SIDEKICK: Genomic data driven analysis and decision making framework

Mark Doderer
PhD Candidate
UTSA Department of Computer Science

Sidekick from Merriam-Webster Online

Pronunciation: \’sīd-kik\

Function: *noun*

: a person closely associated with another as a subordinate or partner

Goals / Background

- There is a need to bridge the gap between laboratory and computational scientists
- Research questions
  - Are there other mechanisms for disease X?
  - Does this set of genes associated with disease X also influence other diseases?
- Sidekick can be an asset to the scientist who is formulating these questions

Outline

- Goals / Background
- Modules
- Theories Enabling the Modules
- Demonstration
- Questions

Goals / Background

Typical computational tasks when answering research questions:

- Evaluate and search literature and databases
- Compile lists / matching identifiers possibly in different formats
- Format lists for use in other analysis tasks
- Combine evidence from different sources and with different scoring mechanisms

Goals / Background

InterologueFinder goals:

- Use known interologues in one species to predict interactions in other species
- Within a species compile evidence for both known and predicted interactions
- Use several data sources to increase coverage
- Score the results to enable comparison

Goals / Background

InterologueFinder tasks:

- Download files from Dip, IntAct, Bind and Ensembl for 5 species
- Query NCBI to ensure up-to-date identifiers
- Combine interactions without duplicates and with unified identifiers
- Prediction within a species
  - Human Gene 1 ←?→ Human Gene2
  - If Mouse Gene 1' ⟷ Mouse Gene 2'
  - Predict Human Gene 1 ←√→ Human Gene 2

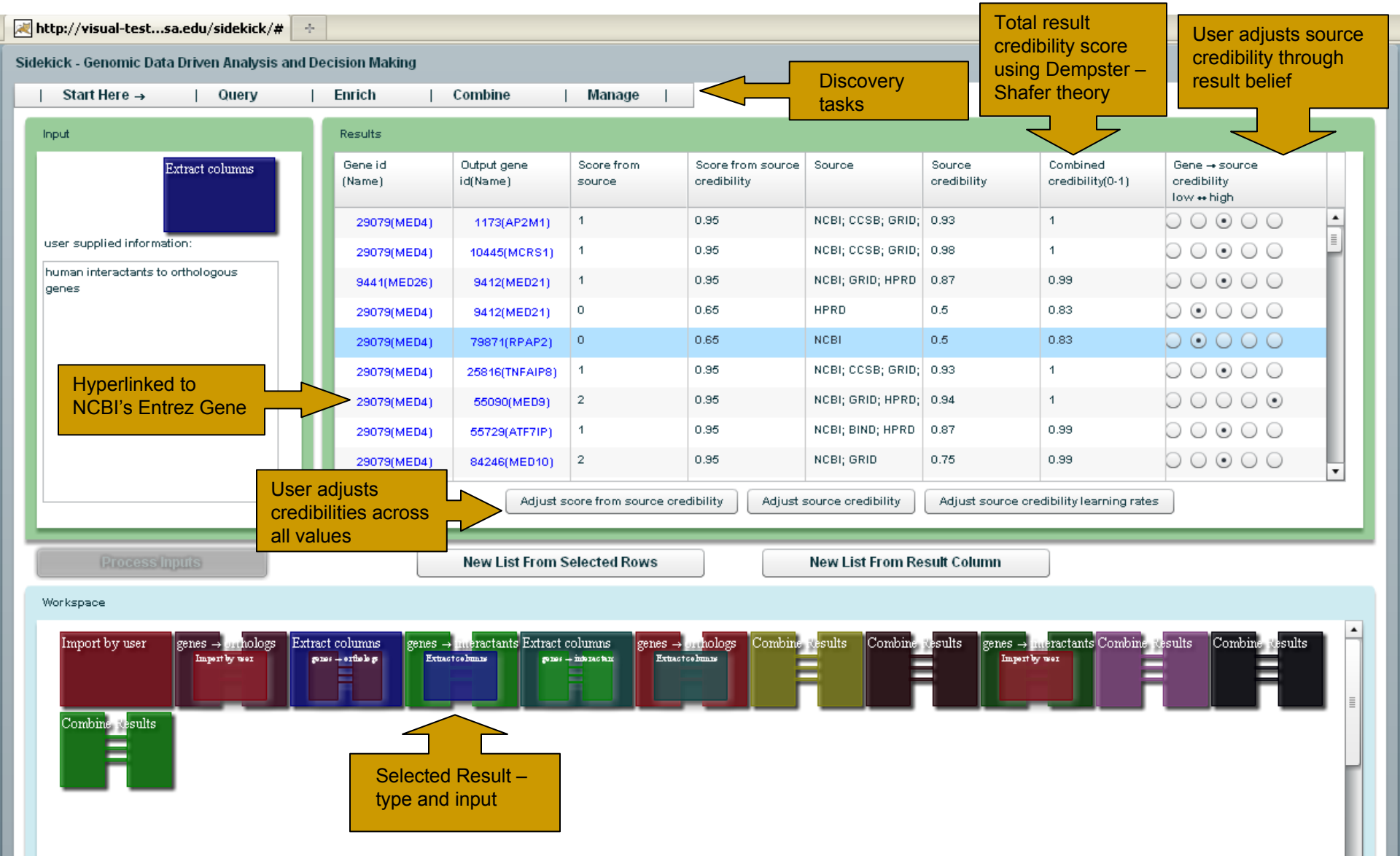
Modules

Sidekick: a web-based tool for genomic exploration

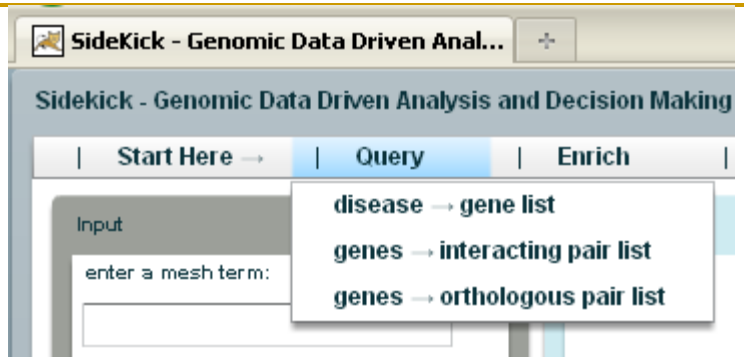
- emphasizes ease-of-use for non-computational users
- stays up-to-date through various web services both local and publicly available
- relies on NCBI for identification and links out
- enables user participation in assigning confidence measurements
- supports combination of results and selection of results
- allows entire research processes or single results to be saved and retrieved
- is modular for easy future expansion

- is not suited to large genome wide studies

Sidekick in Action



Modules - Query



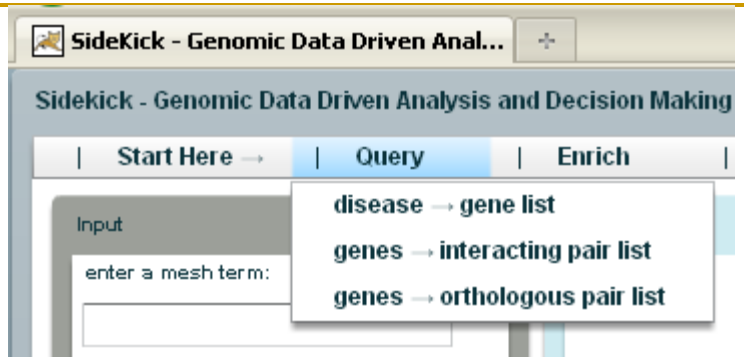
Query: disease → gene list

- generates a list of genes given a search term
- uses NCBI and NCIBI's gene2mesh web services
- p-value to describe the association between gene and term (*Score-from-source*)
- provides two methods to influence the credibility of the source of information

Results

Gene id (Name)	Score from source (Raw P-Value)	Score from source credibility	Source	Source credibility	Combined credibility(0-1)	Gene — source credibility low — high
8151(WT4)	1.301(5.00e-2)	0.8	Ncbi	0.5	0.9	
8136(WT3)	1.301(5.00e-2)	0.8	Ncbi	0.5	0.9	

Modules - Query

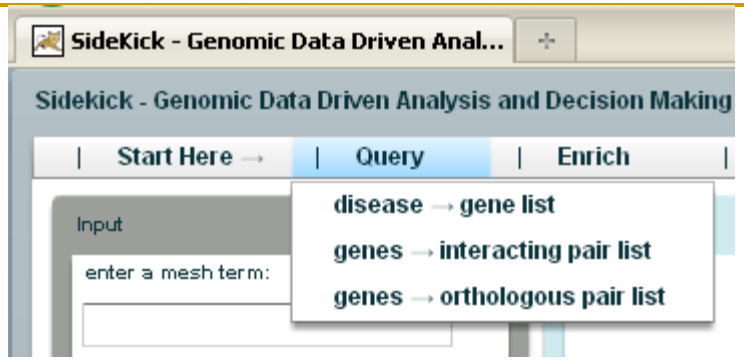


Query: genes → interacting pair list

- generates interactions given a list of genes
- NCBI's and NCIBI's MiMI web services
- *Score-from-source* is number of articles
- provides two methods to influence the credibility of the source of information

Gene id (Name)	Output gene id(Name)	Score from source	Score from source credibility	Source	Source credibility	Combined credibility(0-1)	Gene — source credibility low — high
7490(WT1)	8626(TP63)	1	0.8	HPRD	0.8	0.8	○ ○ ◉ ○ ○
7490(WT1)	374(AREG)	1	0.8	HPRD; MINT	0.8	0.8	○ ○ ◉ ○ ○

Modules - Query



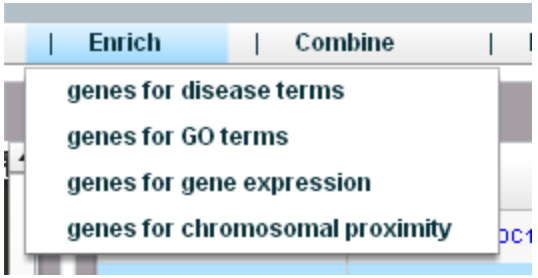
Query: genes → orthologous pair list

- generates orthologues given a list of genes
- local web service that downloads from Ensembl
- *Score-from-source* is percent identity
- enables user to evaluate the credibility of the source of information

Results

Gene id (Name)	Output gene id(Name)	Score from source	Score from source credibility	Source	Source credibility	Combined credibility(0-1)
6774(STAT3)	100045296(LOC100...	99	0.8	Ensembl	0.8	0.8
6774(STAT3)	20848(Stat3)	99	0.8	Ensembl	0.8	0.8

Modules - Enrich

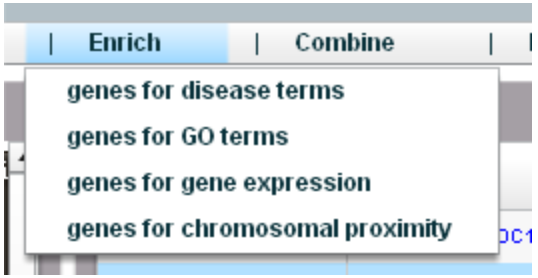


Enrich: genes for GO terms

- explores the relationships between genes in a given a gene list according to their Gene Ontology (GO) annotations
- uses a local web service based on NCBI's gene2go
- user selects the *GO Type*, *Evidence to include, Enrichment strategy, and Maximum # of GO* terms to return
- *Score-from-source* is a p-value
- includes size of group credibility
- grouped by term and genes

Profile	Gene id (Name)	Score from sour...	Score from source credibility	Size of group	Group size credibility	Combined credibility(0-1)
death-inducing signalin			0.8	4	0.8	0.96
death-inducing signa	355(FAS)	4.063(8.64e-5)	0.8	4	0.8	0.96
death-inducing signa	8772(FADD)	4.063(8.64e-5)	0.8	4	0.8	0.96
death-inducing signa	841(CASP8)	4.063(8.64e-5)	0.8	4	0.8	0.96
death-inducing signa	8737(RIPK1)	4.063(8.64e-5)	0.8	4	0.8	0.96
extracellular region par			0.8	43	0.8	0.96

Modules - Enrich

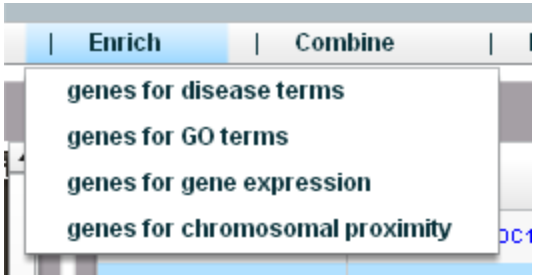


Enrich: genes for disease terms

- enriches gene lists according to their disease Mesh annotations
- NCIBI gene2mesh
- user selects *Enrichment strategy and Maximum # of* terms to return
- *Score-from-source* is a p-value

Profile	Gene id (Name)	Output gene id(Name)	Score from so...	Score from so...	Size of group	Group size credibility	Combined credibility
Kidney Neoplas				0.8	2	0.8	0.96
Kidney Neop	7490(WT1)	5076(null)	1.501(3.15e-2)	0.8	2	0.8	0.96
Kidney Neop	7490(WT1)	9589(WTAP)	1.501(3.15e-2)	0.8	2	0.8	0.96
Male Urogenital				0.8	5	0.8	0.96

Modules - Enrich

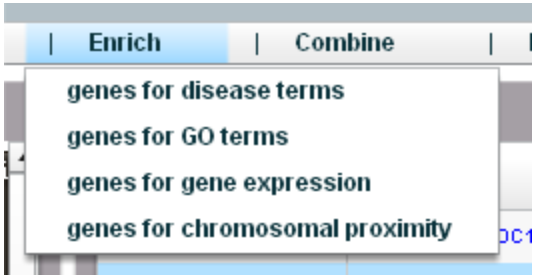


Enrich: genes for gene expression

- genes according to their expression levels
- Gene Expression Atlas contains curated data for gene expression under different biological conditions
- conditions include cell line, disease progression, etc.
- *Score-from-source* is difference between the number of up regulated experiments and the number of down regulated experiments

celltype:CD105+ endothelial(6)			0.8	6	0.8	0.96
celltype:CD105+ endothelial(6)	8718(TNFRSF...	1(1.00)	0.8	6	0.8	0.96
celltype:CD105+ endothelial(6)	472(ATM)	1(1.00)	0.8	6	0.8	0.96
celltype:CD105+ endothelial(6)	3576(IL8)	-1(-1.00)	0.8	6	0.8	0.96
celltype:CD105+ endothelial(6)	332(BIRC5)	-1(-1.00)	0.8	6	0.8	0.96

Modules - Enrich

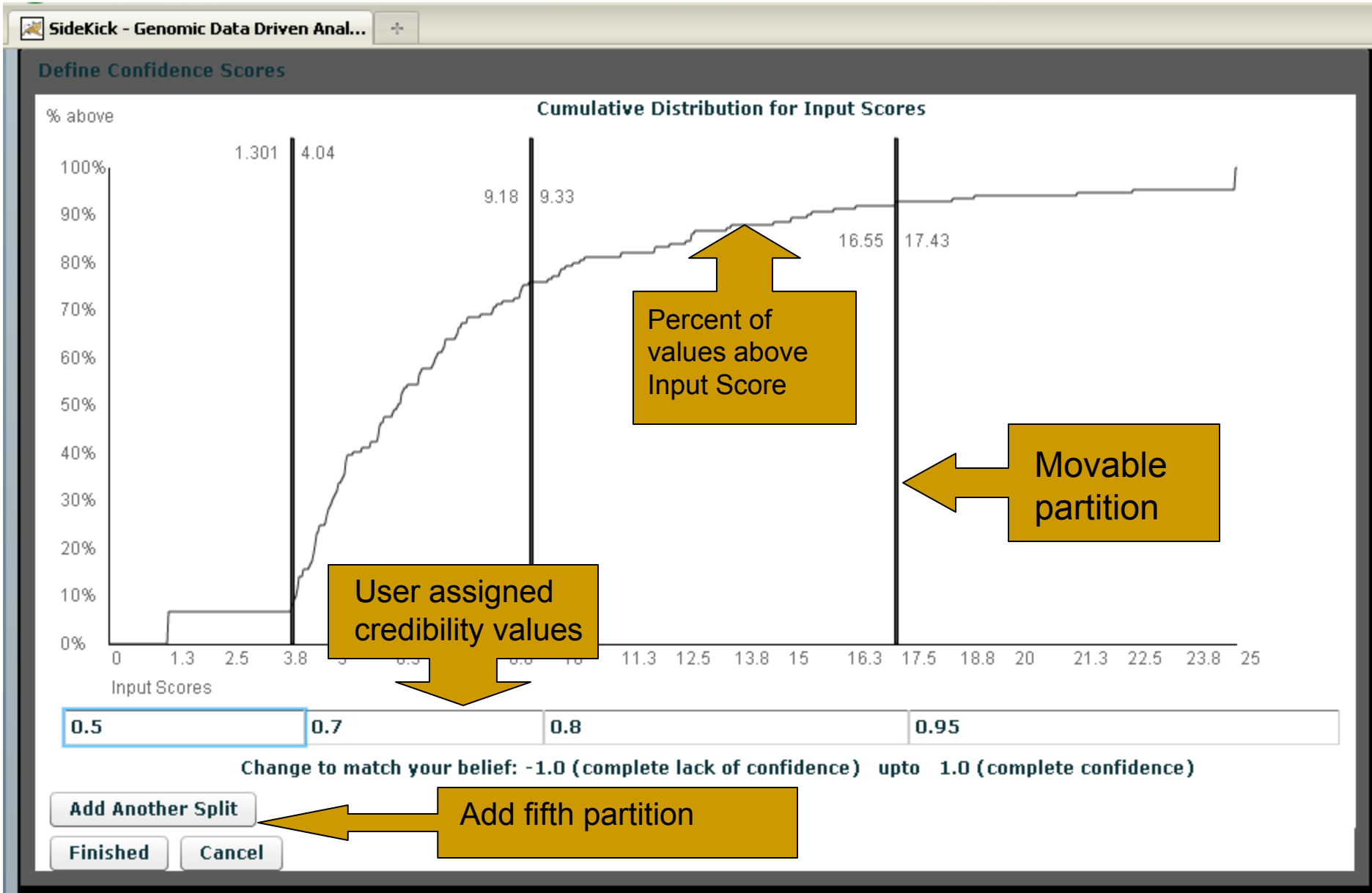


Enrich: genes for chromosomal proximity

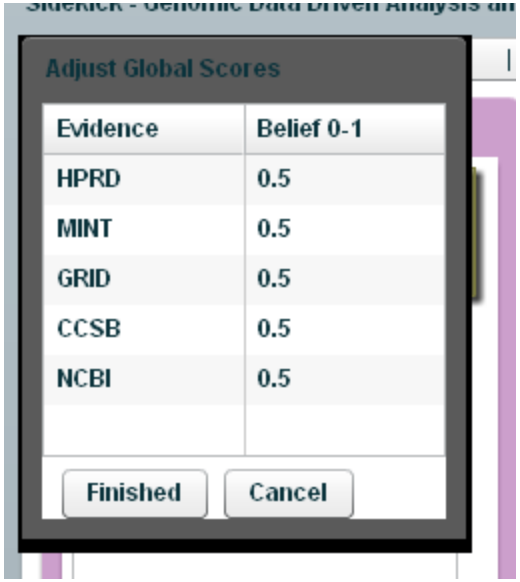
- gene list enriched according to their chromosomal proximity as determined by the number of base pairs separating the start positions of the genes
- source is NCBI for both genome wide and gene specific annotations
- *Score-from-source* is a p-value

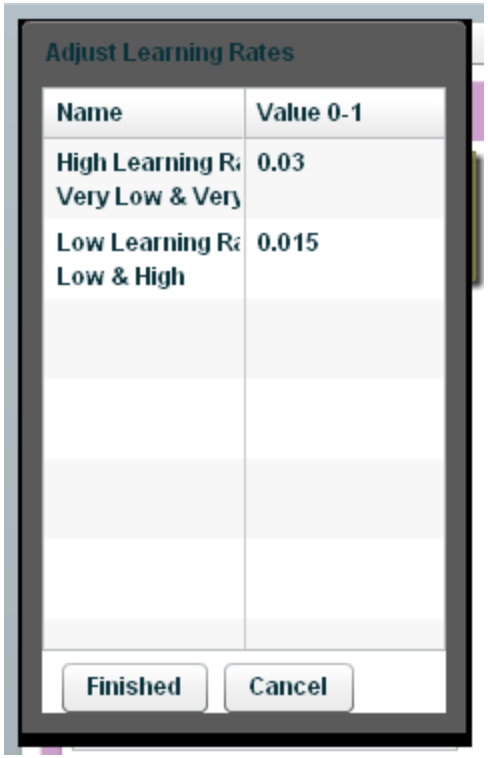
chromosome 1 fro			0.5	5	0.8	0.9
chromosome 1	3814(KISS1)	3.968(1.07e-4)	0.5	5	0.8	0.9
chromosome 1	7832(BTG2)	3.968(1.07e-4)	0.5	5	0.8	0.9
chromosome 1	4194(MDM4)	3.968(1.07e-4)	0.5	5	0.8	0.9
chromosome 1	3586(IL10)	3.968(1.07e-4)	0.5	5	0.8	0.9
chromosome 1	11009(IL24)	3.968(1.07e-4)	0.5	5	0.8	0.9
chromosome 1 fro			0.5	7	0.8	0.9
chromosome 1 fro			0.95	2	0.8	0.99
chromosome 1	11009(IL24)	25(0.00e-16)	0.95	2	0.8	0.99
chromosome 1	3586(IL10)	25(0.00e-16)	0.95	2	0.8	0.99

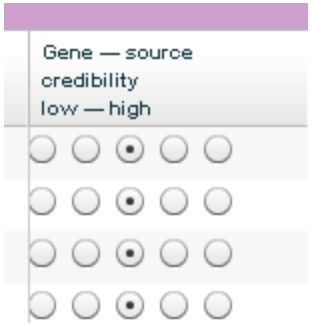
Credibility Visualization



Credibility Assignment

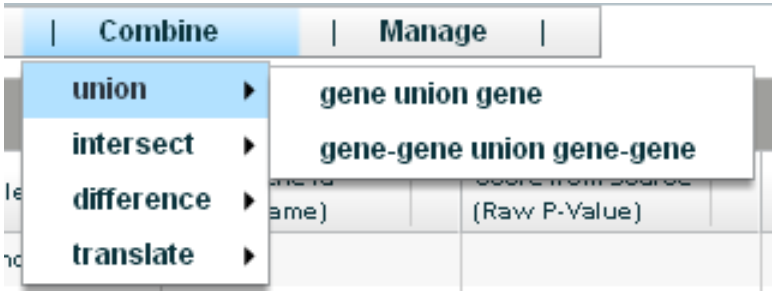


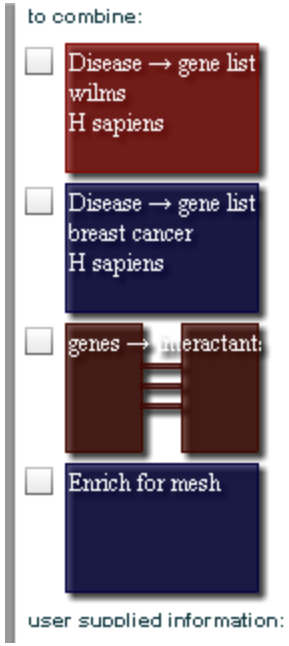




Combined Credibility formed from all of the confidence measurements using Dempster – Shafer theory

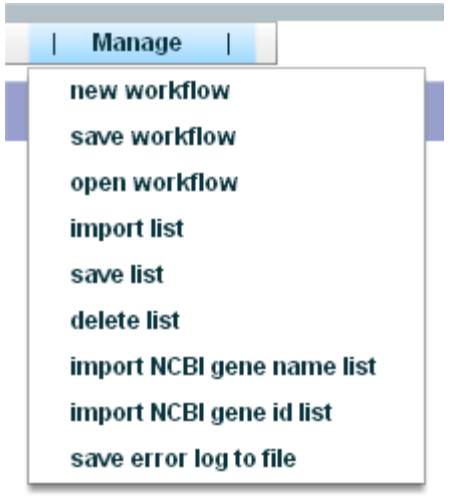
Modules - Combine





- union
  - *gene union gene*
  - *gene-gene union gene-gene*
- intersect and difference
  - *gene intersect gene*
  - *gene-gene intersect gene-gene*
  - *gene intersect gene-gene*
- translate
  - *geneA-geneB translate geneB-geneC → geneA-geneC*

Modules - Manage



Manage:

- new workspace
- save workspace
- open workspace
- import list
- save list
- delete list
- import NCBI gene name list (9606,BRCA1,BRCA2, …)
- import NCBI gene id list (672,675, …)
- save error log to file

Theories Enabling the Modules

Dempster – Shafer theory (credibility scoring)

- Arthur Dempster and Glenn Shafer
- Models belief, disbelief and uncertainty
- Combines available evidences
- Does not equate lack of evidence with disbelief
- Frame of discernment with no uncertainty
  Θ = {T, F}.
- When uncertainty is also modeled,
  Θ = {T, F, TF}
- Use believability, doubt, upper probability functions

$$Bel(A) = \sum_{B \subseteq A} m(B), forall A \subseteq \Theta \quad Dou(A) = Bel(\neg A) \quad P^*(A) 1 - Dou(A)$$

Theories (continued)

Parent-Child Enrichment - Grossmann *et al.*

- terms in the Gene Ontology form a directed acyclic graph.
- simple term-for-term analysis misses “mismatch repair” as a child of “DNA repair”
- Parent-Child intersection, and parent-child union and term-for-term provide
- Sidekick uses an open-source implementation of these algorithms as a local web service

Future Goals

- Expansion of queries
- Incorporating queries built by others
- Expand to other molecules like proteins and protein complexes

- Sidekick Collaborators
  - Kay Robbins PhD.
  - Kihoon Yoon PhD.
- Demonstration: http://visual.cs.utsa.edu/sidekick
- Questions?