Worksheet 1: Introduction to Information Searching from Public Databases and Functional Analysis of Network

From Omics to Functional Analysis

In order to understand biological systems and mechanisms of disease, researchers often perform large scale studies and analyze massive amounts of data. Omics approaches such as Transcriptomics, Metabolomics, and Proteomics, are generating data that is being stored in a number of databases, which are in turned queried by the scientific community.

The necessary step after collecting data from an Omics study or database, is to run a **Functional Analysis**. The goal of the functional analysis is to evaluate the evidence to reach conclusion about what proteins, metabolites or other molecules are doing in the cell.

In the past, the function or description of every gene or protein was annotated manually using comments fields, which were not standardized or compatible with many data mining tools. Efforts to make gene and gene product annotation consistent across multiple species and to facilitate access to the data, led into the creation of the GeneOntology (GO, www.geneontology.org).

Gene Ontology started as collaboration between the databases for model organisms of mouse (MGI), fruit fly (FlyBase), and baker's yeast (SGD). It now receives contributions approximately 20 groups, which include Worm Base, dictyBase, J. Craig Venter, Reactome and others.

Ontology = controlled vocabulary; a set of standard terms or phrases, each with a precise definition

GO describes how gene products behave in a cellular context. The GO consists of three controlled vocabularies:

- Molecular Function (what does the gene product do?):
  GO: 0003872, "6-phophofructokinase activity"
- Biological Process (why does it Biological Process perform the activity?):
  GO: 0006096, "glycolysis"
- Cellular Component (where does it act?):
  GO: 0005737, "cytoplasm"

GO annotators use evidence codes to record the type of evidence that was used to make the

annotation. Some evidence codes are:

EXP: Inferred from **Exp**eriment

IEA: **I**nferred from **E**lectronic **A**nnotation

IDA: **I**nferred from **D**irect **A**ssay

Useful Databases

Numerous databases have been created to store and query information about genes, proteins and other biological molecules. Protein-protein interaction (PPI) and enzyme reactions are two of the main types of data used in networks and pathways analysis. Understanding protein-protein interactions is also important for the investigation of signaling pathways. Below, we will explore the use of protein-protein interaction databases, as well as network tools that will allow us identify affected pathways/networks in pathogen and host.

A brief word on Commercial Databases and Tools for Pathways and Networks Analysis

- Ingenuity Pathways Analysis (IPA) http://ingenuity.com/, and GeneGo, MetaCore, http://www.genego.com/, which are manually curated data bases with mammalian data. Facilitate extracting and comparing functional data from large lists of genes or metabolites.
- Pathway Studio http://www.ariadnegenomics.com/products/pathway-studio/, which is computer generated database of functional relationships between molecules. It also provides tools for functional classification.

Noncommercial Databases and Tools

Protein Databases:

- UniProt http://www.uniprot.org/
  It combines the annotations of SwissProt (manually curated) with TrEM (computationally generated) into one single entry
- NCBI http://www.ncbi.nlm.nih.gov/protein
  All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR (Protein Information Resource) + PRF (Protein Research Foundation).

Enzyme and Metabolic Pathways Databases

- Enzyme Commission http://www.chem.qmul.ac.uk/iubmb/enzyme/ is a collection of enzymatic reactions, which contains official nomenclature of the classification of enzymes by the reactions they catalyze
  For example, E.C.4.2.1.11
- BRENDA www.brenda.enzymes.org (The Comprehensive Enzyme Information System)
  Collection of enzyme functional data

- KEGG http://www.genome.jp/kegg/kegg2.html

  Contains manually drawn pathway maps

  Also contains maps of cellular processes, genetic information processing and others

  KEGG Brite (Functional hierarchies and binary relationships of biological entities) allows classifications of large gene lists into functional categories
- MetaCyc http://biocyc.org/metacyc/index.shtml

  Contains metabolic pathways that have been manually curated for a large number of organisms

  Allows for comparative analyses of pathways and pathway holes

Protein-Protein or Protein-DNA Interaction Databases

- **BioGRID** (The **Bio**logical **G**eneral **R**epository for **I**nteraction **D**atasets)–acurated database http://www.thebiogrid.org/
- Pathways Commons http://www.pathwaycommons.org/pc/

  Integrates data from BioGRID, HumanCyc, IntAct, Reactome, MINT, NCI, SBCNY, HPRD, and CancerCell Map. Important: Of these sources, IntAct and MINT contain data for some parasites.
- STRING http://string.embl.de/

  Combines known and predicted protein-protein interactions

  Predictions are derived from Genomic context, high throughput experiments, coexpression and previous knowledge reported in PubMed.

Exercise: Information Searching from Public Databases

Objective: Work with available resources for information retrieval of proteins, pathways, and learn how to use BLATing algorithm to compare DNA sequence with the genome database and design a RNA probe for a specific gene.

A. Retrieving protein information from UniProtKB

1. Point your browser to http://www.uniprot.org/
2. Select "UnitProtKB" and input "egfr" in the Query box. Hit search button.
3. Select "A2VCQ7" entry.
4. Read the information in "Names and origin", "Protein attributes", and "Ontologies".

Q: Where does this protein locate in the cell? What function it performs? What does the biological process it joins to perform?

B. Find out the 3D structure of a protein from PDB

1. Point your browser to http://www.ncbi.nlm.nih.gov/Structure/index.shtml. In the search box select "structure" and input "egfr" and hit GO button.
2. Select a proper entry and view the crystal structure of the protein.

C. Find out the pathway relevant to a specific protein

1. Point your browser to the KEGG pathway website http://www.genome.jp/kegg/pathway.html.
2. Input "has" (for Homo Sapiens) into the organism box and "EGFR" in the keyword box. Press go button.
5. Press "hsa04012" pathway entry to input ErbB signaling pathway-Homo sapiens.
6. Study the pathway summary and press hsa04012 pathway map to examine the pathway diagram.

Q: How many pathways will be activated by ErbB1 (EGFR) dimerization? Describe the biological functions of the three most important pathways activated by EGFR.

EGFR is a transmembrane glycoprotein that is a member of the protein kinase superfamily. This protein is a receptor for members of the epidermal growth factor (EGF) family. EGFR is a cell surface protein that binds to EGF. Binding of the protein to a ligand induces receptor dimerization and tyrosine autophosphorylation and leads to cell proliferation. Mutations in this gene are associated with lung cancer.

D. Design a RNA probe that recognizes the egfr gene

1. Go to NCBI Gene Search http://www.ncbi.nlm.nih.gov/gene .
2. Select "Gene" database and input "egfr" in the search box. Press search button.

3. Select “the Genomic regions, transcripts, and products” section and select “FASTA” format.
4. Select and copy all or a portion of the DNA sequence.
5. To verify your sequence by BLATing, point your browser to http://genome.ucsc.edu/cgi-bin/hgBlat?command=start. Make sure you select the species “human”. Paste the DNA sequence into the search box and hit “search” button.
6. Select the “Browser” of the result that yields highest score for 100% identity (i.e., for the entire sequence). This will bring you to a screen that displays your target sequence location within the gene. Adequate visualization of the entire gene will require zoom out/in or movement of the cursor.
7. Identify the sequences of the coding exon or intron you wish to target. For targeted knockout, target to the first coding exon. Click on the actual image depicting your gene of interest, this will bring you to a new screen.
8. Select “Sequence and Links”. This will bring you to the “Sequence and Links to Tools and Databases” section
9. Click on the “Genomic Sequence”. This will bring you to the “Get Genomic Sequence Near Gene” screen.
10. Check “CDS” and “Introns” only and “One FASTA record per gene”. Click “submit”.
11. The sequence generated will show coding sequences in upper case and introns in lower case. Copy and paste the exon sequences you desire to target (coding exon 1 is recommended for knockout studies).
12. Go to http://genome.ucsc.edu/cgi-bin/hgBlat?command=start to re-Blat the exon sequence to confirm the exon sequence you wish to target. If your sequence spans the entire or a portion of the exon you wish to target, then your sequence is confirmed.