Gene function annotation

Paul D. Thomas, Ph.D.
University of Southern California

What is function annotation?

• The formal answer to the question: what does this gene do?
• The association between: a description of biological function, in electronic form, with a biological sequence (gene or gene product e.g. protein or functional RNA)
In this lecture

• Introduction to ontologies of gene function

• Methods and online information sources for function annotation
  - Understand what you are getting from each source so you can use it wisely
  - The most useful databases combine sophisticated computational algorithms (coverage) with human curation/review/error correction (accuracy)

• Phylogenetic analysis of function
  - Importance of homology inference

Ontologies

• A formal structuring of knowledge
• Consists of concepts and relations
• Concept (entity, class, term): a class of things in the real world
  - Continuant (thing that exists)
  - Occurrent (process)
• Relation: a type of relationship between concepts
  - E.g. is_a, part_of
Entrez Gene: INSR

Protein function ontologies

- **Gene Ontology (GO)**

- **Pathway Ontologies**
  - Reactome
  - PANTHER
  - BioCyc
  - KEGG

Gene Ontology

- Formal representation of biology knowledge domain, as it relates to genes and gene products (mostly proteins)
- Three knowledge domains:
  - Molecular function: what a gene product does with its direct physical interaction partners, e.g. protein kinase
  - Cellular component: where the protein is located when the function is carried out, e.g. plasma membrane
  - Biological process: “system” function carried out by multiple molecular functions working together in a regulated manner, e.g. pathways, cellular processes, organ functions, organism behavior

Pathway ontologies

- Point of view from the molecular reaction
  - Generalized to include covalent and noncovalent (e.g. binding) reactions
- Concepts are reaction, molecule classes
- Relations are between molecule classes and reactions
  - Catalyst
  - Reactant
  - Product
- Top level structure provided by SBML, BioPAX
  - Systems modeling community vs. Genomics community
Notch signaling pathway in GO

Term lineage (superclasses)

Term children (subclasses)

Notch signaling in PANTHER
**GO vs. pathway ontologies**

- **GO is simpler, and has more biological context**
- **GO is a community-wide effort!**
- Pathway ontologies are more detailed, and can capture dependencies and temporal series
- Pathway ontologies are (currently) relatively isolated and distributed

**GO annotation evidence**

For most model organisms, about half the unique GO annotations are based only on homology

NOTE: evidence codes are available in gene annotation files from http://www.geneontology.org
Direct, literature-based annotation

- Function annotation **inference** based on direct evidence in the scientific literature
  - Experiment performed on that gene product itself
- Text mining and management (Textpresso)
  - Very active area of research
- Curator reads abstract or article and manually enters annotation
- GO annotation is performed at 12 different “model organism databases” and UniProt
- Two types:
  - Primary source: experimental paper (Evidence codes: IMP, IGI, IDA, IEP, IPI)

Homology-based annotation

- Until recently, no coordinated effort by GOC
  - (ISS, some RCA and most IEA)
- Pairwise view
  - If two sequences are similar, they are likely to share some functions in common
  - So if I know the function of one gene, I can make inferences about the function of another gene
    - “transitive annotation” (ISS evidence code in GO)
  - Very commonly applied, in database search algorithms like BLAST, FASTA
  - This success has led to overinterpretation of its meaning by many casual users
BLAST results for human MTHFR vs. SwissProt database

What is transitive annotation?

- Two sequences are similar **because** they are homologous (at least for relatively long, non-repetitive sequences, i.e. almost all genes)

- More properly, transitive annotation of function is inheritance!
  - “related genes have a common function **because** their common ancestor had that function, which was inherited by its descendants”
  - not just an inference about one gene. It is also making inferences about
    - The most recent common ancestor (MRCA)
    - Continuous inheritance since the MRCA
    - Potential inheritance by other descendants of the MRCA
Function annotation by homology: important background

- Protein family evolution
  - Orthology, paralogy
- Modular domain structure of many proteins
- Interpreting sequence divergence (neutral, selected)
- Sequence divergence and structure similarity

Function annotation by homology: computational background

- Computational algorithms
  - Pairwise alignment (e.g. BLAST)
    - Global alignment
    - Local alignment / domain
  - Multiple sequence alignment (e.g. MAFFT)
  - Protein sequence phylogenetic tree (e.g. PHYLIP, PhyML)
  - Hidden Markov Model (e.g. SAM, HMMer)
Protein families

- Arise from copying and divergence
  - A tree is a natural way to represent this (Darwin)
- A family derives from a single common ancestor, and members retain (“conserve”) sequence similarity due to functional constraint
- Proteins are modular: part or all of a protein may be copied and conserved, but a minimum functional unit must remain (a “domain”)

Representing evolution of related genes

- Start with Darwin’s basic model:
  - Copying
    - An ancestral “species” “splits” into two separate species
  - Divergence
    - Each copy (species) changes independently over generations
      - NATURAL SELECTION: adaptation to different environment
Darwin’s species tree

- Number of generations/time along one axis
- Amount of divergence along other axis
- Characters in common are due to inheritance
  - Also tells us something about common ancestor

Representing evolution of related genes

- “Gene families”
- Add detail from population genetics/molecular evolution to apply to genes
  - Copying
    - An ancestral species “splits” into two separate species
      - SPECIATION
    - A gene is duplicated in one population and subsequently inherited
      - DUPLICATION
  - Divergence
    - Each copy (gene sequence) changes independently over generations
      - NATURAL SELECTION: sequence substitutions to adapt to new function/role
      - NEUTRAL DRIFT: accumulation of “neutral” substitutions
A gene tree

- Only one "informative" axis: rate of sequence evolution
  - For neutral changes this can often act as a "molecular clock"
  - Non-neutral changes will speed up the rate of evolution

How does this relate to gene function?

- Copying
  - An ancestral species "splits" into two separate species
    - SPECIATION: likely to continue performing ancestral function ("orthology")
      - BUT not always
  - A gene is duplicated in one population and subsequently inherited
    - DUPLICATION: "redundant gene" free from previous constraints can adapt to a new function ("paralogy")
      - One "conserved" copy may retain ancestral function
      - BUT "redundant" copy still inherits some aspects of ancestral function

- Divergence
  - Each "new" (gene sequence) changes independently over generations
    - NATURAL SELECTION: sequence substitutions adapt to new/modified function/role
    - NEUTRAL DRIFT: sequence changes from accumulation of "neutral" substitutions. This is the MAJOR source of sequence differences!
Gene duplication and functional novelty

- **“Neofunctionalization” model**
  - One copy retains ancestral function
  - One copy adapts to new function
    - More diverged copy often recognizable as having larger branch length

- **“Subfunctionalization” model**
  - Ancestral gene has at least two functions/specificities
  - Each copy adapts to “specialize” in a subset of the ancestral functions

Homology inference in a tree

inheritance and divergence of function
Homology inference in a tree

inheritance and divergence of function

COMBINES:
1. Evolutionary information (tree)
2. Experimental knowledge (GO annotations from literature)
3. Organism-specific biological knowledge (curators)
Orthologs and paralogs

- The term "Orthologs" is often used to denote "the same gene" in different organisms but this is not technically correct, and can lead to confusion
- Orthologs share a MRCA immediately preceding a speciation event
  - i.e. they can be traced to a single gene in the most recent common ancestor population/species
- Paralogs share a MRCA immediately preceding a gene duplication event
  - i.e. they can be traced to a gene duplication event in the most recent common ancestor population/species, and can be traced to distinct ancestral genes in that species

Why do we care about orthologs?

- If there have been no gene duplication events, orthologs are likely to be performing the same function in two different organisms
- Even if there have been duplication events, because there is a common ancestor, the functions are likely to be similar
- Many "ortholog" finding methods (e.g. "bi-directional best hit") will report only one ortholog even if there are more
  - This is the "most similar" homolog and there are scenarios under which this is the one with the most similar function as well
"Ortholog clusters" OrthoMCL

Consider both:
- Orthologs (reciprocal best hits between species)
- In-paralogs (reciprocal better hits within species)

There are numerous other ortholog clustering algorithms. Among the most highly used are:
- InParanoid (Sonnhammer)
- Homologene (NCBI)
- Ensembl Orthologs

Li et al., Genome Res. 13:2178, 2003

Why orthology is confusing

- It is defined technically in relation to a PAIR of genes
- The term “ortholog clusters” is often used, but it is NOT TRANSITIVE
  - An ortholog cluster may contain pairs that are paralogs!
- Proposed solutions are complicated
  - One solution is to only cluster “one-to-one orthologs” where no gene duplication occurs, but this misses many functionally similar genes
  - Another solution is to allow “close paralogs” (“in-
Clusters from different “orthology” methods

- OrthoMCL in red; PhiGs in blue; InParanoid in green
- An "ortholog cluster" is made by one or more "slices" through the protein family tree
- Some combination of evolutionary rates and history of duplications
- Might miss genes that have inherited some but not all functions from the MRCA

Simplifying orthology

- View in terms of phylogenetic tree
  - If the MRCA is a speciation event, the genes are orthologs
  - If the MRCA is a duplication event, the genes are paralogs
- A transitive ortholog cluster can be defined using “least diverged orthologs” (Mi et al., NAR 2010)
  - Following a duplication event, only the most slowly evolving duplicate is included in the cluster
Orthology
only defined for PAIRS of genes

Paralogy
only defined for PAIRS of genes
Tools for homology-based

- InterproScan is among most highly-used automatic method
- Combines most popular web resources into one package
- Most of these are homology-based, searching a library of Hidden Markov Models (HMMs)
- Two distinct types of model
  - Domain-based (e.g. Pfam, SMART, Superfamily)
    - Model divergent groups usually with relatively ancient common ancestor
    - Domain shuffling has often occurred since this ancestor
    - Useful for seeing modular architecture
    - Will often predict only very general function, conserved since MRCA of module
  - Subfamily-based (e.g. PANTHER, TIGRfAMs, PRINTS)
    - Model groups that are more closely related (relatively recent)

Multiple sequence alignment:
used as basis for HMM, and for protein phylogeny modeling

Mouse wild type
Mouse albino

Mammalian tyrosinases excerpted from an alignment spanning vertebrates
Meaning of multiple sequence alignment:
identify homologous SITES

Mammalian tyrosinases excerpted from an alignment spanning vertebrates

HMM: “generative model”, first-order, learn “hidden” states and probabilities

Mammalian tyrosinases excerpted from an alignment spanning vertebrates
Profile-based approach

• Define a group of homologous sequences
  - Family/domain (e.g. Pfam)
  - Subfamily (e.g. PANTHER)
• For most methods, build an HMM to recognize members of the homologous group
• Annotate the group with functions/processes all known members have in common

Profile-based approach

• Driven by sequence relationships first, function later
  - Generally works well for molecular function
    • Sometimes loses specificity, depending on the approach
  - Loses specificity especially for biological process largely because of
    • co-option into new processes during evolution
    • Domain shuffling
Evolution of mutS family

molecular function

Binding of mutL

LOST

DNA mismatch recognition

Binding of mutL

Evolution of mutS family

biological process

Somatic hypermutation of immunoglobulin genes

Apoptosis

Maintenance of DNA repeats

Homologous recombination

DNA repair

Evolution of MSH2 subfamily
Many proteins are constructed of modular units: domains

- Domains are generally independently folding units
  - Minimum modular functional segment of protein sequence
- Domains are found in different combinations
  - They have arisen from “domain shuffling”: duplication and rearrangement

Finding domains: Domainer algorithm

Sonnhammer & Kahn, Protein Sci. 3:482, 1994
Prodom

Build graphs
Define domains:
- Termini
- Shuffling
- Repeats
- (3D structure)
Build MSAs

Build HMMs

Sonhammer & Kahn,
Protein Sci. 3:482, 1994


Pfam: Insulin receptor

http://www.sanger.ac.uk/Software/Pfam/
Pfam: Insulin receptor

INSR_HUMAN

This is the summary of UniProt entry INSR_HUMAN (P00213).

Description: Insulin receptor (EC:2.7.10.1)
Source organism: Homo sapiens (Human) (NCBI taxonomy ID 9606)
View Pfam proteome data.

Length: 1382 amino acids

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you directly to the Pfam database for each of the domains.

Opisthokonts  Metazoa  Eukaryota  LUCA

http://www.sanger.ac.uk/Software/Pfam/

Pfam: Insulin receptor

INSR_HUMAN

This is the summary of UniProt entry INSR_HUMAN (P00213).

Description: Insulin receptor (EC:2.7.10.1)
Source organism: Homo sapiens (Human) (NCBI taxonomy ID 9606)
View Pfam proteome data.

Length: 1382 amino acids

Pfam domains

This image shows the arrangement of the Pfam domain boundaries for each of the domains.

Molecular function
- protein-tyrosine kinase activity (GO:0004713)
- ATP binding (GO:0005524)

Biological process
- protein amino acid phosphorylation (GO:0006468)

Cellular component
- membrane (GO:0016020)

http://www.sanger.ac.uk/Software/Pfam/
Pfam: Protein Y kinase domain is independently folding
Bottom line for homology inference by phylogenetic analysis

- We are not necessarily interested in orthology per se
- We are interested in the history of inheritance, gain, loss and modification of function during gene family evolution
  - “Least diverged orthologs” can generally be identified as the more slowly evolving duplicated gene (more similar to ancestor so more likely to conserve most or all functions)
  - Even “one-to-one” orthologs (or LDOs) can diverge in function; one way to detect this is in accelerated evolutionary rates
  - Even divergent paralogs are likely to retain some subset of ancestral functions and these can be annotated transitively

Phylogenetic analysis

- “Tree” representation of evolutionary relationships between sequences
- Models primarily mutation: substitution and (sometimes) small indel events
- Does not model events prior to shared common ancestor, e.g. prior to a new domain combination or horizontal transfer
Phylogenetic methods

• Often begin with clustering method, starting with pairwise relationships
• Followed by multiple sequence alignment
• Followed by tree reconstruction
  - Character evolution methods
    • E.g. parsimony, Neighbor Joining, UPGMA
  - Sequence evolution methods
    • Maximum likelihood, Bayesian methods
  - Genome evolution methods
    • Build on another method, adding information about

Complications for phylogenetic

• Over “short” (but not too short!) periods of evolutionary time, or high degree of functional constraint, phylogenetic analysis is an excellent tool
• Over longer periods, it can be complicated by events that can accumulate to become important caveats
  - Domain shuffling (partial duplications fused together)
  - “Distant homology”: sequence divergence that masks common ancestry
  - Complementation of function by unrelated gene
    • “functional orthologs”
**Curated GO annotation sources**

- Experimental annotation
  - [http://www.geneontology.org](http://www.geneontology.org)
  - Remember to filter by evidence code
- Homology annotations
  - PANTHER ([http://www.pantherdb.org](http://www.pantherdb.org))
- Note that these two sources will be merging as the collaboration continues over the next couple of years
  - PanTree ([http://www.pantree.org](http://www.pantree.org))

**Pathway annotation sources**

- Still distributed among many sources
- Pathway Commons is starting to collect BioPAX representations with the aim of relating entities and reactions across different data sources
What if similarity to characterized proteins is weak or non-existent?

• Protein structure can be conserved after sequences have diverged past the point of statistically significant similarity
• Structure-based predictions usually give only a very general prediction of function

• Solutions:
  - Make the HMMs more sensitive by careful alignment of globally conserved amino acids (or properties):
    • Superfamily database
  - Use various ways of mapping a sequence onto a structure ("threading")

Crotonase superfamily: conservation of chemistry

Babbitt & Gerlt, J.Biol.Chem. 272:30591, 1997
Structure-function linkage database (SFLD)

- [http://sfld.rbvi.ucsf.edu](http://sfld.rbvi.ucsf.edu)

Other methods

- Short (often repeated) sequence motifs
  - These can arise by convergent evolution, so these are similar but not necessarily homologous
  - Main example is cellular component prediction
    - SignalP for signal peptides
    - TMHMM for transmembrane helix prediction

- “Guilt by association” methods
  - Genes used in the same coordinated process will need to be co-expressed, at least in some organisms
    - Transcriptional co-regulation
    - Genomic structure (operons, similar expression control “modules”/transcription factor binding sites)
Inference of functional relationships from transcriptional co-regulation

- mRNA expression data
  - Genes that are co-expressed over a series of conditions are likely to participate in the same overall biological process (e.g. M. Eisen, LBL)

- Identification of regulatory regions in genomic sequence
  - Genes sharing common blocks of regulatory elements are likely to be controlled by the same set of transcription factors
  - ChIP-chip experiments to find sequences that bind to given transcription factors
  - Evolutionary analysis of noncoding genomic sequence (e.g.

Inference of function from microbial genome structure

- Phylogenetic profiling: genes involved in the same process are often either gained or lost together

- “Rosetta stone”: proteins that are covalently linked (one protein) in one genome are often involved in the same process even in genomes where they are separate genes

- Operons: proteins encoded in the same operon generally participate in the same process (co-regulation)
Inference from network structure

- Sometimes the “least diverged ortholog” does not play the same role in a network
- Functional complementation can be detected by comparing networks in two different species/cell types, etc. (Trey Ideker, UCSD: Bandyopadhyay et al. Genome Res. 2006)

Inference of functional relationships from protein-protein

- Protein complexes
- Transient protein-protein interactions
  - E.g. phosphorylation, proteolysis, ubiquitination
- Databases:
  - IntAct
  - Predictions in Pfam (domain-domain interactions)