BIOINFORMATIK II
Übung 3

Dietmar Rieder

http://icbi.at/bioinf2

Übung III

Einführung

GO Klassifikation / Netzwerkanalyse

Gene Ontology

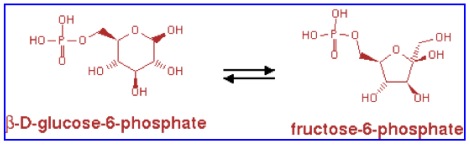
The* Gene Ontology *(GO) is a controlled vocabulary that can be applied to all organisms even as knowledge of* gene *and protein roles in cells is accumulating and changing.

GO is…

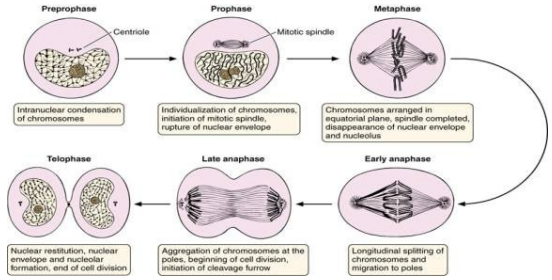
- A collaborative effort to address the need for consistent descriptions of **gene** products in different databases
- Three structured, controlled vocabularies (ontologies) that describe **gene** products in a species-independent manner

GO Categories

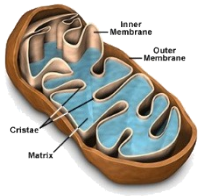
- **Molecular Function** Ontology
  - the tasks performed by individual gene products; examples are carbohydrate binding and ATPase activity



- **Biological Process** Ontology
  - broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions



- **Cellular Component** Ontology
  - subcellular structures, locations, and macromolecular complexes; examples include mitochondria, nucleus, telomere, and origin recognition complex



GO annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as ‘gene associations’ or GO annotations
  - Multiple annotations per gene
- Some GO annotations are created automatically

GO Software Tools

- GO resources are freely available to anyone without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Different groups have used GO to create tools for many purposes
  - http://www.geneontology.org/
  - BiNGO plugin for Cytoscape
  - DAVID (**D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery)

GO Software Tools

BiNGO plugin for Cytoscape

- Determine which Gene Ontology (GO) categories are statistically overrepresented in a set of genes or a subgraph of a biological network.
- Maps the predominant functional themes of a given gene set on the GO hierarchy, and outputs this mapping as a Cytoscape graph

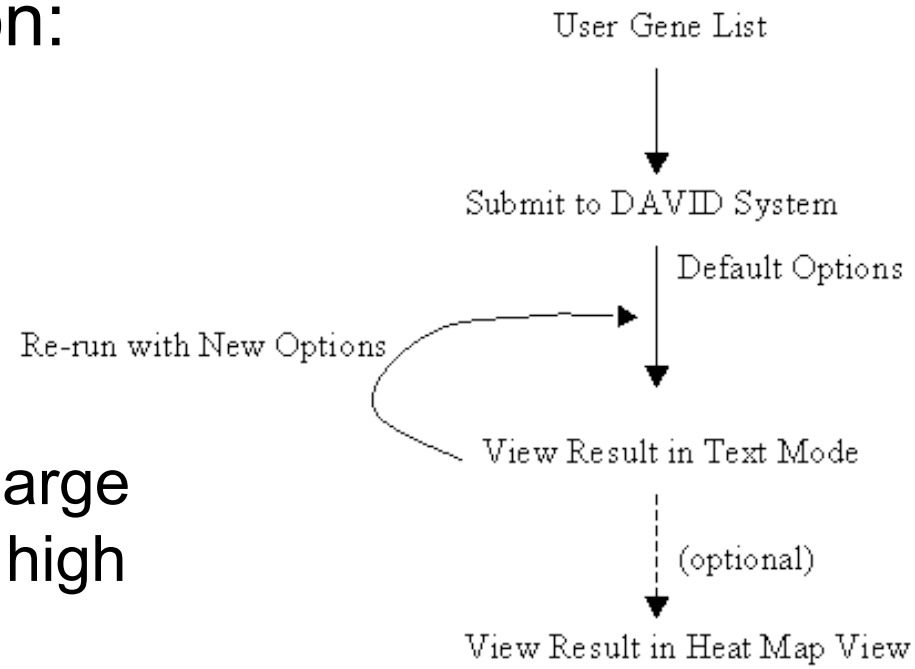
DAVID - http://david.abcc.ncifcrf.gov

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Can do many other tasks

DAVID

- Functional Classification:

Grouping genes based on functional similarity can systematically enhance biological interpretation of large lists of genes derived from high throughput studies.



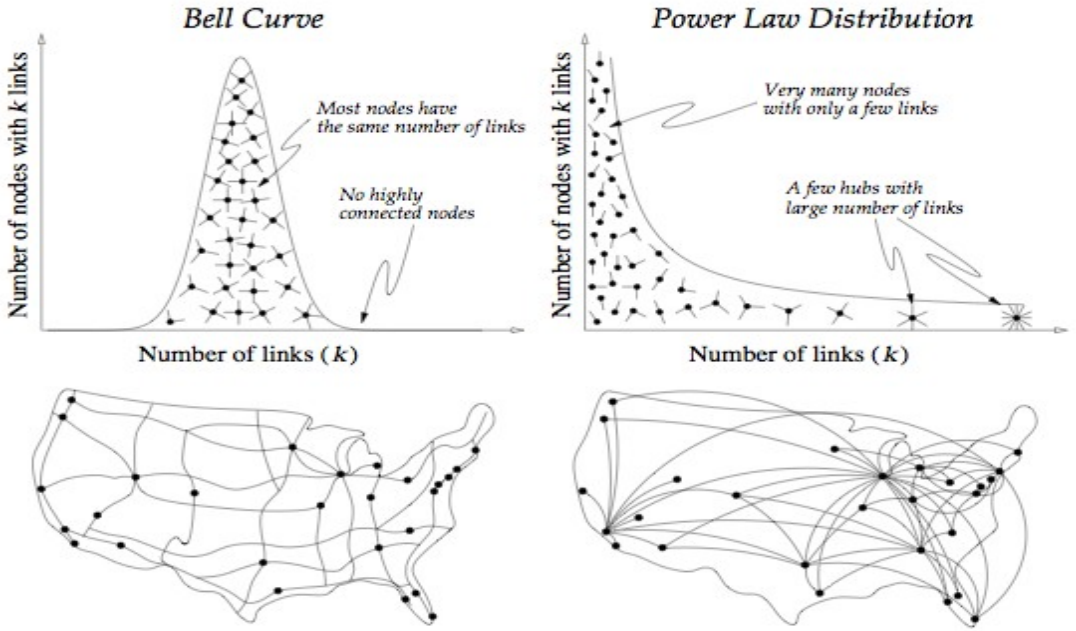
DAVID

- **Functional Annotation Tool**

  - provides typical batch annotation
  - gene-GO term enrichment analysis
  - highlight the most relevant GO terms in a gene list
  - extended annotation content coverage:
    - currently over 40 annotation categories, including GO terms, protein-protein interactions, protein functional domains, disease associations, bio-pathways, sequence general features, homologies, gene functional summaries, gene tissue expressions, literatures, etc.

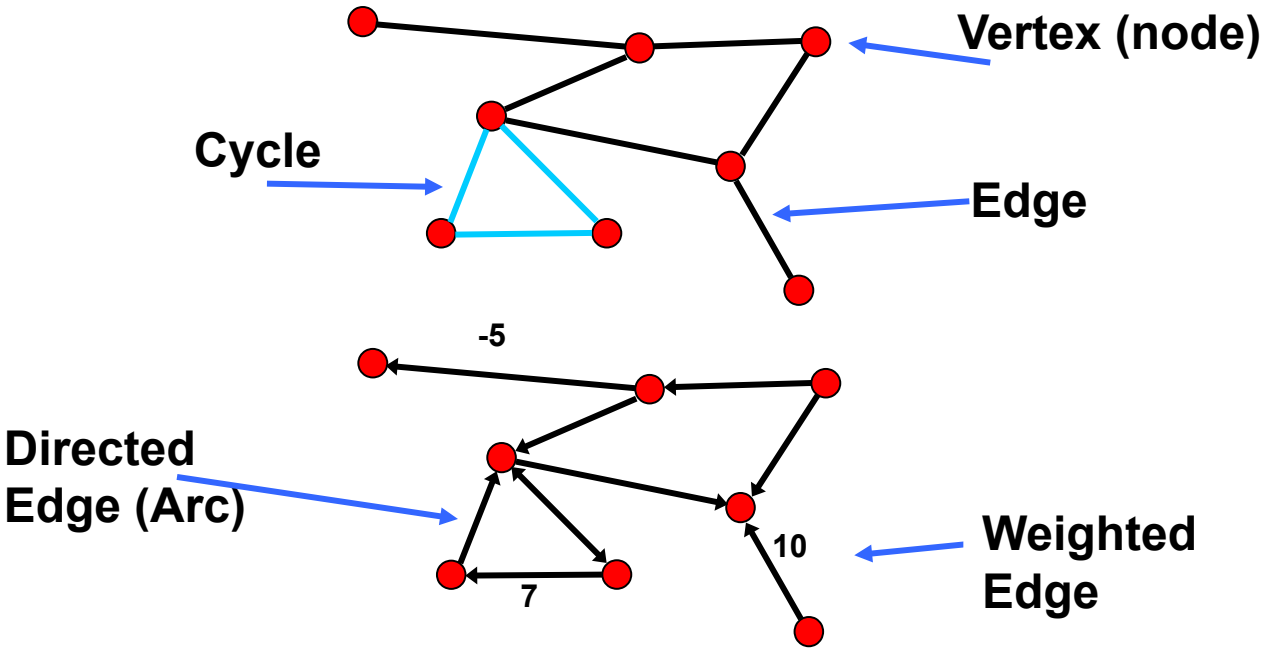
Biomolecular Networks

Scale-free networks



Networks

- E.g. Protein-protein interaction networks
- Useful if we don't know pathways
  - Could discover new pathways



Mapping Biology to a Network

- A simple mapping: Protein-protein interactions
  - one protein/node, one interaction/edge
- Edges can represent other relationships
  - Physical e.g. protein-protein interaction
  - Regulatory e.g. kinase activates target
  - Genetic e.g. epistasis
  - Similarity e.g. protein sequence similarity, expression profile similarity

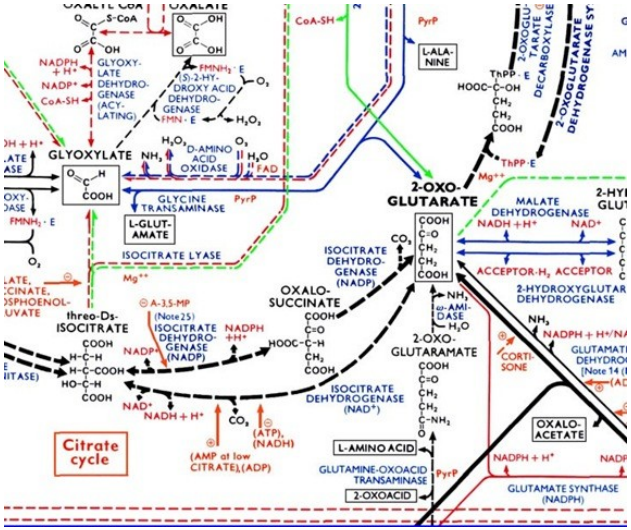
- Critical: understand the mapping for network analysis

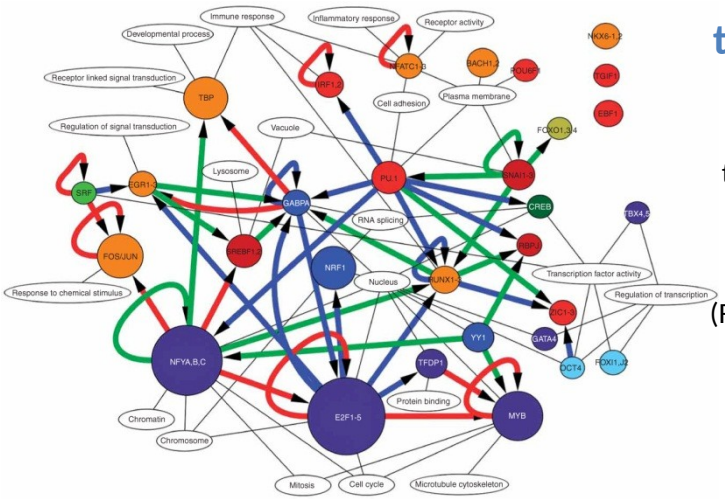
Examples for networks in Biology

metabolic

Nodes
metabolites and enzymes

Edges chemical reactions





transcriptional

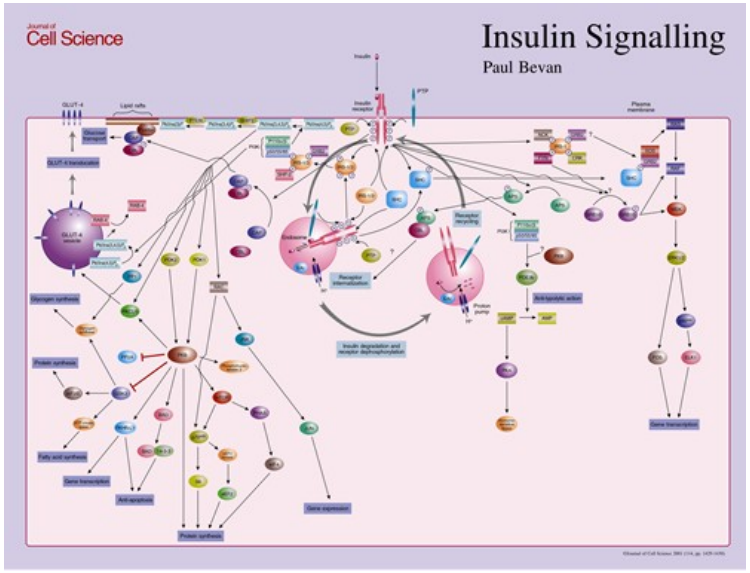
Nodes
genes and transcription factors

Edges
Directed
(Functional activation and inhibition)

signaling

Nodes
proteins

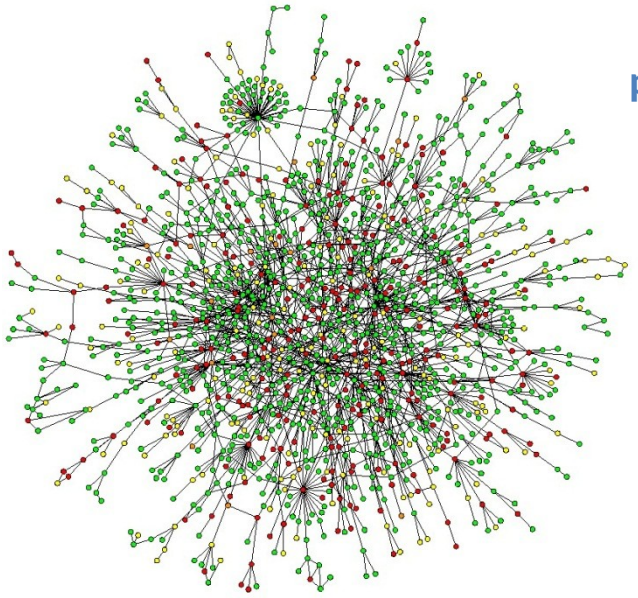
Edges
'physical' interactions (binding and post-translational modifications)



protein-protein Interaction PPI

Nodes
proteins

Edges
'physical' interactions



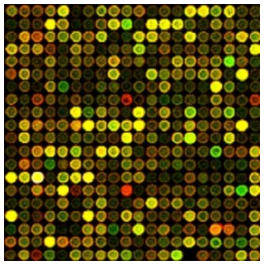
Why Pathway and Network Analysis?

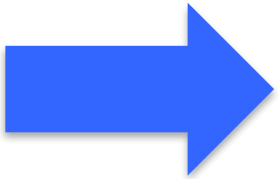
Intuitive to Biologists

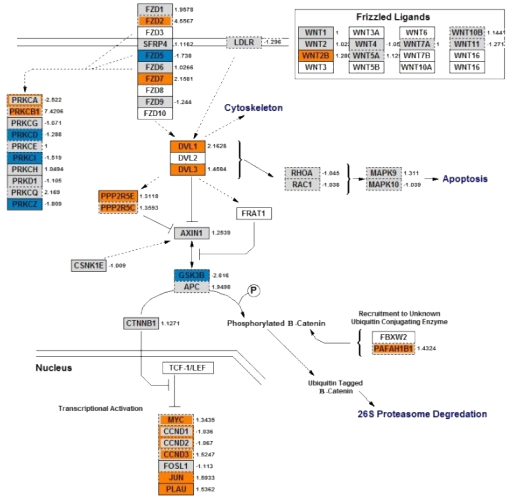
- Provide a biological context for results
- More efficient than searching databases gene-by-gene
- Intuitive display for sharing data

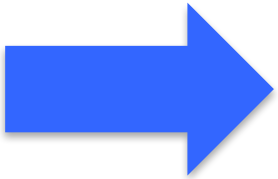
Computation on Pathway Content

- Visualize multiple data types on a pathway or network
- Find active pathways
- Identify potential regulators





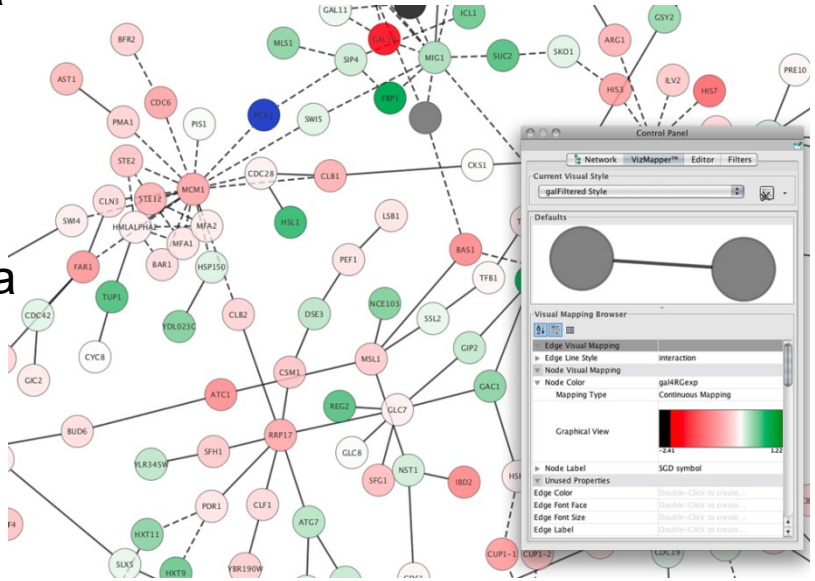




Cytoscape

Cytoscape supports many use cases in molecular and systems biology, genomics, and proteomics

- Load molecular and genetic interaction data
- Project and integrate global datasets and functional annotations
- Establish visual mappings across these data
- Perform analysis and modeling using plugins
- Visualize and analyze human-curated pathway datasets such as Reactome or KEGG.



Exercise III

1. Create a hypothetical protein-protein interaction network in Cytoscape.

2. Create, visualize and analyze a network form gene expression data. (Install ExpressionCorrelation plugin)

3. Use the Cytoscape BiNGO plugin to analyze gene expression data. (Install BiNGO plugin)

4. Map gene expression data to pathways the that you found.

5. Homework: Analyze gene expression data using DAVID GO tools.

DAVID: functional annotation summary

DAVID Documentation: Annotation Summary

The Annotation Summary presents annotation associated with the user's gene list within a compact, category-driven layout with interactive selection controls that allows navigation to Functional Annotation Table, Chart, or Clustering results for single or multiple annotation categories.

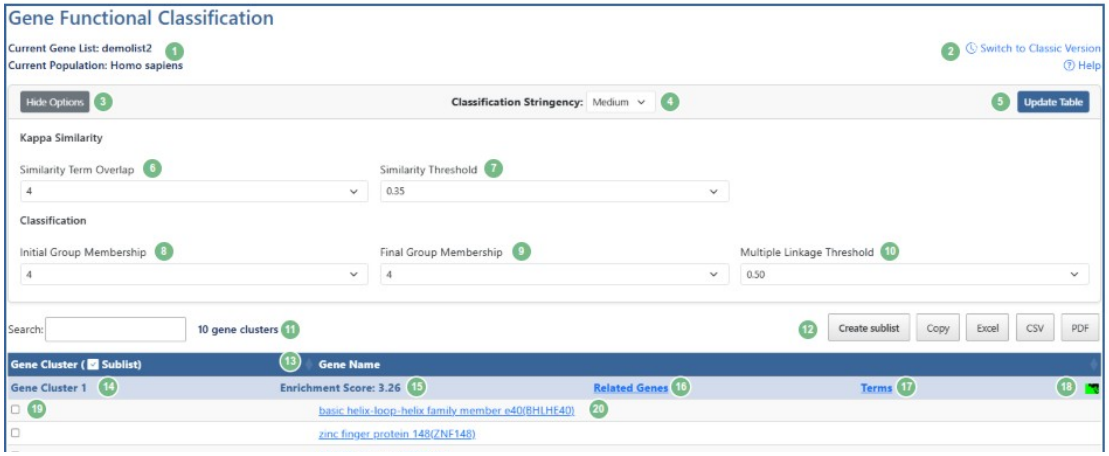
1. **Current Gene List & Population**
   Displays the active gene list and selected species or background population.
2. **Help and Switch to Classic Version**
   Link to this help page and provides access to the legacy Gene List Report interface.
3. **Categories**
   Click each category to expand and view additional annotation categories. Check/uncheck the items to add/remove them from your selection.
4. **Clear All**
   Click this button to clear all selections.
5. **Check Default**
   Click this button to undo any currect selections and revert to default selections.
6. **Functional Annotation Table**
   Click this button to view the **Functional Annotation Table** for all selections. This is a gene-centric view which lists the genes and their associated annotation terms. For more information, see the Functional Annotation Table documentation page.
7. **Functional Annotation Chart**
   Click this button to view the **Functional Annotation Chart** for all selections. Chart Report is an annotation term focused view which lists annotation terms and their associated genes under study. For more information, see the Functional Annotation Chart documentation page.
8. **Functional Annotation Clustering**
   Click this buttn to view the **Functional Annotation Clustering** for all selections. Functional Annotation Clustering integrates Kappa statistics to measure the degree of the common genes between two annotations. For more information, see the Functional Annotation Clustering documentation page.
9. **Gene count**
   Number of genes from the user list involved in this annotation category.
10. **Gene Percentage & Single Table**
    Percentage of genes from user list involved in this category out of total genes from user list.
    Click this percentage bar to be directed to a functional annotation table for user genes in this annotation category ONLY.
11. **Single Chart**
    Click this button to be directed to a functional annotation chart for user genes in this annotation category ONLY.
12. **Single Cluster**
    Click this button to be directed to a functional annotation cluster for user genes in this annotation category ONLY.

DAVID: functional annotation chart report

Chart Report is an annotation term focused view which lists annotation terms and their associated genes under study. To avoid over counting duplicated genes, the Fisher Exact statistics is calculated based on corresponding DAVID gene IDs by which all redundancies in original IDs are removed.

1. **Current Gene List & Population**
   Displays the active gene list and selected species or background population.
2. **Help and Switch to Classic Version**
   Link to this help page and provides access to the legacy Gene List Report interface.
3. **EASE Threshold**
   Minimum P-value of results to be displayed. Click to either type a numberical value or use the up/down arrows to increment by 0.01.
4. **Count Threshold**
   Minimum gene count of annotation result to be displayed. Click to either type a numberical value or use the up/down arrows to increment by 1.
5. **Update Table**
   After making adjustments to filtering options, click this button to update the table using the selected parameters.
6. **Search and Record Count**
   The search box filters table results in real time. The record count indicates total entries and the number of DAVID genes mapped.
7. **Create sublist, Export options & Column Visibility**
   Create a sublist by clicking on checkboxes to the left of desired results, then clicking the **Create sublist** button, where you will be prompted to name the new sublist. Sublists will then appear in the list manager as a new gene list.
   Results can be exported using Copy, Excel, CSV, or PDF. Exports reflect the current filtered and sorted view.
   Additional columns can be added to the current view using the **Column Visibility** drop-down button. Click on available items to add the column to the table.
8. **Sortable Columns**
   Click on any table column header to sort the table by that column in ascending or descending order.
9. **Sublist**
   Click on check boxes to select items for creating a sublist. (7)
10. **Categories**
    Categories of term results, indicating orginal source of information. (e.g. UP = uniprot). Click to sort by this value.
11. **Annotations**
    Annotations available for each item. If available, hyperlinks lead users to original resources for further details. Click to sort by this value.
12. **Related Terms**
    Click **RT** to be directed to a Functionally Related Terms search result for each term. For more information, see the Related Terms documentation page.
13. **Gene Count and Percentage**
    Count and percentage of genes involved in each term.
14. **P-value & Benjamini**
    EASE Score where smaller means more enriched. Benjamini in DAVID requests adjusted p-values by using the linear step-up method of Benjamini and Hochberg

DAVID: functional classification report



1. **Gene List**
   Current gene list selected for view and background species being analyzed.
2. **Help and Switch to Classic Version**
   Link to this help page and provides access to the legacy Gene List Report interface.
3. **Show/Hide Options**
   Click this button to expand or hide additional filtering options.
4. **Classification Stringency**
   Click to access a drop-down menu of stringency levels, then click desired level to select. A high-level single control to establish a set of detailed parameters involved in functional classification algorithms. In general, the higher stringency setting generates less functional groups with more tightly associated genes in each group, so that more genes will be unclustered.
5. **Update Table**
   After making adjustments to filtering options, click this button to update the table using the selected parameters.
6. **Similarity Term Overlap**
   Click to access a drop-down menu of overlap values. The minimum number of annotation terms overlapped between two genes in order to be qualified for kappa calculation. This parameter is to maintain necessary statistical power to make the kappa value more meaningful. The higher the value, the more meaningful the result is.
7. **Similarity Threshold**
   Click to access a drop-down menu of theshrold values. The minimum kappa value to be considered significant. A higher setting will lead to more genes going unclustered, which leads to a higher quality functional classification result with fewer groups and fewer gene members. Kappa value of 0.3 starts giving meaningful biology based on our genome-wide distribution study. Anything below 0.3 has a good chance to be noise.
8. **Initial Group Membership**
   Click to access a drop-down menu of membership values. The minimum gene number in a seeding group, which affects the minimum size of each functional group in the final cluster. In general, the lower value attempts to include more genes in functional groups, and may generate a lot of small size groups.
9. **Final Group Membership**
   Click to access a drop-down menu of membership values. The minimum gene number in one final group after a 'cleanup' procedure. In general, the lower value attempts to include more genes in functional groups and may generate a lot of small size groups. It cofunctions with previous parameters to control the minimum size of functional groups. If you are interested in functional groups containing only 2 or 3 genes, you need to set it to a very low value. Otherwise, the small group will not be displayed and the genes will go unclustered.
10. **Multiple Linkage Threshold**
    Click to access a drop-down menu of threshold values. This parameter controls how seeding groups merge with each other, i.e. two groups sharing the same gene members over the percentage will become one group. A higher percentage, in general, gives sharper separation (i.e. it generates more final functional groups with more tightly associated genes in each group). In addition, changing the parameter does not cause additional genes to go unclustered.
11. **Search & Cluster count**
    The search box filters all table results in real time and shows how many clusters are present in results.
12. **Create sublist, Export options & Column Visibility**
    Create a sublist by clicking on checkboxes to the left of desired results, then clicking the **Create sublist** button, where you will be prompted to name the new sublist. Sublists will then appear in the list manager as a new gene list.
    Results can be exported using Copy, Excel, CSV, or PDF. Exports reflect the current filtered and sorted view.
    Additional columns can be added to the current view using the **Column Visibility** drop-down button. Click on available items to add the column to the table.
13. **Sortable Columns**
    Click on any table column header to sort the table by that column in ascending or descending order.
14. **Cluster**
    A group of genes having similar biological meaning due to sharing similar terms.
15. **Enrichment Score**
    The overall enrichment score for the group based on the EASE scores of each gene. Higher value = more enriched.
16. **Related Genes (RG)**
    Click the **RG** link to explore functionally related genes associated with each entry. For more information, see the Related Genes documentation page.
17. **Terms**
    Click the **Terms** link to explore the Cluster Term Report for each gene. For more information, see the Cluster Term Report documentation page.
18. **Cluster view**
    Click this icon to be directed to a 2D cluster view of the genes in this cluster.
19. **Sublist**
    Click on check boxes to select items for creating a sublist. (12)
20. **Gene Name**
    Click the gene name link to explore the full gene report. For more information, see the Gene Report documentation page.