Disease gene prioritization

Anil Jegga
Biomedical Informatics
http://anil.cchmc.org

GLBIO-2014
What, Why, & How

Computational Disease Gene prioritization

• **What**: Computationally assigning likelihood of gene involvement in generating a disease phenotype

• **Why**: Narrows down the set of genes to be tested experimentally – saves time/resources.

• **How**: “Guilt by Association” - Gene “priority" in disease is assigned in a more “informed” way taking into account a set of relevant features or annotations (e.g., gene expression, function/processes, pathways, model organism phenotype, etc.) - Functional Similarity-based methods
Computational Disease Gene Prioritization

Similarity-based Approaches (functional annotations-based)
- Training set-independent
- Training set-dependent

Network/Topology-based Approaches
- Training set-independent
- Training set-dependent

- Protein-Protein Interactions
- Protein Associations (Functional Linkage)
Guilt by association - Reliable predictions about the disease involvement of a gene can be made if several of its partners (e.g., genes with correlated expression profiles or protein interactants or genes involved in same biological process or pathway) share a corresponding annotation.

Incorporating the prior information or knowledge about a disease (e.g., known disease genes) is critical.

Challenge: Gather, normalize, and integrate heterogeneous data from multiple sources (and keeping them current).
Functional annotation-based candidate disease gene prioritization – General workflow

- **Step 1**: List of candidate genes (Test Set) to prioritize - linkage regions, chromosomal aberrations, association study loci, differentially expressed gene lists or genes identified by sequencing variants, or the complete genome.

- **Step 2**: Seed Genes or Training Set: Prior knowledge about the disease - known disease genes, or disease-relevant keywords, or biological processes or pathways.

- **Step 3**: Prioritization methods: Which one to select/use?

- **Step 4**: Assessment - Are the selected training/seed genes, keywords and tools suitable? Can reliable predictions be made using these?

- **Step 5**: Use multiple tools or multiple sets of seed gene or keywords - Combine the results to obtain a consensus result.
What constitutes a “good” seed gene set?

- **Relevancy**: Review each gene - Domain experts especially for selecting keywords (e.g., disease-relevant phenotypes)

- **Size Matters**: Neither too small nor too large.
  - Too small - may be insufficiently informative
  - Too large - too heterogeneous pattern to be useful.
    - Break them down into multiple random sets
    - Filter them based on additional features (e.g., genes associated with a BP term + MP term)
  - Ideally 6 – 30
What constitutes a “good” seed gene set?

- **Robustness**: How robust are the ranking results using a particular seed set?

**Cross-validation** - Assess whether a set of seed genes provides a coherent pattern

- Create **multiple sets** of seed genes or keywords covering complementary phenotypic aspects of the disease and assess their performance separately.
- **Negative control seed genes**: Use genes for other unrelated diseases as training set.
  - Top-ranking candidates are same with negative control seed genes – suggests some systematic bias and prioritization results are probably unreliable.
Other Quality Control Measures

Tool A ranks my “favorite” gene on/among top – Therefore tool A is the BEST!!!

• Smaller Test Sets: Perform prioritizations both on the actual set of candidates and on the whole genome OR on a larger set that includes the smaller set of candidates.
  - Are the top-ranking candidates from the small subset rank within the top 5–15% of the whole genome?
  - If not, the prioritization might not have been able to capture enough information to identify good candidates

• Functional Coherence: What are the enriched terms for the top ranked candidates? Do they match expectations for the biological process or phenotype of interest?

Moreau & Tranchevent, 2012
Resources commonly used for compiling seed set

OMIM: http://omim.org

Phenopedia: http://hugenavigator.net

GAD: http://geneticassociationdb.nih.gov

KEGG Disease: http://www.genome.jp/kegg/disease/
Additional resources for compiling seed set
Comparative Toxicogenomics Database: http://ctdbase.org
Comparative Toxicogenomics Database: [http://ctdbase.org](http://ctdbase.org)

### Batch Query

1. **Select your input type**
   - Chemicals (MeSH® names, synonyms, or IDs, or CAS RNs)
   - Diseases (MeSH or OMIM names, synonyms, or IDs)
   - Genes (NCBI symbols or IDs)
   - Gene Ontology terms (GO names, synonyms, or IDs)
   - Pathways (KEGG or REACTOME names or IDs)
   - References (PubMed® IDs or DOIs)

2. **Provide query terms (up to 4,000)**
   - Or upload a tab-separated file:
     - Browse: No file selected.
   - Identifiers column: 1

3. **Choose data to download**
   - **Data**
     - Chemical–gene interactions
     - Chemical associations
     - Gene associations
     - Disease associations
     - Pathway associations
     - Gene Ontology associations
   - Format:
     - TSV (tab-separated values)
     - CSV (comma-separated values)
     - JSON
     - XML

4. **Choose data to download**
   - **Data**
     - Chemical–gene interactions
   - **Chemical associations**
     - All
     - Curated
     - Inferred
     - Format:
       - TSV (tab-separated values)
     - CSV (comma-separated values)
     - JSON
     - XML
   - **Gene associations**
     - All
     - Curated
     - Inferred
   - **Disease associations**
     - Inferred
   - **Pathway associations**
     - Inferred
   - **Gene Ontology associations**

[Download] [Clear]
Comparative Toxicogenomics Database: http://ctdbase.org

Ontological tree – Children nodes and their annotations also used
Comparative Toxicogenomics Database: http://ctdbase.org

Explore the Venn utilities – Handy for generating/comparing annotated gene lists (seed set selection)

NCBI's portal to information related to Medical Genetics. Terms from the NIH Genetic Testing Registry (GTR), UMLS, HPO, ClinVar and other sources are aggregated into concepts and their gene annotations where available.
A group of genes that have a pathogenicity or other phenotype associated with them.
Functional Similarity – What features to consider?
• No single source of data can be expected to capture all relevant relations
• Integrate multiple data sources: Better signal-to-noise ratio and improved prediction accuracy
• **Guilt-by-association:** Approaches differ by the strategy adopted in calculating similarity and by the data sources utilized.

• With some exceptions (e.g., ENDEAVOUR, ToppGene), most of the existing approaches mainly focus on the combination of only a few data sources.

• For methodological details & validation see:
  - Aerts et al., 2006 Nature Biotech.
  - Chen et al., 2007 BMC Bioinfo.
ENDEAVOUR

Perform gene prioritization using our Web server

Perform gene prioritization using our Java server

Prioritize your candidates in 4 steps with the following wizard


Species
- Homo sapiens
- Rattus norvegicus
- Mus musculus
- Drosophila melanogaster
- Caenorhabditis elegans

Supports multiple species

Little picky on the input types
• Little picky on the input types - e.g., gene symbols have to be HGNC approved.
• Supports chromosomal regions, e.g. chr:8p or chr:20p13 – will fetch all the genes in that region
• Doesn’t support chromosomal coordinates

What shared features make a test set gene rank at top or What features are shared between the training/seed set and test set gene are not explicit.
ToppGene Suite – http://toppgene.cchmc.org

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis
  Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.

- **ToppGene**: Candidate gene prioritization
  Prioritize or rank candidate genes based on functional similarity to training gene list

- **ToppNet**: Relative importance of candidate genes in networks
  Prioritize or rank candidate genes based on topological features in protein-protein interaction networks

- **ToppGenet**: Prioritization of neighboring genes in protein-protein interaction network
  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network

---

Doesn’t support other than human/mouse
- Supports synonyms
- Presents suggestions/alternatives for unrecognized entries
### Enrichment of Training/seed set:

- Helps in assessing the "quality" of the training set
- Can assist in selecting sub-sets of training set to perform prioritizations (e.g., large training set)
Why is a test set gene ranked at top or what features are shared between the training/seed set and ranked test set gene are presented both as a network and tabular format.

- Select the ranked genes
- Resulting training set, shared annotations, and the ranked gene(s) can be downloaded as an XGMML or GEXF file (Cytoscape/Gephi import)

Download the rankings table
**Gene in bold is ranked test set gene; rest are training/seed genes.**

**Shared annotations between training set & ranked test set gene.**

**Blue nodes: Seed genes**  
**Pink nodes: ranked candidates**  
**All other nodes – shared annotations**
Combining gene level information with genomic variant information – Few case studies

Exome sequencing and gene prioritization

Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis

Yari Eritkh, Simon Edvardsson, Emily Hodges, et al.

Genome Res. 2011 21:598-604. originally published online April 12, 2011

Moreau & Tranchevent, 2012
Some more examples of published studies that used Endeavour and ToppGene for candidate gene prioritization

Original Article

Whole exome sequencing identifies mutation of EDNRA involved in ACTH-independent macronodular adrenal hyperplasia

Jie Zhu, Liang Cui, Wei Wang, Xing-Yi Hang, A-Xiang Xu, Su-Xia Yang, Jing-Tao Dou, Yi-Ming Mu, Xu Zhang and Jiang-Feng Gao.

Circ Cardiovasc Genet, published online May 15, 2013.

Exome Sequencing and Systems Biology Converge to Identify Novel Mutations in the L-Type Calcium Channel, CACNAIC, Linked to Autosomal Dominant Long QT Syndrome

Nicole J. Boczek, Jabe M. ETG, David T. Tran, John R. Gindressey, Stuart Midda, Jared M. Evans, Timothy J. Kamp and Michael J. Ackerman

Circ Cardiovasc Genet, published online May 15, 2013.

Table 3. Gene Ranking Using Bioinformatic/Systems Biology Algorithms

<table>
<thead>
<tr>
<th>Gene</th>
<th>Endeavour</th>
<th>SUSPECTS</th>
<th>ToppGene</th>
<th>Total Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACNAIC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SLC26A1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MAFIA</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DMRTB1</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>DCAMML1</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>OLTSCR1</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>PPP1R1C2</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>MARCKS</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 51. Summary of comparison results from ToppGene and Endeavour gene prioritization analysis

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>ToppGene Results</th>
<th>Endoavour results</th>
<th>Combined score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNAJC5</td>
<td>1.5</td>
<td>0.8</td>
<td>0.13</td>
</tr>
<tr>
<td>CACNAIC</td>
<td>2.0</td>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>CDK19</td>
<td>1.3</td>
<td>2.0</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Variants in GATA4 are a rare cause of familial and sporadic congenital diaphragmatic hernia


Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis

Bruno A. Benitez, David Alvarado, Yefei Cai, Kevin Mayo, Sumitra Chakraverty, Joanne Norton, John C. Morris, Mark S. Sands, Alison Goate, Carlos Cachaca.

AOTX2 Duplication Is Implicated in Hemifacial Microsomia

Dina Zielinski, Barak Markus, Iona Shalek, Melissa Gmousik, Clement Chu, Marta Zako, Batu Srinivasan, Joed D. Hoffman, Dror Alfenbender, Yaniv Erlich.
Limitations & Points to Remember

- **Bias towards the training set**: Disease genes yet to be discovered will be consistent with what is already known about a disease and/or its genetic basis – assumption not always true.

- **Bias towards selecting better annotated genes**: “true” candidate can be missed if it lacks “sufficient” annotations.

- **Accuracy depends on the quality (and coverage)** of underlying original sources from which the annotations are retrieved.

- **Appropriate or “true representative” training set selection**: Using larger training sets (>100 genes) decreases the sensitivity and specificity of the prioritization compared to smaller training sets (6 to 30 genes).

- **Coding-gene-centric**: Complex traits result more often from noncoding regulatory variants than from coding sequence variants.
Gene Prioritization Portal

Gene prioritization

One of the major challenges in human genetics is to find the genetic cause underlying disorders in order to unravel the molecular basis of these diseases and eventually elaborate medical treatments. High-throughput technologies such as linkage analysis and association studies are usually able to associate a chromosomal region with a genetic condition. One can also use expression arrays and obtain a list of transcripts differentially expressed in a disease sample with respect to a reference sample. A common characteristic of these methods is usually the large number of genes returned. For instance, hundreds of differentially expressed genes are often reported. The working hypothesis is often that only one or a few candidates are really of primary interest. Identifying the most promising candidates among such large list of genes is a challenging and time-consuming task. The bioinformatics community has therefore introduced the concept of gene prioritization that uses computational methods and genomic data to identify the most promising candidates.

The gene prioritization portal

In the last decade, many different strategies have been developed, some of which have been implemented into web applications and eventually experimentally validated. This website represents an updated electronic review in this field. The aim is to help researchers who need to select the most promising genes from large gene lists by describing what are the current available options. We describe more than 20 strategies (see the tools section) and give more general information to support user decisions (see the statistics section).

References


Summary:
46 tools
12 data sources
113 publications
39 validations
Gene Prioritization Portal

### General statistics

This section contains different statistics about the gene prioritization portal. These statistics are automatically retrieved from our frequently updated database and thus reflect the current state of the art in gene prioritization. Currently, the portal contains:

- 46 gene prioritization tools
- 12 data sources
- 113 gene prioritization-related publications
- 27 gene prioritization validations

### Data sources statistics

This section contains details about the different data sources and their use by the gene prioritization tools. Further details about the data sources can be found below the table.

<table>
<thead>
<tr>
<th>Tools</th>
<th>Functional annotation</th>
<th>Expression</th>
<th>Text (ontology)</th>
<th>Text (functional)</th>
<th>Integration</th>
<th>SNP</th>
<th>Phenotype</th>
<th>Conservation/homology</th>
<th>Disease</th>
<th>Disease associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>aGeneApart</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioGraph</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomine</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bitola</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caesar</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candid</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGI</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIR</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DomainRBF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endeavour</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eResponseNet</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2D</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneDistiller</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneFriends</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneProspector</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneRank</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneRanker</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneSeeker</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneWanderer</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Inputs / outputs statistics

This section contains details about the different inputs needed by the gene prioritization tools and about the different outputs they produce. The following two tables correspond to Table 2 of the paper [Trenchev et al. (2019)](https://doi.org/10.1093/nar/gkz192). Further details about the inputs / outputs can be found below the tables. The first table describes the inputs of the gene prioritization softwares.

<table>
<thead>
<tr>
<th>Tools</th>
<th>Known genes</th>
<th>Keywords</th>
<th>Expression dataset</th>
<th>Region</th>
<th>DEC</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>aGeneApart</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioGraph</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomine</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bitola</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caesar</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candid</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGI</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIR</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DomainRBF</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endeavour</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eResponseNet</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2D</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneDistiller</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneFriends</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneProspector</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneRank</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneRanker</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneSeeker</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneWanderer</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Génie</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenTrepid</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Gene Prioritization Portal**

**Detailed view of a gene prioritization tool**

<table>
<thead>
<tr>
<th>Name</th>
<th>Pinta</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td><a href="http://www.ebi.ac.uk/pinta/">http://www.ebi.ac.uk/pinta/</a></td>
</tr>
</tbody>
</table>

**Description**

PINTA identifies the most promising candidates within a region when only sparse information about the phenotype is available by replacing this knowledge by experimental data on differential gene expression between affected and healthy individuals. Considering the problem from the perspective of a gene/protein network, we assess the relevance of a candidate gene by considering the level of differential expression in its network neighborhood under the assumption that strong candidates tend to be surrounded by differentially expressed neighbors.

**Software modes**

- Web based software: The software contains a web based interface and therefore can be used from a web browser.

**Datasources**

<table>
<thead>
<tr>
<th>Functional annotations</th>
<th>Keywords associated to the genes and describing their functions or the ones of the corresponding proteins. Often an organized vocabulary (ontology) is used. The most used annotation source is Gene Ontology.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>Data covering the expression level (usually at the transcript level) of many genes (usually genome-wide) over a wide range of tissues/conditions.</td>
</tr>
<tr>
<td>Text (co-citation)</td>
<td>Links between genes obtained by applying text mining techniques in order to find co-citations occurring in the scientific literature.</td>
</tr>
<tr>
<td>Interactions</td>
<td>Collection of gene-gene (or protein-protein) interactions coming from biological experiments. Two examples of such data sources are DIP and BioGrid.</td>
</tr>
</tbody>
</table>

**Inputs / outputs**

- Training data: expression dataset
- Candidate genes: region
- Candidate genes: DEG
- Candidate genes: genome
- Prioritized list of candidates
- Test Statistics

**Publications**

  - Show abstract, link to paper

**Benchmarks**

- Nitsch et al., Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches, BMC Bioinformatics (2010)
  - Show abstract, link to paper

**Tools used:** Pinta

**Critical Assessment**

Results of a critical assessment performed on 8 gene prioritization softwares (reference to come). Regular benchmarks are usually run on known data; oppositely this benchmark mimics the discovery of a novel disease gene by performing predictive prioritizations. More precisely, 42 novel disease gene associations were used to benchmark the tools as soon as they were published and therefore before their inclusions in the databases used by those tools.

This plot presents the results of this benchmark for Pinta. Several performance indices are computed: median over the 42 rank ratios, AUC (Area Under the ROC Curve), and True Positive Rates at 10% and 30%. 
Disease Gene Prioritization - Network-based strategies

- Candidate genes are ranked based on their topological relevance (e.g., distance) to known disease genes (Training/seed genes) in a network.
  - Protein-protein interactions network (BioGrid, BIND, HPRD, etc.)
  - Protein association network (STRING)
- Random-walk (or PageRank) approaches outperform clustering and neighborhood approaches.
ToppNet (http://toppgene.cchmc.org)

ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis
  
  Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.

- **ToppGene**: Candidate gene prioritization
  
  Prioritize or rank candidate genes based on functional similarity to training gene list

- **ToppNet**: Relative importance of candidate genes in networks
  
  Prioritize or rank candidate genes based on topological features in protein-protein interaction network

- **ToppNet**: Prioritization of neighboring genes in protein-protein interaction network
  
  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network based on functional similarity to the "seed" list (ToppGene) or topological features in protein-protein interaction network (ToppNet).

Graph prioritization parameters

- **Prioritization method**: Page Rank With Priors
- **Bias (between 0 and 1)**: 0.3
- **Training gene neighborhood subnetwork visualization parameters**
  
  Neighborhood distance: 1
  
  When training set is big, the training gene neighborhood subnetwork can be huge

Test Genes

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2023</td>
<td>EP300</td>
</tr>
<tr>
<td>2</td>
<td>5399</td>
<td>ID3</td>
</tr>
<tr>
<td>3</td>
<td>3398</td>
<td>ID2</td>
</tr>
<tr>
<td>4</td>
<td>3397</td>
<td>ID1</td>
</tr>
<tr>
<td>5</td>
<td>10014</td>
<td>HDAC5</td>
</tr>
<tr>
<td>6</td>
<td>8928</td>
<td>FOXH1</td>
</tr>
<tr>
<td>7</td>
<td>6965</td>
<td>TBP2</td>
</tr>
<tr>
<td>8</td>
<td>836</td>
<td>CASP3</td>
</tr>
<tr>
<td>9</td>
<td>7603</td>
<td>TEAD1</td>
</tr>
<tr>
<td>10</td>
<td>840</td>
<td>CASP7</td>
</tr>
<tr>
<td>11</td>
<td>5469</td>
<td>MBD1</td>
</tr>
<tr>
<td>12</td>
<td>9734</td>
<td>HDAC9</td>
</tr>
<tr>
<td>13</td>
<td>4776</td>
<td>NIFATC4</td>
</tr>
<tr>
<td>14</td>
<td>23414</td>
<td>ZFP52</td>
</tr>
<tr>
<td>15</td>
<td>23654</td>
<td>NQCA6</td>
</tr>
<tr>
<td>16</td>
<td>5551</td>
<td>PROX1</td>
</tr>
<tr>
<td>17</td>
<td>481</td>
<td>NR3C1</td>
</tr>
<tr>
<td>18</td>
<td>1489</td>
<td>CEBPB1</td>
</tr>
<tr>
<td>19</td>
<td>841</td>
<td>CASP8</td>
</tr>
<tr>
<td>20</td>
<td>93649</td>
<td>MYOD1</td>
</tr>
</tbody>
</table>
New tools – Variant prioritization

http://homes.esat.kuleuven.be/~bioiuser/eXtasy/

http://compbio.charite.de/ExomeWalker

EXOME WALKER is a computational method to prioritise a set of candidates in exome sequencing projects that aim to identify novel Mendelian disease genes. Our approach involves filtering a Variant Call Format (VCF) file according to a number of user-definable criteria, for instance, off-target variants (those that are not located within or close to protein-coding exons) are removed.

Genes are prioritised according to a variant score (predicted pathogenicity, rarity, pattern of variants compatible with the assumed mode of inheritance) and to their vicinity to other genes that belong to the same disease-gene family within the protein-protein interaction (PPI) network, using the Random-Walk method as described in Kuhler et al. (2008) to determine similarity within the PPI network on the basis of the global characteristics of the network.

1. Upload VCF Data

VCF files for single or multiple samples are supported. If you upload a multiple-sample VCF file, you will also be required to enter pedigree data in PED file format. The file suffix must be “VCF” or “vcf”.

Choose VCF file: Browse... No file selected.

Submit
References & Further Reading
