

Transcriptome Functional Annotation

**Dr. Jung Soh
657.001 Transcriptomics (WS 2018/2019)**

Why functional annotation

- ▶ Get biological understanding from features (transcripts)
- ▶ Validate experiments (RNA-seq expectations)
- ▶ Generate new hypotheses (unexpected findings)
- ▶ Limitations
 - Only as good as published or known information about genes and proteins in known species
 - Reliable annotation difficult to achieve for novel or mixed species

Transcriptome functional annotation

- ▶ Coding region prediction
 - ORF (open reading frame) identification
 - Likely coding region prediction
- ▶ Protein search
 - Search transcripts and predicted proteins against reliable protein databases (e.g. Swiss-Prot)
- ▶ Functional information collection
 - Use various annotation sources

Sources of annotation

- ▶ Proteins and protein domains
 - Swiss-Prot, Pfam, InterPro
- ▶ Gene Ontology
 - Most popular, but can be too general
- ▶ Pathways
 - KEGG, Reactome, BioCata, PANTHER
- ▶ Other databases
 - Orthologous groups, diseases

Protein database: Swiss-Prot

- ▶ Most reliable, manually annotated protein database

UniProtKB
UniProt Knowledgebase

Swiss-Prot (558,590)
 Manually annotated and reviewed.

TrEMBL
(126,780,198)
 Automatically annotated and not reviewed.

UniRef
Sequence clusters


UniParc
Sequence archive


Proteomes


Supporting data

Literature citations


Cross-ref. databases


Taxonomy


Diseases


Subcellular locations


Keywords


Protein database: Pfam

► Protein families and domains

The screenshot shows the Pfam 32.0 homepage. At the top left is the EMBL-EBI logo. To its right are navigation links: HOME | SEARCH | BROWSE | FTP | HELP | ABOUT. On the right side is the Pfam logo with a search bar labeled "keyword search" and a "Go" button.

Pfam 32.0 (September 2018, 17929 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [Less...](#)

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

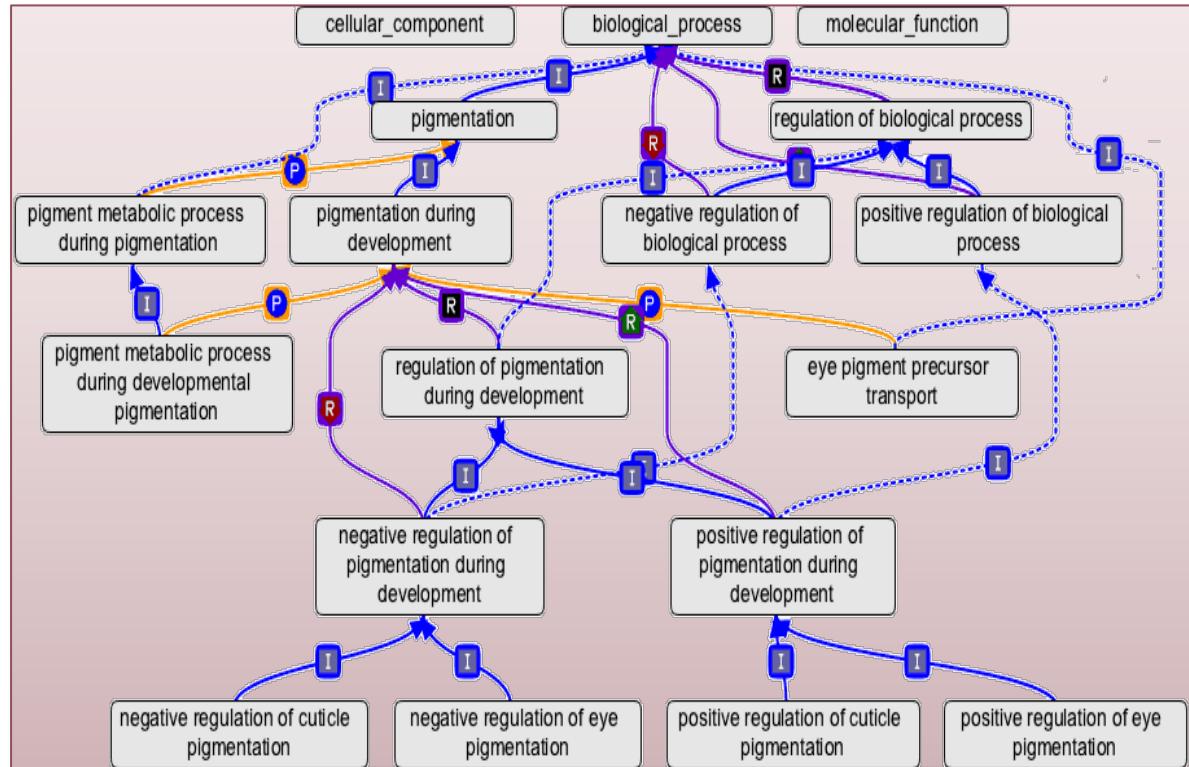
Pfam also generates higher-level groupings of related entries, known as **clans**. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the [UniProt Reference Proteomes](#) but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords

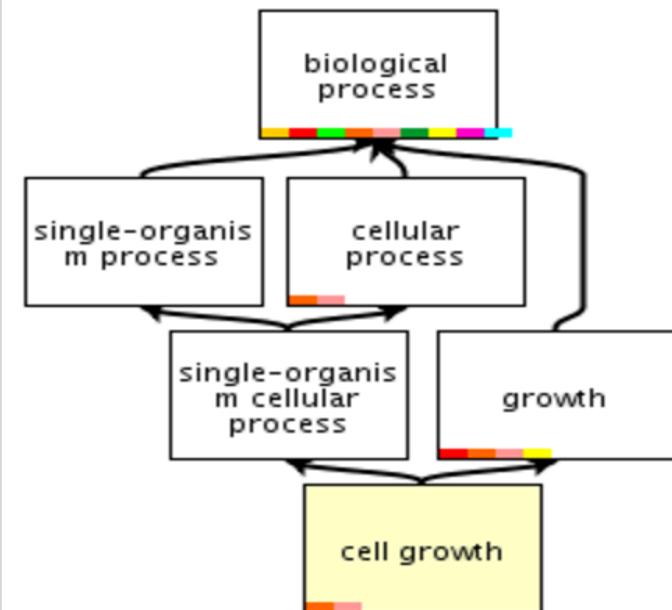
Gene Ontology (GO)

- ▶ GO project provides
 - Controlled vocabularies of terms representing gene product properties
- ▶ 3 ontologies
 - Cellular Component (CC)
 - Biological Process (BP)
 - Molecular Function (MF)



GO term example

- id: GO:0016049
- name: **cell growth**
- namespace: **biological_process**
- def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." [GOC:ai]
- subset: goslim_generic
- subset: goslim_plant
- subset: gosubset_prok
- synonym: "cell expansion" RELATED []
- synonym: "cellular growth" EXACT []
- synonym: "growth of cell" EXACT []
- is_a: GO:0009987 ! cellular process
- is_a: GO:0040007 ! growth
- relationship: part_of GO:0008361 ! regulation of cell size

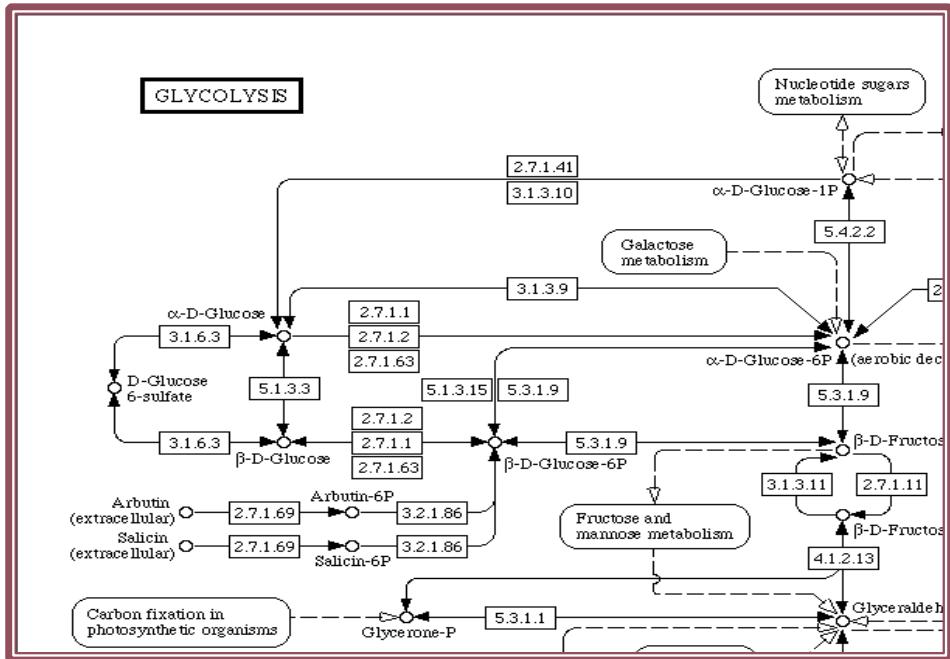
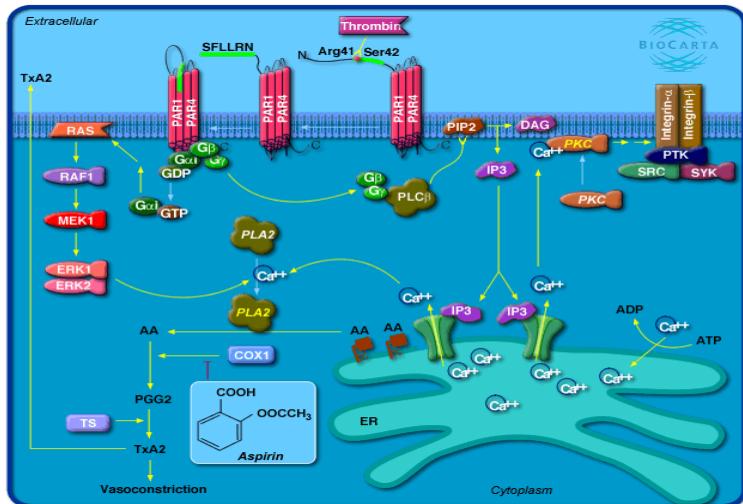


GO annotation

- ▶ Which biological processes or molecular functions are enriched in my gene list?
- ▶ Results given as a set of enriched or over-represented GO terms
 - Degrees of enrichment sometimes available
 - Interpretation is important
 - Often need more information from other annotation sources

Pathways

- ▶ KEGG pathways most popular
- ▶ Many others exist
 - BioCarta
 - MetaCyc
 - Reactome



Pathway annotation

- ▶ Are there specific pathways enriched in my gene list?
- ▶ What are other genes involved in those pathways?
- ▶ Results given as a set of pathways
 - More than simple terms (enzymes, involved genes, reactions)
 - Interpretation and cross-check with other sources of annotation necessary

Orthologous groups: EggNOG

- ▶ Orthologous groups of genes



Lab overview

