ABSTRACT

Gene enrichment analysis is a common technique for highlighting molecular pathways and biological processes of a phenotype. Such technique has recently evolved exploiting the information contained in biological networks. We developed NET-GE, a web server for network-based gene enrichment analyses. NET-GE defines functional associations between a list of genes/proteins and biological processes or pathways by identifying function-specific modules in a molecular interaction network. The peculiarity of NET-GE is the possibility to enrich terms not detectable by standard enrichment procedure. Here, we highlight with two specific applications the performances of NET-GE by computing which functional phenotypes can be associated with two different sets of genes related to Attention Deficit Hyperactivity Disorder and to an Obsessive-compulsive disorder, respectively.

KEYWORDS

Gene enrichment analysis; network-based gene enrichment analysis; functional association

INTRODUCTION

Technologies capable of investigating the organism complexity at different levels of resolution have been revolutionizing healthcare practice [1]. Genomic data are generated more and more to better define, at molecular levels, the origin of the different phenotypes. From a precision/genomic medicine prospective, such phenotypes need annotations in order to reconcile variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder (ADHD) and the Obsessive-compulsive disorder (OCD). One of the main features of NET-GE is the possibility to enrich terms that are not originally present in the annotation of the starting gene/protein set (and thus not detectable through a standard enrichment). When tested on benchmark sets retrieved from the Online Mendelian Inheritance in Man (OMIM) resource (https://www.omim.org), NET-GE was able to enrich sets of genes related to the same disease, also highlighting new terms (i.e. terms not included in the annotations of the input set) [11].

Here, we present two study cases, demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder (ADHD) and the Obsessive-compulsive disorder (OCD).
METHODS

NET-GE background

The network-based enrichment makes use of precomputed annotation terms, as previously described [11]. Briefly, the human molecular-interaction network was downloaded from STRING v.10 (http://string-db.org). A second version of STRING, named STRING0.9, was obtained by retaining only the links with the STRING combined score $\geq 0.9$. The database for annotating features were: GENE ONTOLOGY (as retrieved from the UniProt-GOA human 145 web resource: http://www.ebi.ac.uk/GOA); KEGG PATHWAY v77 and REACTOME PATHWAY v53. For each annotating feature, proteins sharing the same annotation term were collected in a seed set and then extended into a compact and connected module of the molecular-interaction network. Thus, the module was determined by computing all the shortest paths among the seeds genes/proteins and then by reducing the resulting sub-network into the minimal connecting network that preserves the distances among seeds. The minimal connecting network adds to the seeds a set of connecting nodes that are more reliably related to the reference annotation. Details about annotations and module extraction can be found in [11] and [12], respectively.

Over-representation analysis is performed by mapping the input set on each module and determining, through a Fisher’s exact test, whether there are significant overlaps between the input set and the modules (seed sets in the case of standard enrichment). Multiple testing correction is then applied using the Bonferroni or the Benjamini-Hochberg (False Discovery Rate, FDR) procedure [13].

When we consider the standard enrichment, the background set is totally disconnected. On the contrary, with the network-based procedure we rely on the human interactome to precompute the annotation modules. Enrichment is computed over a changed reference set that includes also all the nodes connecting seeds with the same annotation. This may change the p-value.

NET-GE web server

A web server, implementing both a standard and a network-based gene enrichment was implemented as described in [12]. Briefly, NET-GE Web interface takes as input a list of genes/proteins (allowed identifiers are: UniProtKB AC, Ensembl and HGNC gene names). The enrichment can be performed considering the annotation modules based on STRING or STRING0.9. The enriched terms can derive from the GENE ONTOLOGY (all the sub-ontologies), or from the KEGG or the REACTOME PATHWAYS. The user can select between two kinds of multiple testing correction methods (Bonferroni or the Benjamini-Hochberg correction), and the significance threshold. As output NET-GE reports: 1) two enrichment tables (one for the standard enrichment and one for network-based one), 2) a graph visualizing how the enriched terms are linked, and 3) the complete set of annotations (for both the enrichment modes). Terms not included in the annotations of the input proteins are highlighted with a double star.

RESULTS

To test the performance of NET-GE we used sets of proteins involved in Mendelian diseases [11]. We tested 244 different genetic disorders, each one associated to two or more proteins. Our method was able to detect functional associations not detectable by the standard enrichment. Moreover, the newly enriched terms that were absent in the original annotations of the input genes are likely to provide new knowledge on the phenotype under examination [11].

Here, we present two cases of study demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder and the Obsessive-compulsive disorder.

Attention Deficit Hyperactivity Disorder

In the following, we deal with a specific test set (http://net-ge.biocomp.unibo.it/enrich/tutorial) that includes two input proteins related to Attention Deficit Hyperactivity Disorder (ADHD; OMIM #143465), a neurodevelopmental disease of childhood affecting the cognitive and behavioral functions. The genetic disease is associated to variations in the dopamine receptors DRD4 (UniProtKB AC: P21917) and DRD5 (UniProtKB AC: P21918). Using as input the DRD4 and DRD5 genes, we carried out enrichment analyses by setting the significance threshold at 0.05 on the Bonferroni corrected p-values. Standard and network-based enrichments ran over the KEGG database. Terms enriched by NET-GE are shown in Figure 1. The standard enrichment on KEGG highlights neuroactive ligand-receptor interaction and dopaminergic synapse as the most significant pathways. The network-based procedure adds new terms, not associated to the input proteins, and involved in ADHD, considering the statistically significant subnetworks. The pathways sorted by significance are: circadian entrainment, morphine addiction, retrograde endocannabinoid signaling and glutamatergic synapse.

Interestingly enough, the enriched pathways had been previously described in literature as being diseases-related. Different experiments have described different pathways [14–17] and the network-based enrichment method retrieved them all from the inclusion of the connecting nodes in the annotation modules.

In Figure 1 the difference in annotation between the standard enrichment procedure and the network-based is shown. As explained in the Methods section, standard enrichment is computed over a totally disconnected reference set. The network-based procedure relies on the precomputed annotation modules and the reference set includes all the nodes that connect seeds with the same annotation. This may increase the p-value as in
Enriched Terms - Directed Acyclic Graph

<table>
<thead>
<tr>
<th>Significance</th>
<th>0.05</th>
<th>0.05 to 10⁻³</th>
<th>10⁻³ to 10⁻⁵</th>
<th>10⁻⁵ to 10⁻⁷</th>
<th>10⁻⁷ to 10⁻⁹</th>
<th>&lt; 10⁻⁹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Content</td>
<td>≤ 5</td>
<td>5 ≤ IC ≤ 10</td>
<td>&gt; 10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard enrichment

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>Term</th>
<th>N1</th>
<th>N2</th>
<th>corrected p-value (Bonferroni)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>hsa04728</td>
<td>151 Show genes</td>
<td>135 Show genes</td>
<td>Show protein info</td>
<td>3.722E-03</td>
</tr>
<tr>
<td>S</td>
<td>hsa04080</td>
<td>2</td>
<td>2</td>
<td>Show protein info</td>
<td>1.736E-02</td>
</tr>
</tbody>
</table>

Network-based enrichment

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>Term</th>
<th>N1</th>
<th>N2</th>
<th>corrected p-value (Bonferroni)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N**</td>
<td>hsa04713</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>1.230E-02</td>
</tr>
<tr>
<td>N**</td>
<td>hsa05532</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>1.362E-02</td>
</tr>
<tr>
<td>N**</td>
<td>hsa04723</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>1.616E-02</td>
</tr>
<tr>
<td>N**</td>
<td>hsa04724</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>1.908E-02</td>
</tr>
<tr>
<td>N**</td>
<td>hsa04728</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>2.454E-02</td>
</tr>
<tr>
<td>N</td>
<td>hsa04728</td>
<td>2</td>
<td>1</td>
<td>Show genes</td>
<td>2.928E-02</td>
</tr>
</tbody>
</table>

Legend:
- New terms added by NET-GE and not directly associated to the input proteins/genes

Figure 1: Output of NET-GE for the enrichment of KEGG pathways in the ADHD study case. Enrichment analysis was carried out using as input the DRD4 and DRD5 genes. The upper panel shows the graph of the enriched terms and their relations. Box filling color represents the corrected p-value associated to the enriched term, while contour color represents its information content (see [11] and [12] for details). The lower panel presents the enriched terms in a tabular format. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.
the case of the neuroactive ligand receptor interaction that is no longer listed among the terms obtained with the network-based enrichment procedure.

For comparison, we also tried PINA and EnrichNET. Considering as significant p-values<0.05 Benjamini-Hochberg corrected, PINA (tool “Identify enriched Interactome modules”) did not retrieve any significantly over-represented module. EnrichNET authors recommend to analyse sets with at least 10 genes/proteins for reasons of statistical reliability. As a consequence, EnrichNET did not retrieve any significant term.

As evaluate the robustness of the method for small input sets composed of two to ten proteins, we computed the effect on the final stability of the enrichment when doubling (with random additions) the sizes of the input sets. We obtain that under these extreme conditions of noise, the stability of the enrichment ranges from 37 to 52%, depending on the annotation term and the network type (see Figure S1).

In Figure 2, the two input proteins are shown in the graph (purple circles) of the first protein neighbors, after network-based enrichment, detailing protein seeds of the Retrograde endocannabinoid signaling KEGG path (hsa04723, Homo sapiens) in yellow and the connecting nodes in blue (proteins that are retained after NET-GE based enrichment). The whole annotation network is downloadable (all seeds, nodes and arcs) and it is available for display.

Obsessive-compulsive Disorder

Obsessive-compulsive disorder (OCD) is a severe neuropsychiatric disorder characterized by the presence of obsessions and compulsions [18]. This disorder has been recently investigated in [18] by using whole-exome sequencing (WES).

Twenty OCD cases and their unaffected parents (parent-child trios) were screened for de novo missense mutations (i.e. mutation present only in the affected individual), identifying 27 OCD-related genes. Based on Ingenuity software (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), three signaling pathways were identified as disease-related [18] sharing only one patient/one gene. In fact, among the 27 genes, only SMAD4 (gene mutated in only one patient) was present in the three enriched pathways.

With NET-GE we highlight four biological processes as the most significant ones (Figure 3, panel A), all related to the purine metabolism that has been proven to be associated with several neurological disorders [19–21]. However, and interestingly enough, 10 of the 27 initial genes have common annotations. Testing Molecular Functions, the standard enrichment procedure highlighted ATPase activity and the network-based procedure enriched thyroxine 5’-deiodinase activity (Figure 3, panel B), a new term not associated to the input proteins and involved in OCD [22].

Our results highlight the involvement of processes common to the gene panel and corroborates the notion that network-based enrichment consistently derives information from the connected annotation modules, including genes corresponding to 9 of the 14 patients analyzed in [18].

CONCLUSION

In this article, we presented the NET-GE web server [12], developed for tackling the problem of the human biological complexity. Specifically, NET-GE is a tool for associating biological processes and pathways with sets of human genes/proteins involved in the same phenotype. It performs standard and network-based enrichment analysis. The network-based procedure extracts from the STRING human interactome sub-networks of connecting proteins that share the same annotation [11]. We benchmarked NET-GE on two specific test cases, with a phenotype and its biological functions already described in literature. On this benchmark, the network-based procedure, considering genes/proteins in the context of their functional interaction network, enriched functional annotations that are experimentally validated. This version of NET-GE is preliminary to the inclusion of some additional features that can eventually add to the relevance of detecting emerging functional characteristiscs from a set of genes, such as the inclusion of ranking scores (e.g. fold of differentially expressed genes) or the usage of tissue-specific interactomes.

ACKNOWLEDGEMENTS

RC thanks COST Action BM1405 (European Union RTD Framework Program) and FARB UNIBO.
### A) Biological Process

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>TERM</th>
<th>N1</th>
<th>N2</th>
<th>corrected p-value (Bonferroni)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>GO:0009203</td>
<td>10 [+] Show genes</td>
<td>1506 Show genes Show protein info</td>
<td>1.505e-02</td>
<td>ribonucleoside triphosphate catabolic process</td>
</tr>
<tr>
<td>N</td>
<td>GO:0009207</td>
<td>10 [+] Show genes</td>
<td>1506 Show genes Show protein info</td>
<td>1.505e-02</td>
<td>purine ribonucleoside triphosphate catabolic process</td>
</tr>
<tr>
<td>N</td>
<td>GO:0009146</td>
<td>10 [+] Show genes</td>
<td>1510 Show genes Show protein info</td>
<td>1.541e-02</td>
<td>purine nucleoside triphosphate catabolic process</td>
</tr>
<tr>
<td>N</td>
<td>GO:0009143</td>
<td>10 [+] Show genes</td>
<td>1518 Show genes Show protein info</td>
<td>1.616e-02</td>
<td>nucleoside triphosphate catabolic process</td>
</tr>
</tbody>
</table>

### B) Molecular Function

<table>
<thead>
<tr>
<th>Enrichment</th>
<th>TERM</th>
<th>N1</th>
<th>N2</th>
<th>corrected p-value (Bonferroni)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N**</td>
<td>GO:004809</td>
<td>2 [+] Show genes</td>
<td>8 Show genes Show protein info</td>
<td>8.505e-03</td>
<td>thyroxine 5'-deiodinase activity</td>
</tr>
<tr>
<td>N</td>
<td>GO:0017111</td>
<td>9 [+] Show genes</td>
<td>1786 Show genes Show protein info</td>
<td>1.146e-02</td>
<td>nucleoside-triphosphatase activity</td>
</tr>
<tr>
<td>N</td>
<td>GO:0016460</td>
<td>9 [+] Show genes</td>
<td>1919 Show genes Show protein info</td>
<td>2.033e-02</td>
<td>pyrophosphatase activity</td>
</tr>
<tr>
<td>N</td>
<td>GO:0016819</td>
<td>9 [+] Show genes</td>
<td>1924 Show genes Show protein info</td>
<td>2.075e-02</td>
<td>hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides</td>
</tr>
<tr>
<td>N</td>
<td>GO:0016817</td>
<td>9 [+] Show genes</td>
<td>1931 Show genes Show protein info</td>
<td>2.136e-02</td>
<td>hydrolase activity, acting on acid anhydrides</td>
</tr>
</tbody>
</table>

Figure 3: Output of NET-GE for Obsessive Compulsive Disorder for Biological Processes (panel A) and Molecular Function (panel B). Genes are derived from [18]. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.
CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files, together with supplementary items listed below, are available at Genomics and Computational Biology online.

Supplementary Figure S1. Testing the robustness of the network-based enrichment methods. For small input sets comprising from two to ten proteins (derived from OMIM), we computed the effect of doubling (with random additions) the size on the final stability of enrichment. This was done for all the annotation terms and the two different version of STRING (see Methods). Errors bars indicated standard deviations over a reference of 123 gene sets.

REFERENCES