Gene/Protein Function Annotation

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Main topics

- What’s function
  - Gene ontology
  - Functional similarity
- Function annotation
  - Homology-based
  - Guilt-by-association
- Gene-disease association
- What can go wrong
  - Annotation mistakes
  - CAFA & evaluation of function prediction
Which is more difficult to predict?

- Function
- Functional residues
Hypothetical proteins

- New protein sequences come from genome (and metagenome) sequencing projects
- Many have no known functions
Why we need to do function annotation?

Fig from: Network-based prediction of protein function. Molecular Systems Biology 3:88. 2007
What’s function?

- The definition of biological function is ambiguous (context dependent)
  - FOXP2 is involved in human-specific transcriptional regulation of CNS development
  - the transcription factor FOXP2 (forkhead box P2) is the only gene implicated in Mendelian forms of human speech and language dysfunction
  - two human-specific amino acids alter FOXP2 function by conferring differential transcriptional regulation in vitro…
    - Nature 462, 213-217, 2009

- It is obvious that the biological function of a protein has more than one aspect
How to describe function?

- in a computationally amenable way?
- Human language
- **Controlled** vocabulary
  - EC (Enzyme Commission Classification)
    1. -. -.- Oxidoreductases.
    1. 1. -. Acting on the CH-OH group of donors.
    1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
    1.1.1.1 Alcohol dehydrogenase.
    1.1.1.3 Homoserine dehydrogenase.
  - GO (Gene Ontology)
    - [http://www.geneontology.org](http://www.geneontology.org)
The GO is actually three ontologies

**Molecular Function**
- GO term: Malate dehydrogenase.
- GO id: GO:0030060
- \((S)\)-malate + NAD(+) = oxaloacetate + NADH.

**Biological Process**
- GO term: tricarboxylic acid cycle
- Synonym: Krebs cycle
- Synonym: citric acid cycle
- GO id: GO:0006099

**Cellular Component**
- GO term: mitochondrion
- GO id: GO:0005739
- Definition: A semiautonomous, self-replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

Adapted from: http://www.geneontology.org/GO.teaching.resources.shtml
Ontology

- In computer science and information science, an **ontology** is a formal **representation of knowledge as a set of concepts** within a domain, and the **relationships** between those concepts.

- Gene ontology: GO terms (e.g., Malate dehydrogenase), and relationships between the GO terms (is_a, part_of)
Each GO term has 2 Definitions

**Gene Ontology Browser**

<table>
<thead>
<tr>
<th>GO term:</th>
<th>cell differentiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO id:</td>
<td>GO:0030154</td>
</tr>
<tr>
<td>Definition:</td>
<td>The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.</td>
</tr>
</tbody>
</table>

A definition written by a biologist: **necessary & sufficient conditions**

**written definition**
(not computable)

Graph structure: **necessary conditions**

**formal**

(computable)

Adapted from: http://www.geneontology.org/GO.teaching.resources.shtml
Terms are defined graphically relative to other terms.
Appropriate relationships to parents

- **GO currently has 2 relationship types**
  - **Is_a**
    - An is_a child of a parent means that the child is a complete type of its parent, but can be discriminated in some way from other children of the parent.
  - **Part_of**
    - A part_of child of a parent means that the child is always a constituent of the parent that in combination with other constituents of the parent make up the parent.

```
Part_of  \  Is_a relationships
   |    /
   v    

nucleus  chromosome  mitochondrion
```

- **Part_of** relationship
- **Is_a** relationships
Distance between two terms (functions)?

- Why we care
  - We can compare proteins/genes based on their biological role
  - Evaluate if a clustering of genes/genes (based on gene expression level, etc) makes sense at all.

- Different ways of computing the distance
  - Shortest path between two terms
  - Semantic similarity
Semantic similarity

- A definition: a semantic similarity measure is defined as a function that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in meaning between them.

Main approaches for comparing terms: node-based and edge-based and the techniques used by each approach

DCA, disjoint common ancestors;
IC, information content;
MICA, most informative common ancestor
Semantic similarity based on information content

Here the probability of each node is the probability of this term occurring in a database such as SWISS-Prot. Semantic similarity defined as the information content (IC) of shared parents of two terms (-\ln p)

Building the ontologies

- The GO is still developing daily both in ontological structures and in domain knowledge

[Diagram showing relationships between terms like immune system process, immune response, antigen processing and presentation, etc.]

Red part_of
Blue is_a

Adapted from: http://www.geneontology.org/GO.teaching.resources.shtml
Ontology:

PROTEIN FUNCTION

DNA binding
RNA binding
Lyase activity

Functional annotation: consistent subgraph of the ontology

Slide resource: Predrag Radivojac, ACM BCB 2013
### GO annotations

<table>
<thead>
<tr>
<th>Species/datasets</th>
<th>Gene products annotated</th>
<th>Annotations</th>
<th>Submission dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bos taurus GO Annotations @ EBI</td>
<td>23800</td>
<td>106735 (4138 non-IEA)</td>
<td>11/7/2009</td>
</tr>
<tr>
<td>Caenorhabditis elegans WormBase</td>
<td>18617</td>
<td>103445 (47582 non-IEA)</td>
<td>10/6/2009</td>
</tr>
<tr>
<td>Drosophila melanogaster FlyBase</td>
<td>12484</td>
<td>71813 (56890 non-IEA)</td>
<td>11/7/2009</td>
</tr>
<tr>
<td>Gallus gallus GO Annotations @ EBI</td>
<td>16306</td>
<td>70674 (2035 non-IEA)</td>
<td>11/7/2009</td>
</tr>
<tr>
<td>Homo sapiens GO Annotations @ EBI</td>
<td>18587</td>
<td>165741 (69048 non-IEA)</td>
<td>11/7/2009</td>
</tr>
</tbody>
</table>

GO evidence code

-- Experimental Evidence Codes
EXP: Inferred from Experiment
IDA: Inferred from Direct Assay
IPI: Inferred from Physical Interaction
IMP: Inferred from Mutant Phenotype
IGI: Inferred from Genetic Interaction
IEP: Inferred from Expression Pattern

-- Computational Analysis Evidence Codes
ISS: Inferred from Sequence or Structural Similarity
ISO: Inferred from Sequence Orthology
ISA: Inferred from Sequence Alignment
ISM: Inferred from Sequence Model
IGC: Inferred from Genomic Context
RCA: inferred from Reviewed Computational Analysis

-- Author Statement Evidence Codes
TAS: Traceable Author Statement
NAS: Non-traceable Author Statement

-- Curator Statement Evidence Codes
IC: Inferred by Curator
ND: No biological Data available

-- Automatically-assigned Evidence Codes
IEA: Inferred from Electronic Annotation
Mappings to GO

- UniProt2GO
- Pfam2GO
- MetaCyt2GO
- EC2GO
- COG2GO (outdated; last updated June 2004)
Annotating gene products using GO

Adapted from: http://www.geneontology.org/GO.teaching.resources.shtml
Gene ontology tools

- **Annotation tools**
  - Blast2GO
  - GOanna
  - GOtcha
  - ...

- **Tools for gene expression/microarray analysis**
  - BiNGO
  - ...

What information can be used for function annotation?

- **Sequence based approaches**
  - Protein A has function X, and protein B is a homolog (ortholog) of protein A; Hence B has function X

- **Structure-based approaches**
  - Protein A has structure X, and X has so-so structural features; Hence A’s function sites are …. 

- **Motif-based approaches (sequence motifs, 3D motifs)**
  - A group of genes have function X and they all have motif Y; protein A has motif Y; Hence protein A’s function might be related to X

- **“Guilt-by-association”**
  - Gene A has function X and gene B is often “associated” with gene A, B might have function related to X
  - Associations
    - Domain fusion, phylogenetic profiling, PPI, etc.

- **Meta-approaches**
Homology-based function prediction

Image from http://genomebiology.com/2009/10/2/207
Different ways of “transferring” functions
Protein function annotation as a classification problem

- **Protein classifications**
  - Domain based
    - Sequence only (Pfam)
    - Structure based (SCOP, CATH)

- **How many protein families?**
  - Superfamily, family & subfamily
Annotation transfer by homology

- Database searching using sequence-based alignment approaches
  - BLAST
  - PSI-BLAST, profile-profile alignment
  - Hmmpfam against Pfam database
- Significance evaluation in database searching
- Ortholog / paralog
  - Phylogeny analysis
  - Ortholog -- same function
  - Paralog -- different function
Database for searching

- Protein family databases
  - Pfam
  - PANTHER: A Library of Protein Families and Subfamilies Indexed by Function (http://www.pantherdb.org/)
  - SEED gene family
  - KEEG gene family
  - etc
Similar vs orthologous

A1⇒B1  √
B2⇒B1  ×

A1, B1: Orthologs
A1, B2: Paralogs

Query sequence
Structure-based function prediction

- Structure-based methods could possibly detect remote homologues that are not detectable by sequence-based method
  - using structural information in addition to sequence information
  - protein threading (sequence-structure alignment) is a popular method

*Structure-based methods could provide more than just “homology” information*
Using sequence-structure alignment method, one can predict a protein belongs to a SCOP family / superfamily / fold

family (same function)
superfamily (similar functions)
fold (different functions)
Structural Genomics: structure-based functional predictions

*Methanococcus jannaschii* MJ0577 (Hypothetical Protein)

Contains bound ATP => ATPase or ATP-Mediated Molecular Switch

Confirmed by biochemical experiments

Modified from: http://pir.georgetown.edu/pirwww/about/presentations/nihworkshop2007/NIH-mar2307.an.ppt
Motif-based function prediction

- **Sequence motif (pattern)**
  - PROSITE (ScanPROSITE)
  - BLOCK
    - Multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins
  - PRINTS
    - Collection of protein fingerprints -- a fingerprint is a group of conserved motifs used to characterize a protein family; more powerful than can single motifs

- **Motif finding -- a well-defined bioinformatics problem**
  - Alignment based / alignment independent
  - MEME
PROSITE & ScanPROSITE

- **PROSITE** contains patterns specific for more than a thousand protein families.

- **PROSITE Examples**
  - **PKC_PHOSPHO_SITE**, PS00005; Protein kinase C phosphorylation site
    - Consensus pattern: [ST] - x - [RK]
    - S or T is the phosphorylation site
  - **URICASE**, PS00366; Uricase signature (PATTERN)
    - [LV] - x - [LV] - [LIV] - K - [STV] - [ST] - x - [SN] - x - F - x(2) - [FY] - x(4) - [FY] - x(2) - L - x(5) - R

- **ScanPROSITE** -- it allows to scan a protein sequence for occurrence of patterns and profiles stored in **PROSITE**
Function prediction based on local structure patterns

- 3D motif (spatial patterns of residues)
- Clefts / pockets (Prediction of ligand binding sites)
  - For ~85% of ligand-binding proteins, the largest cleft is the ligand-binding site
  - For additional ~10% of ligand-binding proteins, the second largest cleft is the ligand-binding site
A typical example of 3D motif: catalytic triad

- A catalytic triad: 3 amino acid residues found inside the active site of certain protease enzymes: serine (S), aspartate (D) and histidine (H). They work together to break peptide bonds on polypeptides.

- The residues of a catalytic triad can be far from each other in the primary structure, but are brought close together in the tertiary structure.
Local structure pattern resources

- PINTS -- Patterns In Non-homologous Tertiary Structures (3D motif)
  - http://www.russell.embl.de/pints/

- eF-site -- electrostatic-surface of Functional site
  - a database for molecular surfaces of proteins' functional sites, displaying the electrostatic potentials and hydrophobic properties together on the Connolly surfaces of the active sites
  - http://ef-site.protein.osaka-u.ac.jp/eF-site/

- Catalytic site atlas
  - http://www.ebi.ac.uk/thornton-srv/databases/CSA/
Guilty-by-association

- Phylogenetic profiling (co-evolution pattern)
- Protein-protein interaction
- Domain fusion
- Genomic context
  - Neighbor genes (operon) / Gene team
- Gene expression (protein expression level) etc
- Integration
Phylogenetic profiling approach

- A non-homologous approach using co-evolution pattern
- The **phylogenetic profile** of a protein is a string that encodes the presence (1) or absence (0) of the protein in every sequenced genome (0/1 string)
- Proteins that participate in a common structural complex or metabolic pathway are likely to co-evolve, the phylogenetic profiles of such proteins are often “similar”
- Similarity of phylogenetic profiles -- similarity of functionality
Phylogenetic profiling approach

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Genomes</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EC</td>
<td>HI</td>
<td>BS</td>
</tr>
<tr>
<td>P1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P4</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Genes with similar phylogenetic profiles have related functions or functionally linked – Eisenberg and colleagues (1999)
Sequence co-evolution

Protein Family A

Protein Family B

Similarity
Gene (domain) fusion for PPI prediction

- **Gene (domain) fusion** is an effective method for predicting protein-protein interactions.
  - If proteins A and B are homologous to two domains of protein C, A and B are predicted to interact with each other.

- **Rosetta stone methods**

Gene-fusion has low prediction coverage, but it has low false-positive rate.
Genomic-context based approaches

Gene cluster

Gene neighborhood

Genome 1

Genome 2

Genome 3
Functional inference at systems level

- Function prediction of individual genes could be made in the context of biological pathways/networks.
- By doing homologous search, one can map a known biological pathway in one organism to another one; hence predict gene functions in the context of biological pathways/networks.
- Example – phoB is predicted to be a transcription regulator and it regulates all the genes in the pho-regulon (a group of co-regulated operons); and within this regulon, gene A is interacting with gene B, etc.
Gene-disease association

- **Phenotypic function:**
  - A higher-level functional organization than protein cellular function
  - Relationship to multi-scale modeling
- **Genes whose disruption is causative?**
  - Mutations and indels in protein-coding or regulatory regions
- **Genes shown to be “associated” in experimental/genetic studies?**
  - Linkage analysis, association studies
  - Copy number variations
  - Chromosomal rearrangement
  - Protein biomarkers
- **Genes whose function is disrupted downstream?**
  - Protein biomarkers
  - Still good for drug targets
  - But, how much downstream?

_Predrag Radivojac’s, “Protein function prediction: formulation, methodology, evaluation, and challenges” tutorial at ACM-BCB2013_
A network of disorders and disease genes linked by known disorder–gene associations
The story of disease genes


## Why different from other genes

Differences between disease-associated and non-disease associated genes

- disease-associated genes are on average longer than non-disease associated genes
- disease-associated genes are more likely to have homologs in distant species, but less likely to have close paralogs than non-disease associated genes
- disease-associated genes have more interacting partners on average than non-disease associated genes, but fewer than essential genes (approximated by using housekeeping genes)
- certain biological functions are overrepresented or underrepresented in disease-associated genes
- disease-associated genes have more exons and greater total exon length than non-disease genes, but they have similar total intron length and 3’ and 5’ UTR region length
- disease-associated genes do not have distinct $K_a/K_s$ values as compared to the remaining genes (except slightly for essential genes)
- disease-associated genes are less conserved than the essential genes, but similarly conserved as the remaining non-disease genes
- genes associated with the same diseases tend to interact more frequently, are co-expressed in the same tissue and tend to share GO terms
- disease-associated genes that correspond to certain GO terms have similar modes of inheritance (dominant vs. recessive)
- disease-associated genes generally have higher expression levels than non-disease genes but are expressed in a narrower range of tissues

- **evolutionarily**
  - purifying selection on genes for fetal survival or reproductive capacity

- **functionally**
  - close paralogs could take over functions

- **statistically**
  - longer proteins likelier to acquire deformities

- **Open questions**
  - Is there a collection bias?
  - Why aren’t all (non-essential) genes involved in some disease?

Predicting disease genes

• **Statistical and machine learning methods (gene prioritization)**
  • Based on the same principles used for prediction of molecular function
  • Task: given disease-associated genes and various data, find other disease-associated genes
  • Data: sequences, structures, PPI networks, microarray data, evolutionary data…

• **Methods that use patient samples (e.g. two cohorts)**
  • Microarray-based methods
  • Mass spectrometry-based methods

• **Methods that attempt to understand molecular cause of disease**
  • Studies of SNPs, given sequence, structure, and evolutionary information
  • Simulations of protein folding and dynamics

*Predrag Radivojac’s, “Protein function prediction: formulation, methodology, evaluation, and challenges” tutorial at ACM-BCB2013; Dalkilic et al. Front Biosci. 13: 3391 (2008).*
Integration of multiple data sources for function annotation

SAMBA framework

Fig from: Network-based prediction of protein function. Molecular Systems Biology 3:88. 2007
Be aware of the easy mistakes one can make

New sequence

Chorismate mutase

BLAST

Chorismate mutase domain

ACT domain
Should we go with whole proteins, domains, or motifs?

PIRSF006256

- **Acylphosphatase** - ZnF - ZnF - **YrdC** - Peptidase M22

On the basis of domain composition alone, biological function was predicted to be:
- RNA-binding translation factor
- maturation protease

Actual function:
- [NiFe]-hydrogenase maturation factor, carbamoyltransferase

Whole Protein != Sum of its Parts?

Modified from: http://pir.georgetown.edu/pirwww/about/presentations/nihworkshop2007/NIH-mar2307.an.ppt
Be aware of the propagation of mistakes

Functional annotation could be very messy

A protein (ZP_06741787.1) from *Bacteroides vulgatus* is annotated as **integron integrase**; similarity search shows that it shares 98% sequence identity with protein ZP_07940359.1 from *Bacteroides sp. 4_1_36*, which is annotated as a **phage integrase**, and shares 87% identity with protein, ZP_05415972.1, annotated as a **tyrosine type site-specific recombinase** from *Bacteroides finegoldii*...
CAFA: Critical Assessment of Function Annotation

Evaluation metric: $F_{\text{max}}$, a harmonic mean between precision and recall

It is a meta-method, combing predictions from

- GO term prediction using PSI-BLAST
  - PSI-BLAST search against the UniRef90 [21] data-base, using 3 iterations, and an E-value threshold of 1x10-3 for both hit selection and profile inclusion

- GO term prediction from UniProtKB/Swiss-Prot text-mining (a naïve Bayes text-mining approach)
  - \( p(\text{GO} | \text{word}) \approx f(\text{word} | \text{GO}) / \{f(\text{word} | \text{GO}) + f(\text{word} | \sim \text{GO})\} \)

- GO term prediction from amino acid trigram mining
  - Similar to the previous one, but using trigrams of amino acids

- GO term prediction from sequence features (FFPRED)
  - FFPRED starts by predicting a diverse range of sequence features, which include secondary structure elements, disordered regions, signal peptides, glycosylation sites, and several others. These features are then analysed by a series of Support Vector Machines (SVMs) to assign GO terms from a subset of 197.

- GO term prediction from orthologous groups

- GO term prediction from profile-profile comparison

Cozzetto et al, BMC Bioinformatics, 2013;14 Suppl 3:S1. Protein function prediction by massive integration of evolutionary analyses and multiple data sources.
and

- GO term prediction from high-throughput data sources (FunctionSpace)
  - Human proteins are assigned coordinates in an 11-dimensional feature space and GO terms assigned from annotated close neighbours in this space using SVM regression. The 11 dimensions represent pairwise sequence similarity, predicted cellular localization, secondary structure similarity, transmembrane topology, disordered segment features, sequence-derived domain architecture, structure-based domain architecture, sequence domain fusion patterns, structural domain fusion patterns, protein-protein interactions and microarray data.

Integration

- Predictions from component methods are combined using a network propagation algorithm based on the GO graph structure.
Automated annotation of protein function is challenging

- today's best protein function prediction algorithms substantially outperform widely used first-generation methods, with large gains on all types of targets
- although the top methods perform well enough to guide experiments, there is considerable need for improvement of currently available tools.