic switch is mediated by the formation of a Gal4p-Gal80p-Gal3p complex. EMBO J 17: 4086–4091.
38. Schlesser A, Ulaszewski S, Ghislain M, Goffeau A (1988) A second transport ATPase gene in *Saccharomyces cerevisiae*. J Biol Chem 263: 19480–19487.
39. Hand R, Jia N, Bard M, Craven R (2003) *Saccharomyces cerevisiae* Dap1p, a novel DNA damage response protein related to the mammalian membrane-associated progesterone receptor. Eukaryot Cell 2: 306–317.
40. Quinn J (2008) Di_erent stress response between model and pathogenic fungi. In: Avery S, Stratford M, van West P, eds. Stress in yeasts and filamentous fungi Academic Press. pp 67–86.
41. Gulshan K, Rovinsky S, Moye-Rowley W (2004) YBP1 and its homologue YBP2/YBH1 inuence oxidative-stress tolerance by nonidentical mechanisms in *Saccharomyces cerevisiae*. Eukaryot Cell 3: 318–330.
42. Kikuchi Y, Oka Y, Kobayashi M, Uesono Y, Tohe A, et al. (1994) A new yeast gene, HTR1, required for growth at high-temperature, is needed for recovery from mating pheromone-induced G1 arrest. Mol Gen Genet 245: 107–116.
43. Kobayashi A, Kang M, Watai Y, Tong KI, Shibata T, et al. (2006) Oxidative and electrophilic stresses activate *Nrf2* through inhibition of ubiquitination activity of Keap1. Mol Cell Biol 26: 221–229.
44. Tachihara K, Uemura T, Kashiwagi K, Igarashi K (2005) Excretion of putrescine and spermidine by the protein encoded by YKL174C (TPO5) in *Saccharomyces cerevisiae*. J Biol Chem 280: 12637–12642.
45. Wang Q, Chang A (2002) Sphingoid base synthesis is required for oligomerization and cell surface stability of the yeast plasma membrane ATPase, *PMA1*. PNAS 99: 12853–12858.
46. Breton AM, Schaeffer J, Aigle M (2001) The yeast *RVS161* and *RVS167* proteins are involved in secretory vesicles targeting the plasma membrane and in cell integrity. Yeast 18: 1053–1068.
47. Durrens P, Revardel E, Bonneu M, Aigle M (1995) Evidence for a branched pathway in the polarized cell division of *Saccharomyces cerevisiae*. Curr Genet 27: 213–216.
48. Sivadon P, Bauer F, Aigle M, Crouzet M (1995) Actin cytoskeleton and budding pattern are altered in the yeast *RVS161* mutant: The *RVS161* protein shares common domains with the brain protein amphiphysin. Mol Gen Genet 246: 485–495.
49. Munn AL, Stevenson BJ, Geli MI, Riezman H (1995) *END5*, *END6*, and *END7*: Mutations that cause actin delocalization and block the internalization step of endocytosis in *Saccharomyces cerevisiae*. Mol Biol Cell 6: 1721–1742.
50. Liu L, Trimarchi J, Navarro P, Blasco M, Keefe D (2003) Oxidative stress contributes to arsenicinduced telomere attrition, chromosome instability, and apoptosis. J Biol Chem 278: 31998–32004.
51. Grune T, Reinheckel T, Joshi M, Davies K (1995) Proteolysis in cultured liver epithelial cells during oxidative stress. Role of the multicatalytic proteinase complex, proteasome. J Biol Chem 270: 2344–2351.
52. Wu C, Roje S, Sandoval F, Bird A, Winge D, et al. (2009) Repression of sulfate assimilation is an adaptive response of yeast to the oxidative stress of zinc deficiency. J Biol Chem 284: 27544–27556.
53. Garcia-Ruiz C, Colell A, Morales A, Kaplowitz N, Fernandez-Checa J (1995) Role of oxidative stress generated from the mitochondrial electron transport chain and mitochondrial glutathione status in loss of mitochondrial function and activation of transcription factor nuclear factor-kappa B: studies with isolated mitochondria and rat hepatocytes. Mol Pharmacol 48: 825–834.
54. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416.
55. Myer VE, Young RA (1998) RNA polymerase II holoenzymes and subcomplexes. J Biol Chem 273: 27757–27760.
56. Stevens SW, Abelson J (1999) Purification of the yeast U4/U6-U5 small nuclear ribonucleoprotein particle and identification of its proteins. PNAS 96: 7226–7231.
57. Hach A, Hon T, Zhang L (1999) A new class of repression modules is critical for heme regulation of the yeast transcriptional activator *HAP1*. Mol Cell Biol 19: 4324–4333.
58. Tai SL, Daran-Lapujade P, Walsh MC, Pronk JT, Daran J (2007) Acclimation of *Saccharomyces cerevisiae* to low temperature: A chemostat-based transcriptome analysis. Mol Biol Cell 18: 5100–5112.
59. Heidke P (1926) Berechnung des erfolges und der gte der windstrkevorhersagen im sturmwarnungsdienst. Geografika Annaler 8: 301349.
60. Gerrity J (1992) A note on Gandin and Murphy's equitable skill score. Mon Weather Rev 120: 2709–2712.
61. Peirce C (1884) The numerical measure of the success of predictions. Science 4: 453–454.
62. Maere S, Heymans K, Kuiper M (2005) BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics 21: 3448–3449.
63. Shah N, Fedoroff N (2004) CLENCH: A program for calculating cluster ENriCHment using the Gene Ontology. Bioinformatics 20: 1196–1197.
64. Khatri P, Drăghici S (2005) Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics 21: 3587–3595.
65. Castillo-Davis CI, Hartl DL (2003) GeneMerge|post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19: 891–892.
66. Berriz GF, King OD, Bryant B, Sander C, Roth FP (2003) Characterizing gene sets with FuncAssociate. Bioinformatics 19: 2502–2504.
67. Zhang B, Schmoyer D, Kirov S, Snoddy J (2004) GOTree machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 5: 16.
68. Alexa A, Rahnenführer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22: 1600–1607.
69. Jupiter D, Sahutoglu J, VanBuren V (2010) TreeHugger: A new test for enrichment of Gene Ontology terms. Informs J Comput 22: 210–221.

70. Young A, Whitehouse N, Cho J, Shaw C (2005) OntologyTraverser: An R package for GO analysis. Bioinformatics 21: 275–276.
71. Al-Shahrour F, Díaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578–580.
72. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R (2003) Identifying biological themes within lists of genes with EASE. Genome Biol 4: R70.
73. Beißbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20: 1464–1465.
74. Apostolico A, Guerra C, Istrail S, Pevzner P, Waterman M, et al. (2006) An improved statistic for detecting over-represented Gene Ontology annotations in gene sets. In: Research in Computational Molecular Biology, Springer Berlin/ Heidelberg, volume 3909 of *Lecture Notes in Computer Science*. pp 85–98.
75. Carmona-Saez P, Chagoyen M, Tirado F, Carazo J, Pascual-Montano A (2007) GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8: R3.
76. Zeeberg B, Feng W, Wang G, Wang M, Fojo A, et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28.
77. Masseroli M, Martucci D, Pinciroli F (2004) GFINDer: Genome function integrated discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res 32: 293–300.
78. Lin D (1998) An information-theoretic definition of similarity. In: In Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann. pp 296–304.
79. Zheng B, Lu X (2007) Novel metrics for evaluating the functional coherence of protein groups via protein semantic network. Genome Biol 8: R153.
80. Raychaudhuri S, Altman RB (2003) A literature-based method for assessing the functional coherence of a gene group. Bioinformatics 19: 396–401.
81. Raychaudhuri S, Schütze H, Altman RB (2002) Using text analysis to identify functionally coherent gene groups. Genome Res 12: 1582–1590.
82. Luo F, Li B, Wan XF, Scheuermann R (2009) Core and periphery structures in protein interaction networks. BMC Bioinformatics 10: S8.
83. Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, et al. (2010) Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol 8: e1000329.
84. Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. Comput Geosci 10: 191–203.
85. Tsekouras G, Sarimveis H, Kavakli E, Bafas G (2005) A hierarchical fuzzy-clustering approach to fuzzy modeling. Fuzzy Set Syst 150: 245–266.
86. Geva A (1999) Hierarchical unsupervised fuzzy clustering. IEEE Trans Fuzzy Syst 7: 723–733.
87. Horng Y, Chen S, Chang Y, Lee C (2005) A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. IEEE Trans Fuzzy Syst 13: 216–228.
88. Lance GN, Williams WT (1967) A general theory of classificatory sorting strategies. The Computer Journal 9: 373–380.