Homework for module 4:

https://www.youtube.com/watch?v=ag74g6jhTOQ
Basic bioinformatics concepts, databases and tools

Module 4

Beyond sequences
Biological data does not only consist of sequences:

- Functional annotation
- Pathways
- Interaction data
- Expression data
- 3D structures
- Variation data

**Relation between modules 1, 2, 3 and 4**

- **Module 1**: Sequence databases
  - `>DNAseq1`
  - GACAAACTTAT
  - `>DNAseq2`
  - CTATATCTTTAAA
  - ...

- **Module 2**: Sequence comparison
  - GACAAACTTAT
  - CTATATCTTTAAA
  - ...

- **Module 3**: Sequence analysis
  - `>DNAseq1`
  - GACAAACTTAT
  - `>proteinseq1`
  - MDKLMHLEGENMESEQENCING
  - Secreted? Phosphorylated?

- **Module 4**: Beyond sequences
  - `>PKN1`
  - GO Biol proc: glycolysis
  - 3D structure: PDB 1PKN
  - Interactions: TBS4
  - ...

...
Functional annotation of a gene

Gene name
Gene symbol
Biological function
Gene Ontology terms
Pathways
Interaction partners
Domains
Functional motifs

…
How to functionally annotate a gene?

Homology searches -> BLASTP against SwissProt/OrthoMCL

MSA (+ HMMER)

Motif searches -> InterProScan / PfamScan / ScanProsite

Prediction of characteristics -> SignalP (secretion)

TargetP (localisation)…

Mine biological databases -> Gene Ontology

Pathway databases

Interaction databases…
Gene Ontology is a biological ontology...

Ontology = representation of knowledge

GO: hierarchy of biological terms

tree structure: general -> specific

http://www.geneontology.org/
...representing knowledge on protein location and function

Provides terms to describe how gene products behave in a cell

Gene products are associated with terms from GO based on evidence

GO consists of 3 root ‘ontologies’:

1. molecular function
2. biological process
3. cellular component

Description and tutorial on GO
http://www.arabidopsis.org/help/tutorials/go1.jsp
Goal of GO

Provide a common language to describe functionality of gene products

Use of consistent vocabulary allows to compare genes from different species

Make functional annotation understandable to computer (program)s

Characterize lists of genes from RNASeq, proteomics, interaction studies
Where do the data come from?

GO is developed by 15 contributing model organism databases.

Terms / Structure / Associations are continuously updated:
none of the files is older than 1 year!

⇒ if you use GO-based tools you need to check if they update their GO copies

⇒ In publications always provide version of GO that was used

Not enough curators to cope with enormous amount of publications

⇒ if your protein lacks GO associations it does not mean it cannot perform a certain function

10 tips for using GO  http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003343
Search Gene Ontology using dedicated tools

QuickGO at EBI:  http://www.ebi.ac.uk/QuickGO/

AmiGO at GO:  http://amogo.geneontology.org

Slimmer-type tools

Tools that use 'shrunken' GO sets (GO-slims)

GO-slims = subsets of GO:

- Terms that are relevant for species
- Too general and too detailed terms are removed

Recommended for microarray / RNA-Seq analysis
Not all functional information is captured by GO

Pathway databases:

KEGG:  http://www.kegg.jp
  many species, many pathways

Reactome at EBI: http://www.reactome.org

Metacyc and other cyc-databases: http://metacyc.org
  many organisms, less pathways

SignaLink: http://signalink.org/
  human, worm and fly, signaling pathways only

Where does the data come from?

**Experts** in a particular area in biology describe pathways and provide references. Curators work with the experts to convert this information into the right format. When the module is ready, it is sent for peer-review by another expert in the field.
Experts publish pathways so databases can refer to them

<table>
<thead>
<tr>
<th>Reference</th>
<th>PMID:18245809</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Kovacs GG, Budka H</td>
</tr>
<tr>
<td>Title</td>
<td>Prion diseases: from protein to cell pathology.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>PMID:15695097</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Campana V, Sarnataro D, Zurzolo C</td>
</tr>
<tr>
<td>Title</td>
<td>The highways and byways of prion protein trafficking.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>PMID:14661267</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Nunziante M, Gilch S, Schatzl HM</td>
</tr>
<tr>
<td>Title</td>
<td>Prion diseases: from molecular biology to intervention strategies.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>PMID:17051207</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Caughey B, Baron GS</td>
</tr>
<tr>
<td>Title</td>
<td>Prions and their partners in crime.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>PMID:18391177</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Linden R, Martins VR, Prado MA, Cammarota M, Izquierdo I, Brentani RR</td>
</tr>
<tr>
<td>Title</td>
<td>Physiology of the prion protein.</td>
</tr>
</tbody>
</table>
Reactome is only curated for human

Computational inferred events

We use the set of manually curated human reactions to electronically infer reactions in twenty evolutionarily divergent eukaryotic species for which high-quality whole-genome sequence data are available, and hence a comprehensive and high-quality set of protein predictions exists. These species include the laboratory mouse and rat, the nematode *C. elegans*, budding and fission yeasts, and two plants. The estimated success rates of our orthology inference strategy can be stated as ‘the percentage of eligible reactions, defined in step 2 below, in the current human data set for which an event can be inferred in the model organism. By this measure, success rates range from 89.1% for the laboratory mouse to 6.7% for the protozoan *P. falciparum*.

Orthologs obtained from Ensembl

All human – human reactions are translated to other species

Via mapping of human proteins to orthologs
Signalink uses published pathways + info from interaction dbs

INTERACTIONS BETWEEN PATHWAY MEMBERS

AREG ➔ EGFR

- This is a stimulatory, direct interaction
- Sources:
  - Biogrid (experimentally verified, 21071413 PubMed)
  - Signalink 2.0 (manual curation) (experimentally verified, 23331499 PubMed)
- References:
  - 10085134 PubMed
  - 10209155 PubMed
- Confidence score:
  - GO Semantic Similarity: 0.601198 PubMed
  - GO Semantic Similarity: 0.601198 PubMed
- AREG belongs to pathways: RTK (core)
- EGFR belongs to pathways: RTK (core), JAK/STAT (core)
Not much consistency between pathway databases

Number of overlapping entities: genes, enzymes, reactions, metabolites

of 5 metabolic pathways

in 5 pathway databases KEGG - Reactome – HumanCyc – BIGG - EHMM

- entity found in all databases
- entity found in one database

Efforts to improve consistency

Always use and compare results of multiple databases!

WikiPathways: http://www.wikipathways.org/index.php/WikiPathways

Wikipedia for pathways

Every pathway has its page and can be edited by anyone from scientific community

Central curators

PathwayCommons: http://www.pathwaycommons.org/

composite pathway database (human only)

Very low consistency among pathway databases http://www.biomedcentral.com/1471-2105/11/449
Databases of interactions

Difficult to make distinction between pathway and interaction databases

**Protein - DNA**: see module 3

**Protein - Protein:**

Intact: [http://www.ebi.ac.uk/intact](http://www.ebi.ac.uk/intact)

STRING: [http://string-db.org/](http://string-db.org/)


PSIQUIC: [http://www.ebi.ac.uk/Tools/webservices/psicquic/view/](http://www.ebi.ac.uk/Tools/webservices/psicquic/view/)

Composite database

**HIPPIE**: [http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/](http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/)

composite database for human protein-protein interactions

...
Where do data come from?

**Experimental:**
- yeast two-hybrid
- protein-fragment complementation assays (PCA)
- affinity purification/mass spectrometry
- protein microarrays
- fluorescence resonance energy transfer (FRET)

**Computational:**
- inference based on orthology
- analysis of fusion proteins
- classification
- based on sequence motifs
- based on structural patterns

https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction_prediction
Intact

**Experimentally proven direct** interactions from **publications**

Thorough curation

Organisms: mainly human, mouse, yeast, Arabidopsis, E. coli, worm, fruit fly

![Interaction detection methods](image)

- tandem affinity purification
- two hybrid array
- anti tag coimmunoprecipitation
- pull down
- two hybrid
- two hybrid prey pooling approach
- anti bait coimmunoprecipitation
- two hybrid pooling approach
- others

100.000 Interactors
700.000 Interactions
BioGRID

Experimentally proven direct interactions from publications

Shallow curation

Organisms: around 100 species

1.400.000 Interactions

- Current Build Statistics (3.4.146) - March 2017
- Physical and Genetic Interaction Statistics
- Chemical Interaction Statistics
- Post Translational Modification (PTM) Statistics
Published experimental interactions from databases + computational predictions

Nice visualization

Organisms: more than 2000 species

Data Sources

Interactions in STRING are derived from five main sources:

- Genomic Context Predictions
- High-throughput Lab Experiments
- (Conserved) Co-Expression
- Automated Textmining
- Previous Knowledge in Databases

10,000,000 Interactors
180,000,000 Interactions
PSICQUIC is a composite database on interaction data

Fetch data on-the-fly: always most recent versions of databases

Extra info on what kind of data
Chemical – Protein:

STITCH: [http://stitch.embl.de/](http://stitch.embl.de/)

Drug – Target: based on literature

DrugBank: [http://www.drugbank.ca/](http://www.drugbank.ca/)

Therapeutic Target Database: [http://bidd.nus.edu.sg/group/cjttd/](http://bidd.nus.edu.sg/group/cjttd/)

PharmGKB: [http://www.pharmgkb.org/](http://www.pharmgkb.org/)

Drug Gene Interaction database: [http://dgidb.genome.wustl.edu/](http://dgidb.genome.wustl.edu/) ...

Drug – Target: based on high throughput compound screening


The Binding Database: [http://www.bindingdb.org/bind/index.jsp](http://www.bindingdb.org/bind/index.jsp)

Drug – Protein that influences response but is not necessarily a target:

T3DB: [http://www.t3db.org/](http://www.t3db.org/)

SIDER: [http://sideeffects.embl.de/](http://sideeffects.embl.de/) ...
Databases specific for drug – target interactions in cancer

CAncerREsource2: http://data-analysis.charite.de/care/

My Cancer genome: http://www.mycancergenome.org/

MD Anderson: https://pct.mdanderson.org/

CIVIC: https://civic.genome.wustl.edu/
Docking: modeling of interactions

Predicting the 3D structure of the complex formed by the interaction

1. Millions of possible complex structures are generated

Prior knowledge required to limit the number of possible structures

- e.g. Location of active site, info from mutants or homologs, stereochemical clashes
- Still millions of possibilities remain

2. The most likely complex structure is selected

= the structure with the lowest free energy

Similar to creation of phylogenetic tree but there you look for the highest score
Functional characterization of a set of genes

What’s the biology behind a list of genes?

RNASeq / Proteomics generate lists of 100s significant potential targets
Impossible to evaluate each target individually
See what the genes in the list have functionally in common
e.g. RNASeq mutant vs wt: upregulated genes
  -> 60% are involved in MAP kinase pathway
  -> MAP kinase pathway is affected by mutation

Gene enrichment analysis
Gene set enrichment analysis
Gene enrichment analysis

Find functions and pathways that **are-enriched-in** potential targets

1. Add functional annotations to the data
2. Define background: full set of genes measured by the platform
3. Perform test to identify enriched 'functions', 'diseases', 'pathways' ...

Enriched = Over-represented

Occurring more frequently than expected based on background data
Tools for gene enrichment analysis

Most popular tool:
DAVID: http://david.abcc.ncifcrf.gov/

Cool tools:
Enrichr: http://amp.pharm.mssm.edu/Enrichr/index.html#
WebGestalt: http://www.webgestalt.org
ToppGene: https://toppgene.cchmc.org/

All tools work with GO-slims but not always the up-to-date version
Overview of tools: http://nar.oxfordjournals.org/content/37/1/1

Video tutorial on DAVID  https://www.youtube.com/watch?v=x1u9mm6b7N0
DAVID uses info from GO and pathways to interpret gene lists

**Most popular** tool

**Poor visualization** of results

Redundant: duplications because of multiple annotation resources

Includes non-specific terms e.g. Biological process, Cell, Metabolism...

**Proprietary DAVID Knowledgebase:** The higher integration level and comprehensive annotation data coverage of the DAVID Knowledgebase, the backend database of DAVID analytic tools, are rarely found in other similar works (which usually only use one or a few public resources as their backend database). In the DAVID Knowledgebase, >20 gene identifier types and >40 functional annotation categories form dozens of heterogeneous public databases are comprehensively integrated by a unique single linkage method developed by DAVID team. Without doubts, the DAVID Knowledgebase, as a 'one-stop shopping' warehouse, maximizes the analytic potential for all functional annotation algorithms.

Many IDs supported

> 40 functional annotations from 12 public databases: GO, KEGG, SwissProt, Pubmed, Interpro

<table>
<thead>
<tr>
<th>Ontology (≥40 million records)</th>
<th>Protein Domain/Family (≥ 15 millions)</th>
<th>Sequence Features (≥21 millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO BIOLOGICAL PROCESS</td>
<td>BLOCKS_ID</td>
<td>ALIAS_GENE_SYMBOL</td>
</tr>
<tr>
<td>GO MOLECULAR FUNCTION</td>
<td>COG_KOG_NAME</td>
<td>CHROMOSOME</td>
</tr>
<tr>
<td>GO CELLULAR COMPONENT</td>
<td>INTERPRO_NAME</td>
<td>CYTOBAND</td>
</tr>
<tr>
<td>PANTHER BIOLOGICAL PROCESS</td>
<td>PDB_ID</td>
<td>GENE_NAME</td>
</tr>
<tr>
<td>PANTHER MOLECULAR FUNCTION</td>
<td>PFAM_NAME</td>
<td>GENE_SYMBOL</td>
</tr>
<tr>
<td>COG_KOG_ONTOLOGY</td>
<td>PIR_ALN</td>
<td>HOMOLOGOUS_GENE</td>
</tr>
<tr>
<td><strong>P-P Interaction (≥ 4 millions)</strong></td>
<td>PIR_HOMOLOGY_DOMAIN</td>
<td>ENTREZ_GENE_SUMMARY</td>
</tr>
<tr>
<td>BIND</td>
<td>PIR_SUPERFAMILY_NAME</td>
<td>OMIM_ID</td>
</tr>
<tr>
<td>DIP</td>
<td>PRINTS_NAME</td>
<td>PIR_SUMMARY</td>
</tr>
<tr>
<td>MINT</td>
<td>PRODOM_NAME</td>
<td>PROTEIN_MW</td>
</tr>
<tr>
<td>NCICB CAPATHWAY</td>
<td>PROSITE_NAME</td>
<td>REFSEQPRODUCT</td>
</tr>
<tr>
<td>TRANSFAC_ID</td>
<td>SCOP_ID</td>
<td>SEQUENCE_LENGTH</td>
</tr>
<tr>
<td>HIV_INTERACTION</td>
<td>SMART_NAME</td>
<td>SP_COMMENT</td>
</tr>
<tr>
<td>HIV_INTERACTION_CATEGORY</td>
<td>TIGRFAMS_NAME</td>
<td></td>
</tr>
<tr>
<td>HPRD_INTERACTION</td>
<td>PANTHER_SUBFAMILY</td>
<td></td>
</tr>
<tr>
<td>REACTOME_INTERACTION</td>
<td>PANTHER_FAMILY</td>
<td></td>
</tr>
<tr>
<td><strong>Disease Association (≥9,000)</strong></td>
<td><strong>Pathways (≥50,000)</strong></td>
<td><strong>Functional Category (≥6.9 millions)</strong></td>
</tr>
<tr>
<td>GENETIC_ASSOCIATION_DB</td>
<td>BioCarta</td>
<td>PIR_SEQ_FEATURE</td>
</tr>
<tr>
<td>OMIM DISEASE</td>
<td>KEGG_PATHWAY</td>
<td>SP_COMMENT_TYPE</td>
</tr>
<tr>
<td><strong>Literature (≥2.8 millions)</strong></td>
<td>PANTHER_PATHWAY</td>
<td>SP_PIR_KEYWORDS</td>
</tr>
<tr>
<td>GENERIF_SUMMARY</td>
<td></td>
<td>UP_SEQ_FEATURE</td>
</tr>
<tr>
<td>PUBMED_ID</td>
<td>PANTHER_PATHWAY</td>
<td></td>
</tr>
<tr>
<td>HIV_INTERACTION_PUBMED_ID</td>
<td>BBID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KEGG_REACTION</td>
<td>GNF Microarray</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UNIGENE_EST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGAP SAGE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGAP EST</td>
</tr>
</tbody>
</table>
Welcome to DAVID 6.8 with updated Knowledgebase (more info). ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Release & Version Information


-- The DAVID Knowledgebase completely rebuilt
-- Entrez Gene integrated as the central identifier to allow for more timely updates
while still incorporating Ensembl and Uniprot as integral data sources
-- New GO category (GO Direct) provides GO mappings directly annotated by the source database (no parent terms included)
-- New annotation categories
-- New list identifier systems added for list uploading and conversion
-- A few bugs fixed

Annotations not very up to date !!
Publish adjusted p-values

<table>
<thead>
<tr>
<th>Sublist</th>
<th>Category</th>
<th>Term</th>
<th>RT</th>
<th>Genes</th>
<th>Count</th>
<th>%</th>
<th>P-Value</th>
<th>Benjamin</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>ribosome biogenesis</td>
<td>RT</td>
<td>10</td>
<td>3.5</td>
<td>1.2E-12</td>
<td>1.9E-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>rna processing</td>
<td>RT</td>
<td>9</td>
<td>3.1</td>
<td>4.6E-11</td>
<td>3.7E-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>ribonucleoprotein</td>
<td>RT</td>
<td>19</td>
<td>6.6</td>
<td>1.2E-7</td>
<td>6.6E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>ribosomal protein</td>
<td>RT</td>
<td>17</td>
<td>5.9</td>
<td>9.6E-7</td>
<td>3.8E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>wd repeat</td>
<td>RT</td>
<td>13</td>
<td>4.5</td>
<td>2.6E-6</td>
<td>8.5E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>nucleus</td>
<td>RT</td>
<td>40</td>
<td>13.8</td>
<td>1.3E-4</td>
<td>3.4E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>rna-binding</td>
<td>RT</td>
<td>11</td>
<td>3.8</td>
<td>4.9E-4</td>
<td>1.1E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>helicase</td>
<td>RT</td>
<td>7</td>
<td>2.4</td>
<td>2.1E-3</td>
<td>4.1E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>protein biosynthesis</td>
<td>RT</td>
<td>6</td>
<td>2.1</td>
<td>3.1E-2</td>
<td>4.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>glycosidase</td>
<td>RT</td>
<td>8</td>
<td>2.8</td>
<td>3.7E-2</td>
<td>4.6E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>hydrolase</td>
<td>RT</td>
<td>23</td>
<td>8.0</td>
<td>4.4E-2</td>
<td>4.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>polysaccharide</td>
<td>RT</td>
<td>3</td>
<td>1.0</td>
<td>8.3E-2</td>
<td>6.9E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>degradation</td>
<td>RT</td>
<td>2</td>
<td>0.7</td>
<td>8.8E-2</td>
<td>6.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP_PIR_KEYWORDS</td>
<td>abscisic acid</td>
<td>RT</td>
<td>3</td>
<td>1.0</td>
<td>8.9E-2</td>
<td>6.6E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Enrichr: more up-to-date, nicer visuals, limited species

http://amp.pharm.mssm.edu/Enrichr/index.html#stats

Human, mouse, rat

Enrichr tutorial  http://amp.pharm.mssm.edu/Enrichr/#help
Enrichr: more up-to-date, nicer visuals, limited species

New ARCHS4 and GO Libraries, Counters, and Combined Score – August 24th, 2017

For this release we added five libraries generated from the ARCHS4 project. ARCHS4 contains processed RNA-seq data from over 100,000 publicly available samples profiled by the two major deep sequencing platforms HiSeq 2000 and HiSeq 2500. After alignment and normalization, we computed co-expression correlation for all human genes. From this co-expression correlation matrix, we generated three new libraries: a) top 300 genes that are co-expressed with transcription factors; b) top 300 genes that are co-expressed with kinases; and c) top 300 genes that are co-expressed with under-studied drug targets from the Illuminating the Druggable Genome (IDG) project. We also added two additional libraries created from ARCHS4: genes that are highly expressed in human cell-lines and tissues. These two libraries were created by z-scoring the expression of each gene across all cell-lines or tissues. In addition, we updated the Gene Ontology libraries by removing high level terms and following a more rigorous process based on an Enrichr user suggestion. Two new counters were added to the landing page showing the number of libraries, and the number of terms across all libraries. We also changed the way the combined score is calculated by multiplying the unadjusted, instead of the adjusted, p-values with the z-scores.

New and Updated Libraries – May 5th, 2017

In this release of Enrichr we added and updated several gene set libraries. The MGI Mammalian Phenotype library was updated and now contains 5231 terms that describe phenotypes. This library has many more terms than the old MGI library made of 476 terms. The old version was created in 2013 and can now be found in the Legacy category for provenance. We also added a new library to the Crowd category. All the signatures in the...
Webgestalt: more up-to-date, many species

http://www.webgestalt.org/

Data Source

Data sources for WebGestalt 2017 was updated on 1/27/2017, which supports 12 organisms, 324 gene identifiers from various databases and technology platforms, and 150,937 functional categories from public databases and computational analyses. Information in this version was collected from the following resources:

- ID mapping

- Functional categories
  - Gene Ontology (Release Oct24, 2016)

  - Pathway
    - KEGG (Release October 1, 2016), Wikipathway (Release 05/10/2018), Reactome (Version 2016), PANTHER (Release 6/14/16)

  - Network
    - Hierarchical mRNA co-expression modules: The modules are computationally derived from the RNA-Seq data sets across 33 TCGA (The Cancer Genome Atlas, Release 01/28/2016) cancer types. Based on the correlation and co-expression analysis.
ToppGene: very up-to-date but human (mouse/rat) only

**ToppGene Suite**

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

### GO Biological Process

<table>
<thead>
<tr>
<th></th>
<th>Human Phenotype</th>
<th>Mouse Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotations:</td>
<td>15.171</td>
<td>8.400</td>
</tr>
<tr>
<td>Genes:</td>
<td>18.623</td>
<td>4.707</td>
</tr>
</tbody>
</table>

Updated 11-feb-2018

### GO Cellular Component

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotations:</td>
<td>1.801</td>
<td></td>
</tr>
<tr>
<td>Genes:</td>
<td>19.081</td>
<td></td>
</tr>
</tbody>
</table>

Updated 11-feb-2018

### GO Molecular Function

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotations:</td>
<td>4.961</td>
<td></td>
</tr>
<tr>
<td>Genes:</td>
<td>18.661</td>
<td></td>
</tr>
</tbody>
</table>

Updated 11-feb-2018

Gene symbols of human, mouse and rat orthologs are the same

So you can use it for mouse too but analysis is based on data from human
ToppFun for functional enrichment analysis

ToppFun: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis

Select your gene identifier type, paste your sets below or select example set, then submit.

Entry Type: HGNC Symbol
Example gene sets:
- HGNC Symbol
- HGNC Symbol and Synonyms
- Entrez ID
Training Gene Set:
- Ensembl ID
  - RefSeq
  - Uniprot
  - ENSMUSG00000069833
  - ENSMUSG00000033763
  - ENSMUSG00000058254

Annotations:
Transcriptome (gene expression)
Proteome (protein domains and interactions)
Regulome (TFBS and miRNA), Ontologies (GO, Pathway),
Phenotype (human disease and mouse phenotype)
Pharmacome (Drug–Gene associations)
Bibliome (literature co-citation)

Tutorial: https://toppgene.cchmc.org/help/supplimental.jsp#ToppFun
ToppFun functional enrichment analysis output

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR B&amp;H</th>
<th>FDR B&amp;Y</th>
<th>Bonferroni</th>
<th>Genes from Input</th>
<th>Genes in Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0005925 focal adhesion</td>
<td>7.26E-6</td>
<td>4.41E-4</td>
<td>2.49E-3</td>
<td>1.14E-3</td>
<td>393</td>
<td>6</td>
<td>393</td>
</tr>
<tr>
<td>2</td>
<td>GO:0005924 cell-substrate adherens junction</td>
<td>7.81E-6</td>
<td>4.41E-4</td>
<td>2.49E-3</td>
<td>1.23E-3</td>
<td>398</td>
<td>6</td>
<td>398</td>
</tr>
<tr>
<td>3</td>
<td>GO:0030055 cell-substrate junction</td>
<td>8.38E-6</td>
<td>4.41E-4</td>
<td>2.49E-3</td>
<td>1.32E-3</td>
<td>403</td>
<td>6</td>
<td>403</td>
</tr>
<tr>
<td>4</td>
<td>GO:0005912 adherens junction</td>
<td>2.37E-5</td>
<td>9.30E-4</td>
<td>5.24E-3</td>
<td>3.74E-3</td>
<td>484</td>
<td>6</td>
<td>484</td>
</tr>
<tr>
<td>5</td>
<td>GO:0070161 anchoring junction</td>
<td>2.94E-5</td>
<td>9.30E-4</td>
<td>5.24E-3</td>
<td>4.65E-3</td>
<td>503</td>
<td>6</td>
<td>503</td>
</tr>
</tbody>
</table>

Show 5 more annotations

number of genes in list that were linked to this annotation
number of genes in genome that are linked to this annotation
report these p-values
You can revert the question: Gene set enrichment analysis

Determines if an *a priori* defined set of genes shows statistically significant concordant differences between two biological sources

Instead of starting from a set of DE genes and looking for overrepresented categories, you start from a set of genes belonging to a category and see if they are DE.
Tools for gene set enrichment analysis


Can do both types of analysis

*Selection of enrichment analysis method:*

- Over- / Under-representation Analysis
- Gene Set Enrichment Analysis

Tools for network analysis of gene lists: IPA/Cytoscape

http://www.cytoscape.org/

Free, many apps for additional analyses in Cytoscape: GeneMania, iRegulon

Tutorials: http://opentutorials.cgl.ucsf.edu/index.php/Portal:Cytoscape3

Training: on demand

VIB scientists can use Ingenuity Pathway Analysis for free!


Heavily curated, high quality

Human, mouse, rat

Custom training: bits@vib.be

ModuleBlast to compare networks between organisms

http://www.expression.cs.cmu.edu/module.html
Where do variant data come from?

Small and large sequencing projects e.g.

**NHLBI Exome Sequencing Project:** Sequencing of 6500 human exomes from Americans
http://evs.gs.washington.edu/EVS/

**1000 Genomes Project:** Sequencing of 2504 human genomes from 26 populations
Data can be viewed in Ensembl or NCBI 1000 genomes browser
Data can be downloaded at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

**GENOMEOFTHENETHERLANDS:** Sequencing of 250 Dutch trios
http://www.nlgenome.nl/

...
Sources of variants in variant databases?

**dbSNP**

Scientific community. Individuals, small and large sequencing projects

No curation (~ Genbank)

Any organism

**Ensembl**

Various databases among others dbSNP

Many organisms

[http://www.ensembl.org/info/genome/variation/sources_documentation.html](http://www.ensembl.org/info/genome/variation/sources_documentation.html)
NCBI variation databases


53 organisms

Data is submitted by scientific community (~ Genbank)

Quality is questionable

**dbSNP**

Single-base nucleotide substitutions (SNPs)

**Small** multi-base deletions or insertions (deletion insertion polymorphisms or DIPs)

Retroposon insertions and microsatellite repeat variations (short tandem repeats or STRs)

**dbVar**

Genomic structural variations = insertions / deletions in DNA regions > 1 kb:

Inversions

Translocations

Genomic imbalances: copy number variants (more or less than 2 copies of a region)
Validation status tells you where the data come from

Reference SNP (refSNP) Cluster Report: rs15869

<table>
<thead>
<tr>
<th>RefSNP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism:</td>
<td>human (Homo sapiens)</td>
</tr>
<tr>
<td>Molecule Type:</td>
<td>Genomic</td>
</tr>
<tr>
<td>Created/Updated in build:</td>
<td>52/147</td>
</tr>
<tr>
<td>Map to Genome Build:</td>
<td>107/Weight 1</td>
</tr>
<tr>
<td>Validation Status:</td>
<td>[Validation Status Icon]</td>
</tr>
<tr>
<td>Citation:</td>
<td>PubMed</td>
</tr>
</tbody>
</table>

Validation status description:
- Validated by multiple, independent submissions to the refSNP cluster
- Validated by frequency or genotype data: minor alleles observed in at least two chromosomes
- Validated by submitter confirmation
- All alleles have been observed in at least two chromosomes apiece
- Genotyped by HapMap project
- SNP has been sequenced in 1000Genome project
- Suspect SNPs: SNP suspected from paralogous region (PMID: 21030649). Added to dbSNP on 01/21/2011.
Searching dbSNP can be done based on

SNP-ID (rsXXXXXX), gene name, keywords

Genomic location

SNP Advanced Search Builder

(X[Chromosome]) AND 2000:3000[Base Position]) AND Drosophila[Organism]
NCBI variation tools


Which tool do I use for .... ?

1. I have 3000 variant calls. How do I find out which is possibly significant? - Use Variation Reporter.
3. I want to know if my variant has been observed in the 1000 Genomes study. - Use the 1000 Genomes Browser.
4. I have a VCF file in GRCh37 coordinates. How do I move those to GRCh38 coordinates? - Use the NCBI Genome Remapping Service.
5. I want to search the NHGRI GWAS Catalog for studies related to Basal Cell Carcinoma. - Use PheGeni.
Sequence variation in Ensembl

Variation pages

SNPs: many source databases for many species

http://www.ensembl.org/info/genome/variation/sources_documentation.html

<table>
<thead>
<tr>
<th>Source</th>
<th>Version</th>
<th>Description</th>
<th>Data type(s)</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbSNP</td>
<td>142</td>
<td>Variants (including SNPs and indels) imported from dbSNP</td>
<td>Variation 12 million+</td>
<td></td>
</tr>
<tr>
<td>Archive dbSNP</td>
<td>142</td>
<td>Former variants names imported from dbSNP</td>
<td>Synonym 10 million+</td>
<td></td>
</tr>
<tr>
<td>AMDGC</td>
<td>1</td>
<td>The AMD Gene consortium is an international collaboration that seeks to</td>
<td>Phenotype Study 3,000+</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>identify genetic loci associated with age-related macular degeneration.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClinVar</td>
<td>06/03/2015</td>
<td>Variants of clinical significance imported from ClinVar</td>
<td>Variation 1,000+</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Phenotype 112,000+</td>
<td></td>
</tr>
<tr>
<td>COSMIC</td>
<td>71</td>
<td>Somatic mutations found in human cancers from the COSMIC project (mapped to</td>
<td>Variation 2.1 million+</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GRCh38)</td>
<td>Phenotype 2.2 million+</td>
<td></td>
</tr>
<tr>
<td>dbGaP</td>
<td>05/2014</td>
<td>The database of Genotypes and Phenotypes.</td>
<td>Phenotype Study 35,000+</td>
<td></td>
</tr>
<tr>
<td>DDG2P</td>
<td>18/11/2014</td>
<td>Developmental Disorders Genotype-to-Phenotype Database</td>
<td>Phenotype 2,000+</td>
<td></td>
</tr>
<tr>
<td>DGVa</td>
<td>01/2015</td>
<td>Database of Genomic Variants Archive</td>
<td>SV 3.2 million+</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Phenotype 8,010+</td>
<td></td>
</tr>
</tbody>
</table>
# Sequence variation in Ensembl

## Sequence variants

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example (Reference / Alternative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
<td>Ref: \ldots TTG\text{ACGTA}\ldots Alt: \ldots TTGG\text{CGTA}\ldots</td>
</tr>
<tr>
<td>Insertion</td>
<td>Insertion of one or several nucleotides</td>
<td>Ref: \ldots TTG\text{ACGTA}\ldots Alt: \ldots TTG\text{ATGCGTA}\ldots</td>
</tr>
<tr>
<td>Deletion</td>
<td>Deletion of one or several nucleotides</td>
<td>Ref: \ldots TTG\text{ACGTA}\ldots Alt: \ldots TTGG\text{GTA}\ldots</td>
</tr>
<tr>
<td>Indel</td>
<td>An insertion and a deletion, affecting 2 or more nucleotides</td>
<td>Ref: \ldots TTG\text{ACGTA}\ldots Alt: \ldots TTG\text{GCTCGTA}\ldots</td>
</tr>
<tr>
<td>Substitution</td>
<td>A sequence alteration where the length of the change in the variant is the same as that of the reference.</td>
<td>Ref: \ldots TTG\text{ACGTA}\ldots Alt: \ldots TTG\text{TAGTA}\ldots</td>
</tr>
</tbody>
</table>

## Structural variants

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example (Reference / Alternative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV</td>
<td>Copy Number Variation: increases or decreases the copy number of a given region</td>
<td>Reference:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inversion</td>
<td>A continuous nucleotide sequence is inverted in the same position</td>
<td>Reference:</td>
</tr>
<tr>
<td>Translocation</td>
<td>A region of nucleotide sequence that has translocated to a new position</td>
<td>Reference:</td>
</tr>
</tbody>
</table>

http://www.ensembl.org/info/genome/variation/index.html
Ensembl Assembly converter

http://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter?db=core

Assembly Converter

New Assembly Converter job:

This online tool currently uses CrossMap, which supports a limited number of formats (see our online documentation for details of the individual data formats listed below). etc, will be lost on conversion.

Important note: CrossMap converts WIG files to BedGraph internally for efficiency, and also outputs them in BedGraph format.

Input

Species: Human (Homo sapiens)

Assembly mapping: GRCh37 -> GRCh38

Name for this data (optional):

Input file format: BED

Either paste data:

Or upload file: Bladeren... Geen bestand geselecteerd.

Or provide file URL:

Run  Reset
Why SNPs are important

What is an SNP?
Different people can have a different nucleotide or base at a given location on a chromosome.

What is an SNP map?
Location of SNPs on human DNA

How can an SNP map be used to predict medicine response?

Section of SNP genotype profile

- Patients with efficacy in clinical trials
- Patients without efficacy in clinical trials
- Predictive of efficacy
- Predictive of no efficacy
Specific sequence variation databases

Database of Genomic Variants

http://dgv.tcag.ca/dgv/app/home

Curated database of structural variations (> 50 bp) in healthy humans

Can be used as control data

6500 individuals is not an accurate representation of the population


Kaviar

http://db.systemsbiology.net/kaviar/

Kaviar contains 162 million SNV sites (including 25M not in dbSNP) from 35 sources

http://db.systemsbiology.net/kaviar/cgi-pub/Kaviar.pl?show=sources

To check if your variants are novel or known

Specific sequence variation databases

Genome Variation Server

http://gvs.gs.washington.edu/

Easy combination of dbSNP, HapMap, Ensembl... data

Human only

<table>
<thead>
<tr>
<th>Gene Name: BRCA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene ID: 675</td>
</tr>
<tr>
<td>Chromosome 13: 32315480 - 32399672 (+)</td>
</tr>
<tr>
<td>Allele Frequency Cutoff (%): 0, monomorphic sites excluded</td>
</tr>
<tr>
<td>Data Merging: combined samples with common variations</td>
</tr>
<tr>
<td>Population: PDR90, Submitter: EGP_SNPS</td>
</tr>
<tr>
<td>Population: ENSEMBL_Watson, Submitter: ENSEMBL</td>
</tr>
<tr>
<td>Population: HapMap-HCB, Submitter: CSHL-HAPMAP</td>
</tr>
</tbody>
</table>

Combined list over 3 data sets

<table>
<thead>
<tr>
<th>SNP</th>
<th>Sample</th>
<th>Allele1</th>
<th>Allele2</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs206118</td>
<td>693:1208</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>rs206118</td>
<td>693:1209</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>rs206118</td>
<td>693:1210</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>rs206118</td>
<td>693:1211</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>rs206118</td>
<td>693:1212</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Databases of mutations linked to cancer

BioMart Cancer portal (see module 5)

COSMIC: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

My Cancer genome: http://www.mycancergenome.org/
Tools for studying mutations linked to cancer

CGWB:  https://cgwb.nci.nih.gov/

  UCSC genome browser with custom tracks for cancer genetics

UCSC cancer genomics tool: http://xena.ucsc.edu/

  Find expression patterns between genes in cancer samples
  Survival analysis: which genes affect survival in cancer?
  Compare your sample to public cancer samples
OMIM database: mutations linked to genetic disorders

Online Mendelian Inheritance in Man
http://www.omim.org

**Single gene** genetic disorders in human

Link gene – phenotype

<table>
<thead>
<tr>
<th>Location</th>
<th>Phenotype</th>
<th>Phenotype MIM number</th>
<th>Inheritance</th>
<th>Phenotype mapping key</th>
<th>Gene/Locus</th>
<th>Gene/Locus MIM number</th>
</tr>
</thead>
<tbody>
<tr>
<td>22q11.23</td>
<td>Opitz GBBB syndrome, type II</td>
<td>145410</td>
<td>AD</td>
<td>3</td>
<td>SPECC1L</td>
<td>614140</td>
</tr>
</tbody>
</table>

A number sign (#) is used with this entry because of evidence that autosomal dominant Opitz GBBB syndrome (GBBB2) is caused by heterozygous mutation in the SPECC1L gene (614140) on chromosome 22q11.23.

wANNOVAR: webtool for functional annotation of variants

http://wannovar.wglab.org/: one vcf file from one sample

Web version of popular command line tool ANNOVAR

Input: VCF file of variants

cromosome, position, reference nucleotide, observed nucleotide

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>871334</td>
<td>G</td>
<td>T</td>
<td>328.93...</td>
</tr>
<tr>
<td>1</td>
<td>881918</td>
<td>G</td>
<td>A</td>
<td>861.50</td>
</tr>
</tbody>
</table>

ANNOVAR: Select variants with **functional consequences**: AA change, stop, frameshift

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>Ref</th>
<th>Alt</th>
<th>Func</th>
<th>Gene</th>
<th>ExonicFunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>871334</td>
<td>G</td>
<td>T</td>
<td>intronic</td>
<td>SAMD11</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>881918</td>
<td>G</td>
<td>A</td>
<td>exonic</td>
<td>NOC2L</td>
<td>nonsynonymous SNV</td>
</tr>
</tbody>
</table>
wANNOVAR: webtool for functional annotation of variants

ANNOVAR Check if variant is already in a database and remove frequent variants

PHENOLYZER From this filtered list find genes that are linked to the phenotype
Based on annotations from OMIM, HPO, GWAS_Catalog, ClinVar...

Output: small list of candidate genes (and their variants) for further study
SNPEff and SNPSift: tools for functional annotation of variants

http://snpeff.sourceforge.net/index.html: combine vcf files from multiple samples from multiple groups (healthy – disease)

Command line tools

Major research institutes uses it: http://snpeff.sourceforge.net/SnpEff.html#who
PDB is the main 3D structure repository

http://www.pdb.org

Database of **3-dimensional structures** of biomolecules:

- Protein
- DNA
- RNA
- Ligands

Obligatory deposit of coordinates in the PDB before publication

Training: on demand

Joost Van Durme, SWITCH Lab
3D structure info is limited

PDB contains 120000 structures but high level of redundancy (mutants)
Where do the data come from?

**X-ray crystallography** (crystals)

**Nuclear Magnetic Resonance** (NMR) (in solution)

No purely theoretical or ab initio models since 2006
X-ray crystallography to solve 3D structures

You measure electron densities:

You observe regions where many electrons are located = positions of atoms

You also measure the uncertainty on the positions of the atoms
Interpretation of X-ray model results

Main results: x, y, z coordinates of the mean atom positions

Disorder around this mean: B-factor and occupancy

Variations in time: molecules vibrate

Variations in space: crystal contains 1000s molecules: not all completely equal

B-factor reflects the uncertainty of the mean atom position

Higher B-factor means more uncertainty about position

Occupancy reflects if we should consider alternative conformations of a side chain

How often do we find a side chain in one or another conformation?
Amine Terminal End

Back bone

Side chains

gly  ala  leu

C. Ophardt, c. 2003
B-factor reflects the certainty of an atom’s position

Atoms with large B-factors have large positional uncertainty

Indication of mobility of an atom

- \(0 < B < 20\): position is most likely OK
- \(20 < B < 40\): position is probably OK, but errors up to 0.5 Ångstrom are possible
- \(40 < B < 60\): position is reasonably OK, but errors up to 1.0 Ångstrom are possible
- \(B > 60\): atom is not likely to be within 1.0 Ångstrom from where you see it
- \(B\) around 100: atom is guaranteed not within 1.0 Ångstrom from where you see it
PDB records show all these factors

<table>
<thead>
<tr>
<th>info concerning atom</th>
<th>amino acid atom belongs to</th>
<th>position in protein sequence</th>
<th>occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr &amp; type of atom</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>ATOM 625 C</td>
<td>ILE A</td>
<td>77 -11.322 28.374 -1.179</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM 626 O</td>
<td>ILE A</td>
<td>77 -11.946 29.453 -1.112</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM 627 CA A</td>
<td>ILE A</td>
<td>77 -11.432 27.329 -0.087</td>
<td>0.70</td>
</tr>
<tr>
<td>ATOM 628 CB A</td>
<td>ILE A</td>
<td>77 -12.918 26.874 0.087</td>
<td>0.70</td>
</tr>
<tr>
<td>ATOM 629 CG1A</td>
<td>ILE A</td>
<td>77 -13.042 25.758 1.141</td>
<td>0.70</td>
</tr>
<tr>
<td>ATOM 630 CG2A</td>
<td>ILE A</td>
<td>77 -13.516 26.421 -1.241</td>
<td>0.70</td>
</tr>
<tr>
<td>ATOM 631 CD1A</td>
<td>ILE A</td>
<td>77 -13.378 26.302 2.501</td>
<td>0.70</td>
</tr>
<tr>
<td>ATOM 632 CA B</td>
<td>ILE A</td>
<td>77 -11.423 27.327 -0.082</td>
<td>0.30</td>
</tr>
<tr>
<td>ATOM 633 CB B</td>
<td>ILE A</td>
<td>77 -12.874 26.775 0.117</td>
<td>0.30</td>
</tr>
<tr>
<td>ATOM 634 CG1B</td>
<td>ILE A</td>
<td>77 -13.519 26.423 -1.227</td>
<td>0.30</td>
</tr>
<tr>
<td>ATOM 635 CG2B</td>
<td>ILE A</td>
<td>77 -13.748 27.739 0.916</td>
<td>0.30</td>
</tr>
<tr>
<td>ATOM 636 CD1B</td>
<td>ILE A</td>
<td>77 -14.720 25.518 -1.100</td>
<td>0.30</td>
</tr>
<tr>
<td>ATOM 637 N</td>
<td>ARG A</td>
<td>78 -10.521 28.048 -2.183</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM 638 CA</td>
<td>ARG A</td>
<td>78 -10.258 28.952 -3.268</td>
<td>1.00</td>
</tr>
<tr>
<td>ATOM 639 C</td>
<td>ARG A</td>
<td>78 -10.857 28.469 -4.584</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Naming of Cs in amino acids

isoleucine

\[
\begin{align*}
    &H \quad H \\
    &\text{N} - C - C - O \\
    &\text{H} \quad \text{B} \\
    &\text{H}_3 \text{C} \\
    &\text{G2} \\
    &\text{G1} \\
    &\text{CH}_2 \\
    &\text{D} \\
    &\text{CH}_3
\end{align*}
\]
You can search PDB using keywords or sequence.
Easy access to PDB via UniProt

Which PDB record do you choose? X-ray over NMR

Low **resolution** (how detailed is the structure)

Which **portion** of the structure is represented?
PDBe gives a detailed report of the quality of the structure

[http://www.ebi.ac.uk/pdbe/](http://www.ebi.ac.uk/pdbe/)

European version of PDB

Also include structures obtained by cryo EM

4ezx › Experiments and Validation

**X-ray diffraction**

*Source organism: Escherichia coli K-12*

- **Resolution:** 1.7 Å
- **Reported R values:**
  - \( R \): 0.17
  - \( R_{\text{free}} \): 0.17

**R:** quality of the model (the lower the better)

**R-free:** unbiased quality of the model
How to interpret the quality report?

Solid sliders: how a given structure ranks relative to all structures in PDB
Open sliders: comparison with structures derived in a similar fashion

X-ray compared with other X-ray structures solved at a similar resolution
PDB-REDO improves the structures from PDB

[https://pdb-redo.eu/](https://pdb-redo.eu/)

Automatically re-refined deposited structures, using the latest methods

<table>
<thead>
<tr>
<th>Validation metrics from PDB-REDO</th>
<th>PDB</th>
<th>PDB-REDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystallographic refinement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>0.1712</td>
<td>0.1680</td>
</tr>
<tr>
<td>$R$-free</td>
<td>0.1996</td>
<td>0.1961</td>
</tr>
<tr>
<td>Bond length RMS Z-score</td>
<td>1.040</td>
<td>0.776</td>
</tr>
<tr>
<td>Bond angle RMS Z-score</td>
<td>1.109</td>
<td>0.879</td>
</tr>
</tbody>
</table>

Z-score: # SDs from mean or 'ideal' value

Many highly accurate small molecule structures contain small peptides

They are used to calculate standard bond lengths and angles in proteins


[http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/explain.html](http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/explain.html)
PDB-REDO improves the structures from PDB

<table>
<thead>
<tr>
<th>Model quality (raw scores</th>
<th>percentiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramachandran plot appearance</td>
<td>97</td>
</tr>
<tr>
<td>Rotamer normality</td>
<td>53</td>
</tr>
<tr>
<td>Coarse packing</td>
<td>N/A</td>
</tr>
<tr>
<td>Fine packing</td>
<td>97</td>
</tr>
<tr>
<td>Bump severity</td>
<td>93</td>
</tr>
</tbody>
</table>

Ramachandran plot: the first serious verification tool available for protein structures

Bumps: Atoms that overlap (impossible)

Percentiles: compared to similar structures with similar resolution

the higher the better
Other useful websites related to protein structures

EBI: http://www.ebi.ac.uk/Tools/structure/

MolProbity: http://molprobity.biochem.duke.edu/
Assess quality of protein structures (bond lengths, angles, Ramachandran)

NMR Constraint Analyser: http://molsim.sci.univr.it/bioinfo/tools/constraint/
Assess quality of NMR derived structures

Nucleic Acid Database: http://ndbserver.rutgers.edu/
DNA and RNA structures
You can protein structure images from PDB/PDBe/PDB redo.
Click an Image to download
Tools for visualization of PDB structures

1. YASARA: http://www.yasara.org

2. SwissPDBViewer: http://spdbv.vital-it.ch/

3. Chimera: http://www.cgl.ucsf.edu/chimera
tutorial: http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/

4. Pymol: http://www.pymol.org

High quality output

Good examples on wiki: http://www.pymolwiki.org

...
Protein folds are the structures of domains

Similarities in assembly of secondary structure elements
So not based on sequence like motifs but on 3D structure

Examples:
- alpha solenoid
- DNA clamp
- Thioredoxin fold

Chlorophyll proteins
Nuclear pores

Binds DNA polymerase to DNA

Enzymes that catalyze S-S bond formation

Folds represent the shapes of protein domains!
Databases of protein folds

SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/

CATH: http://www.cathdb.info/

Manually curated
Protein structure prediction

Homology modeling
Predict structure based on experimental 3D structure of homolog
Quality of alignment is crucial

Fold recognition
Predict structure based on experimental 3D structure of protein with same fold
Also works when sequence similarity is low

Ab initio prediction
Predict structure based on sequence
Only works for small proteins, requires massive computing power
Rosetta project: http://boinc.bakerlab.org/rosetta/
Predict protein structures by homology modeling

Similarity on sequence level projected to a structure

1. Search for a homolog with known 3D structure
BlastP your query **sequence** against **sequences** from PDB


OR use PSI-BLAST to detect sequences with similar structures

DELTA-BLAST

2. Make alignment

3. Model the structure of your protein based on the known 3D structure of homolog
YASARA/FoldX: see wiki for tutorial

https://swissmodel.expasy.org/: see wiki for tutorial

http://www.ii.uib.no/~slars/bioinfocourse/PDFs/structpred_tutorial.pdf
Predict protein structures by fold recognition

Similarity on fold level projected to a structure

1. Search SCOP/CATH for protein with same fold and known 3D structure

Search CATH

Find folds that are compatible with the sequence: not based on sequence similarity
Predict protein structures by fold recognition

2. Align each amino acid of query sequence to a position in the template structure

Evaluate how well the sequence fits the fold and select best-fit fold

Build structural model of query based on alignment with selected fold


HHpred: [http://toolkit.lmb.uni-muenchen.de/hhpred](http://toolkit.lmb.uni-muenchen.de/hhpred)


Works because:

Number of different folds in nature is fairly small (approximately 1300)

90% of new submissions in PDB have similar folds to those already in PDB

Not always accurate
Guidelines to improve fold recognition results

- Run as many methods as you can
- Run each method on many sequences from your homologous protein family
- After all of these runs, build up a consensus picture of the likely fold
- Compare function of your protein to function of the proteins with the likely fold
- Compare secondary structure of your protein to that of the likely fold
Similarity searches based on 3D structure

Similarity on structural level: aligning 3D structures

Structure of query protein is known and aligned to PDB structures

VAST+

DALI
http://ekhidna.biocenter.helsinki.fi/dali_server/

Compare proteins with low sequence similarity:

  similar structure implies homology -> same function

Can help to find active sites