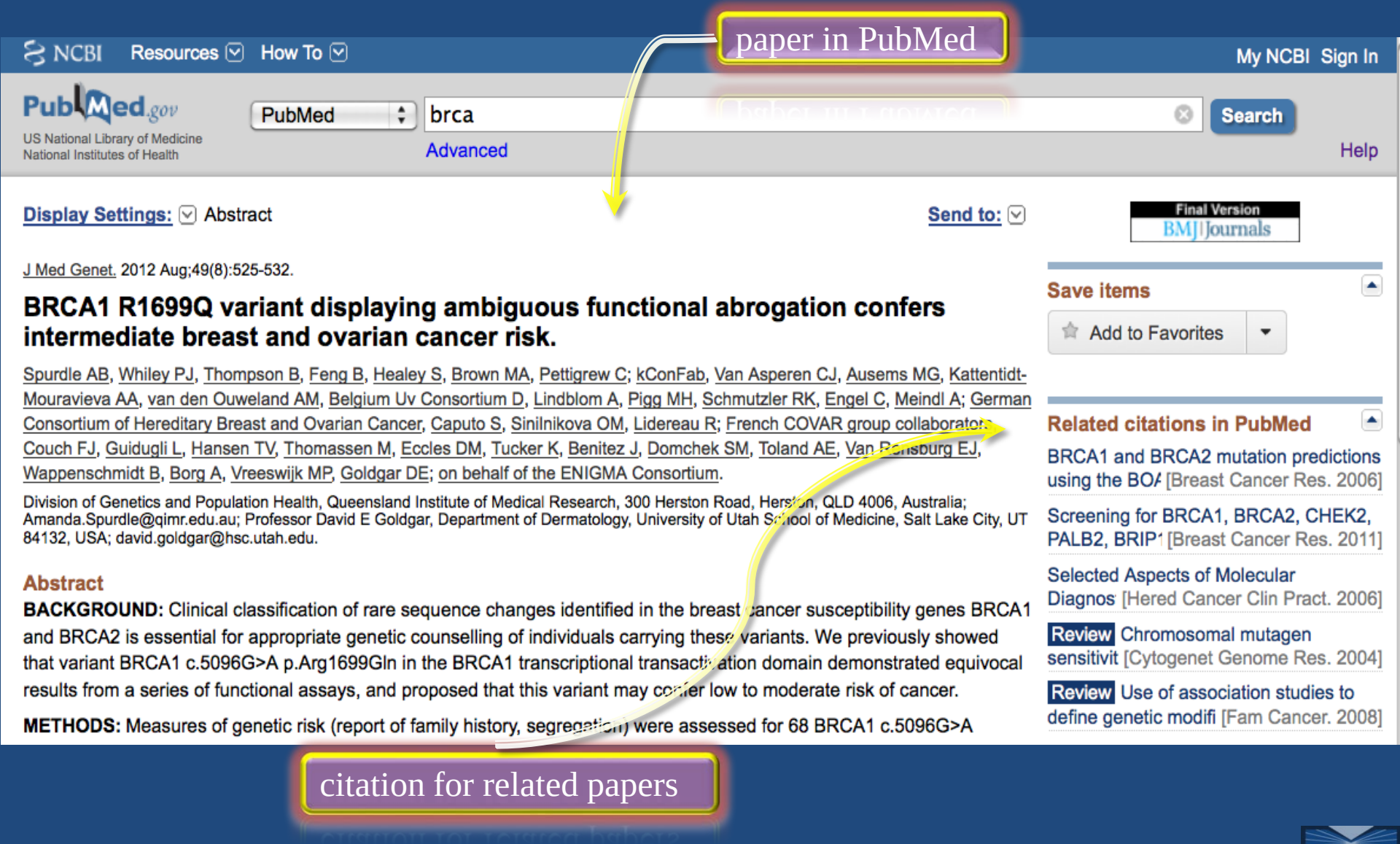
A MUlti-Dimensional integrative approach to comparing genes for knowledge DIScovery

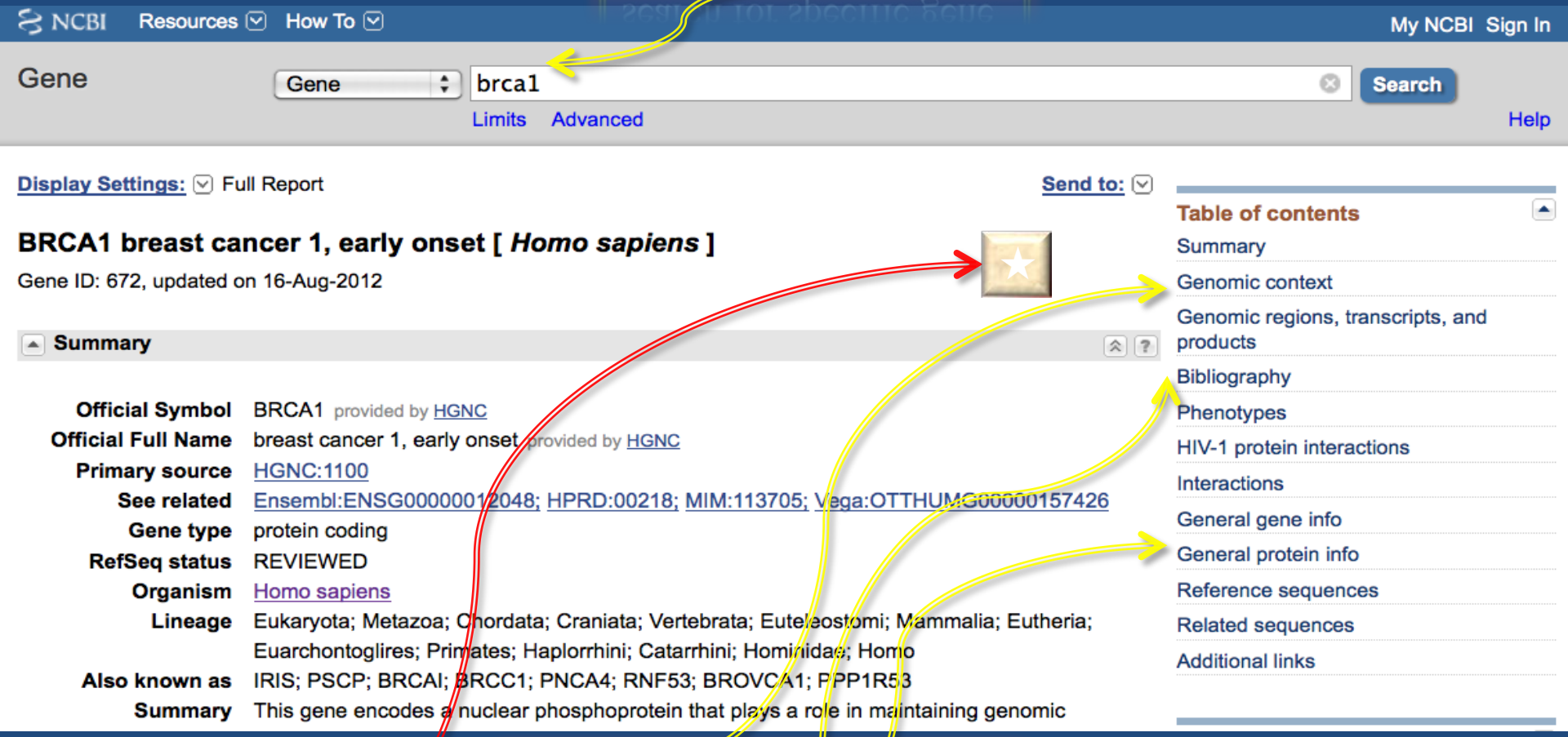
Maryam Panahiazar
PhD Candidate
Computer Science
Bioinformatics
Oct 2012

Motivation , Background



Motivation , Background

search for specific gene

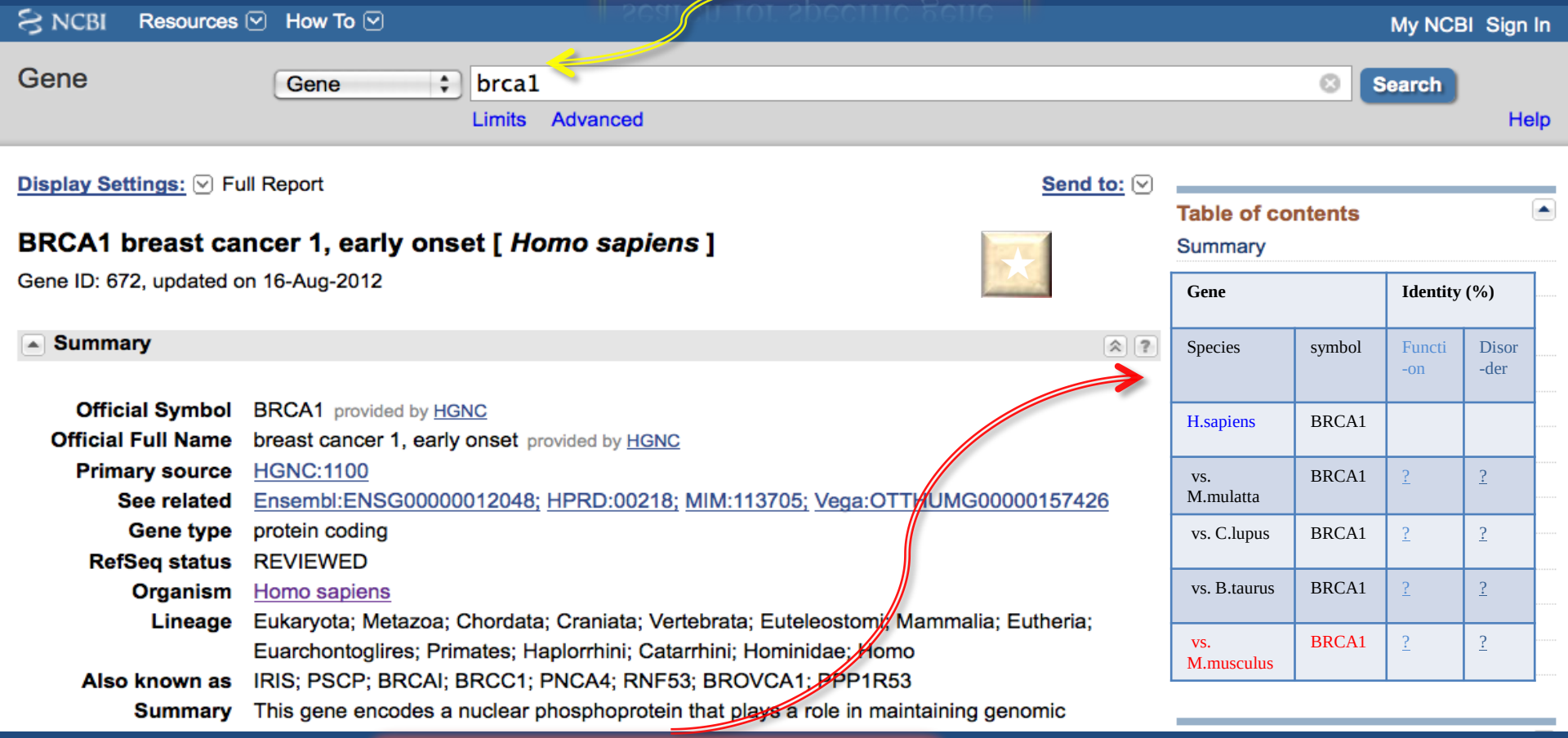


there are links for information

no citation for similar gene records!

Motivation , Background

search for specific gene



citation for similar gene records!

Gene		**Identity (%)**				
Species	symbol	HomoloGene		Muddis		
H.sapiens	BRCA1	protein	DNA	function	disorder	drug
vs. M.mulatta	BRCA1	23	45	?	?	?
vs. C.lupus	BRCA1	67	54	?	?	?
vs. B.taurus	BRCA1	33	45	?	?	?
vs. M.musculus	BRCA1	85	90	?	?	?

Motivation

- Similarity between genes from different aspects

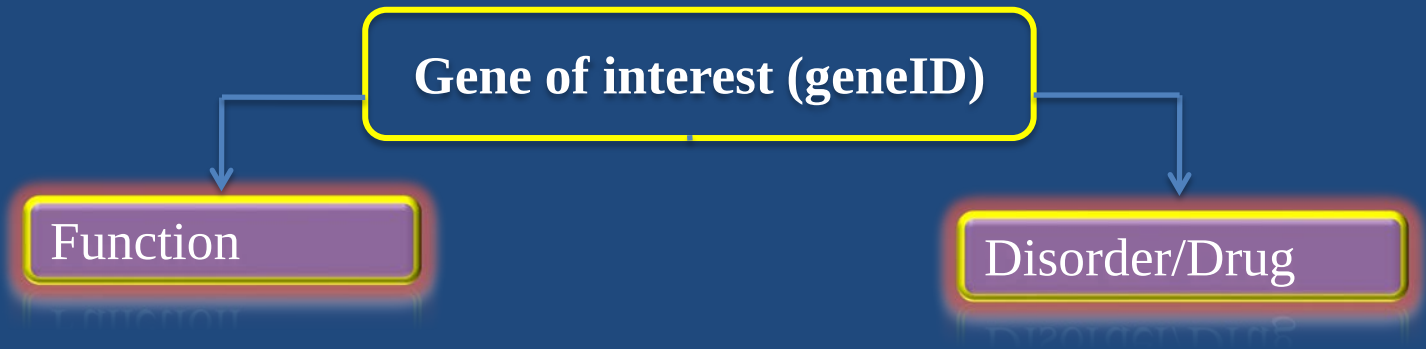
- Curated data bases are limited
  - Not all species
  - Not all aspects

- Semantic knowledge discovery
  - Literature
  - Curated data bases

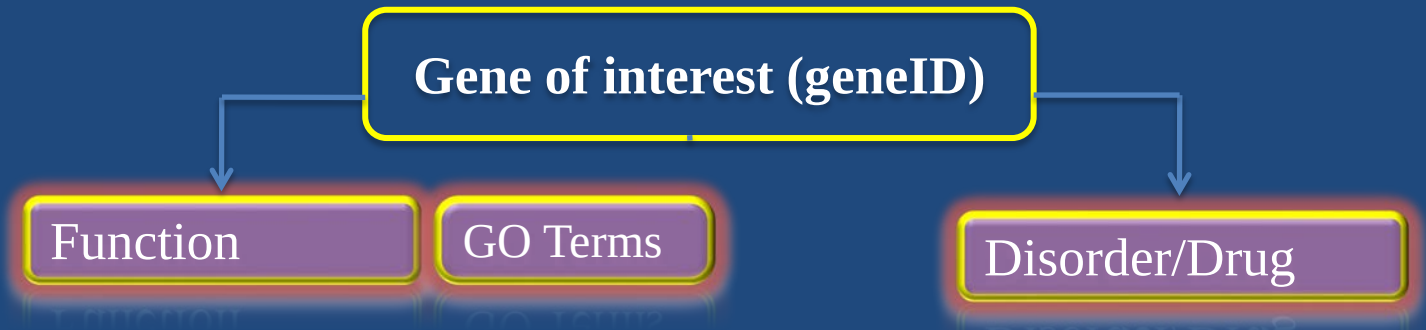
Approach

- Collect data for features
- Define the similarity functions for each feature
- Find the gene-gene similarity based on multiple features

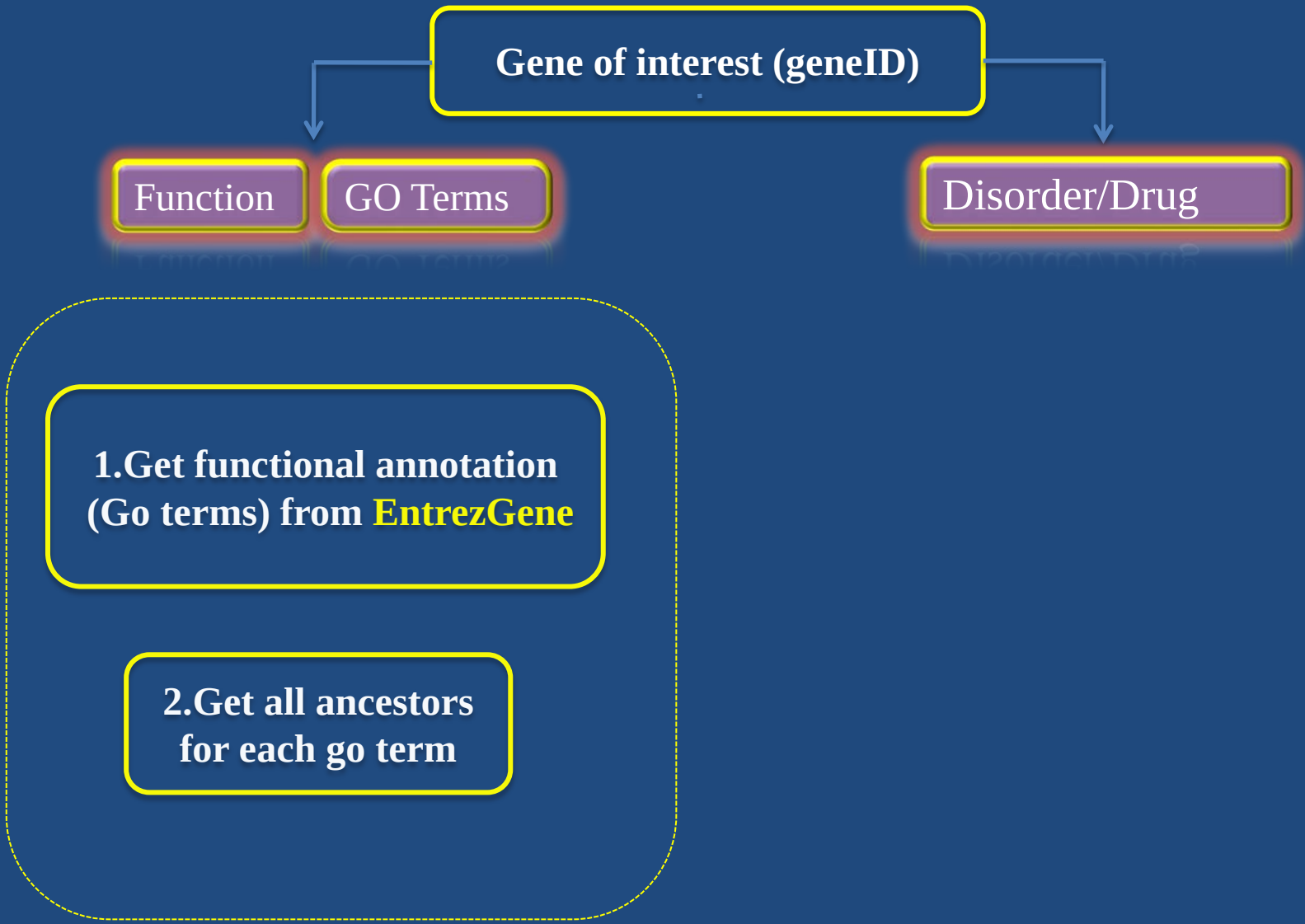
Collect data for features



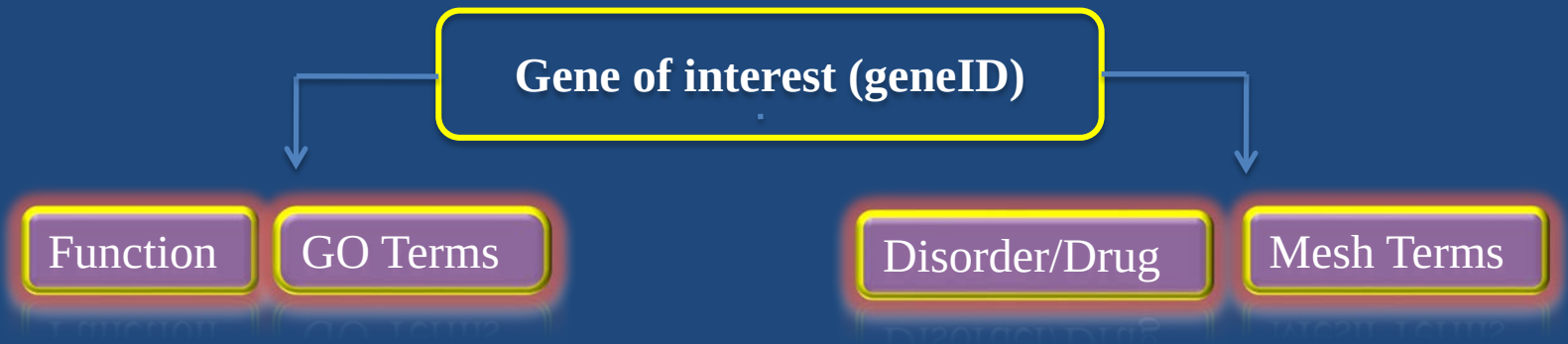
Collect data for features



Collect data for features



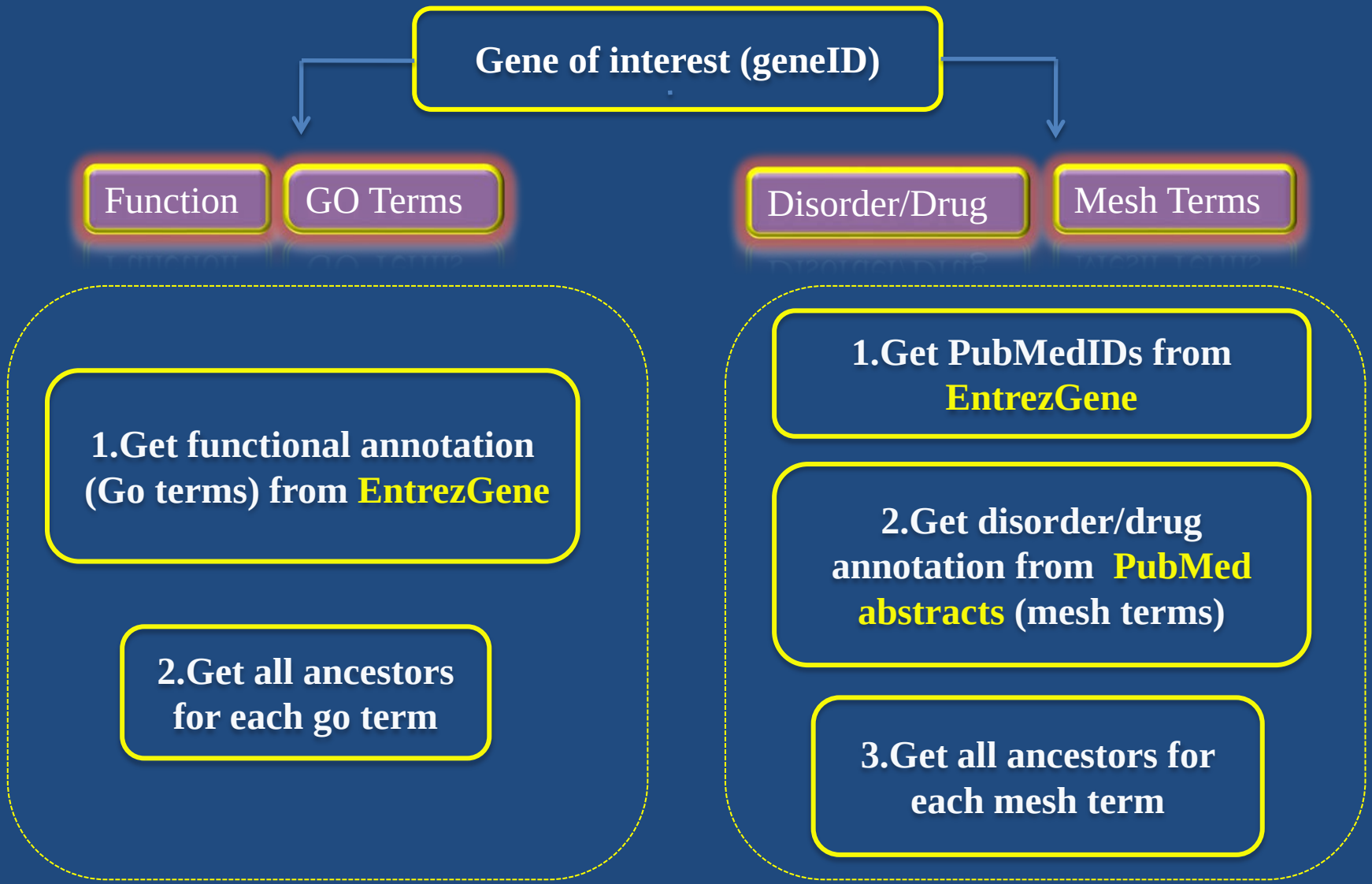
Collect data for features



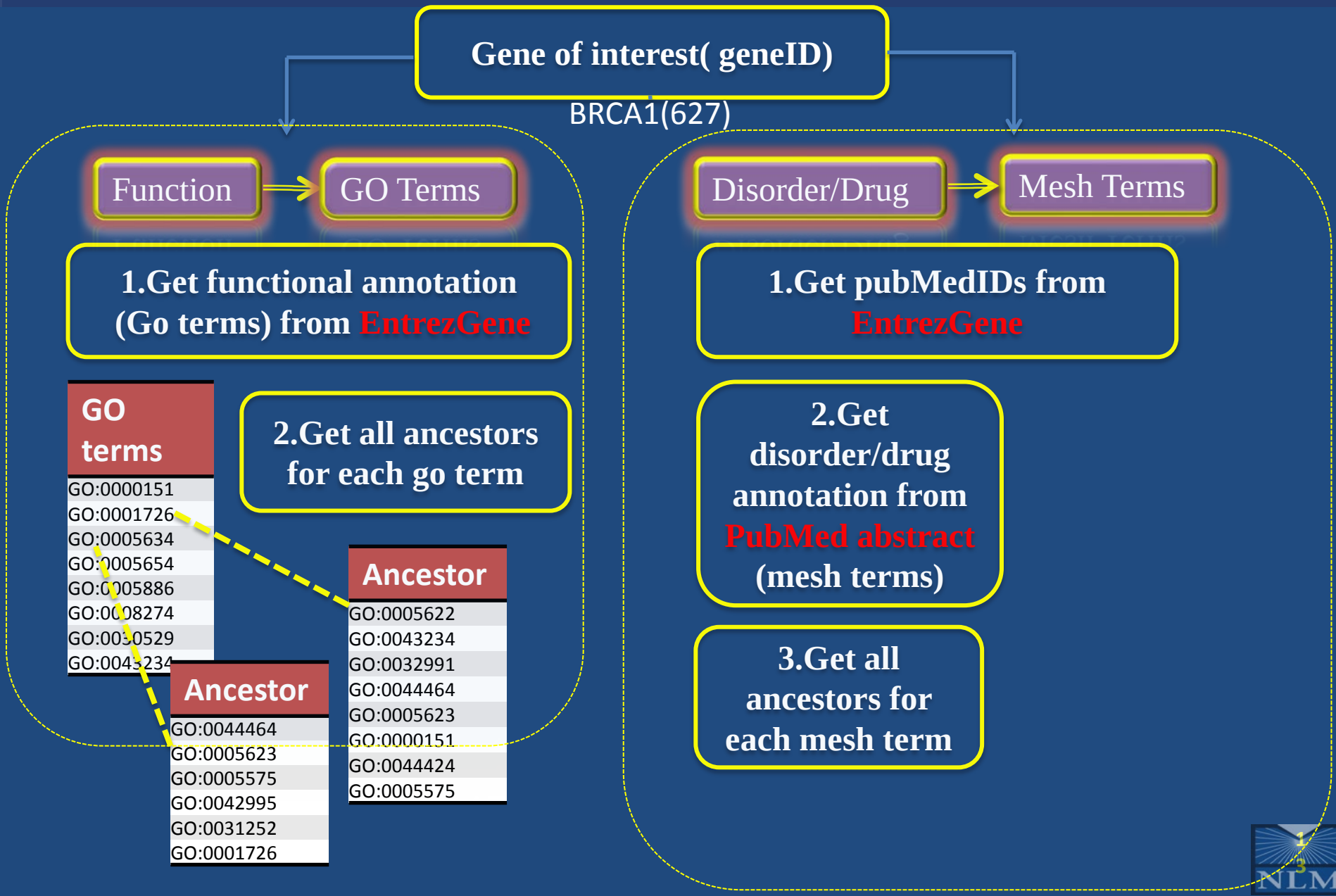
1.Get functional annotation (Go terms) from **EntrezGene**

2.Get all ancestors for each go term

Collect data for features



Materials-Features and resources for each features



Define similarity function between 2 genes

Sim(g1,g2)=?

- Gene of interest: BRAC1

- Species
  - g1: geneID: 672 from H.Sapiens
  - g2: geneID:12189 from M.musculus

- Features
  - f1: Function of the gene
  - f2: Disorder
  - f3: Drug

Sim(f1)	Sim(f2)	Sim(f3)	--------	Sim(fn)

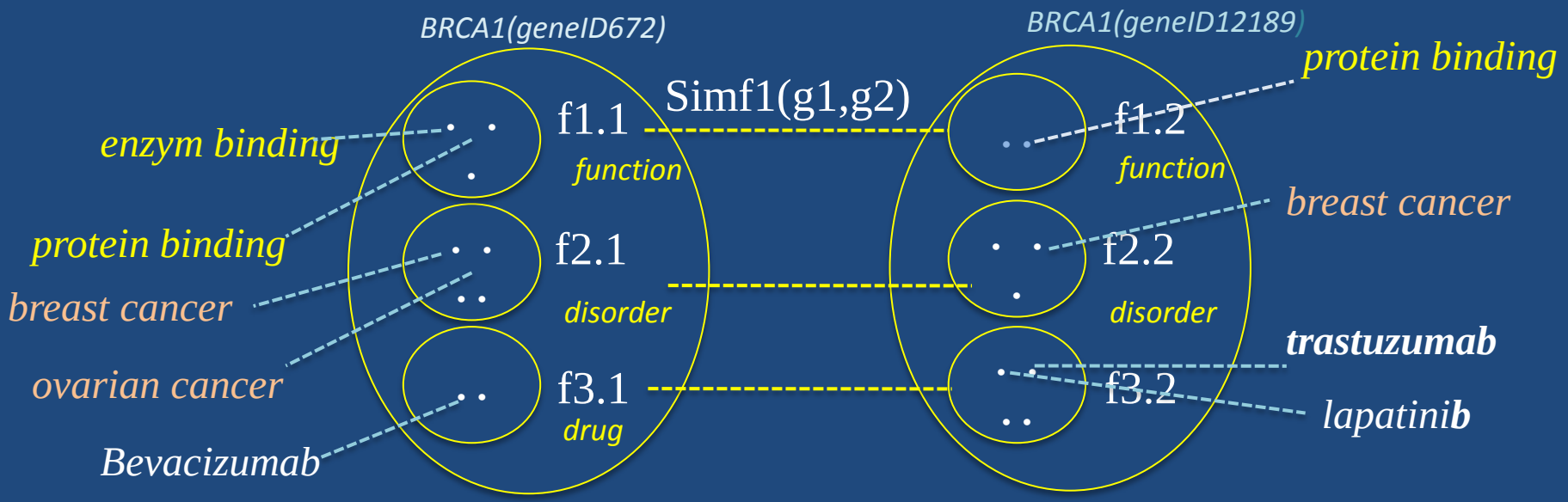
$$sim(g_1,g_2) = \sum_{i=1}^{n} w_i sim_i(g_1,g_2)$$

$$w_1 + w_2 + ...w_i = 1$$

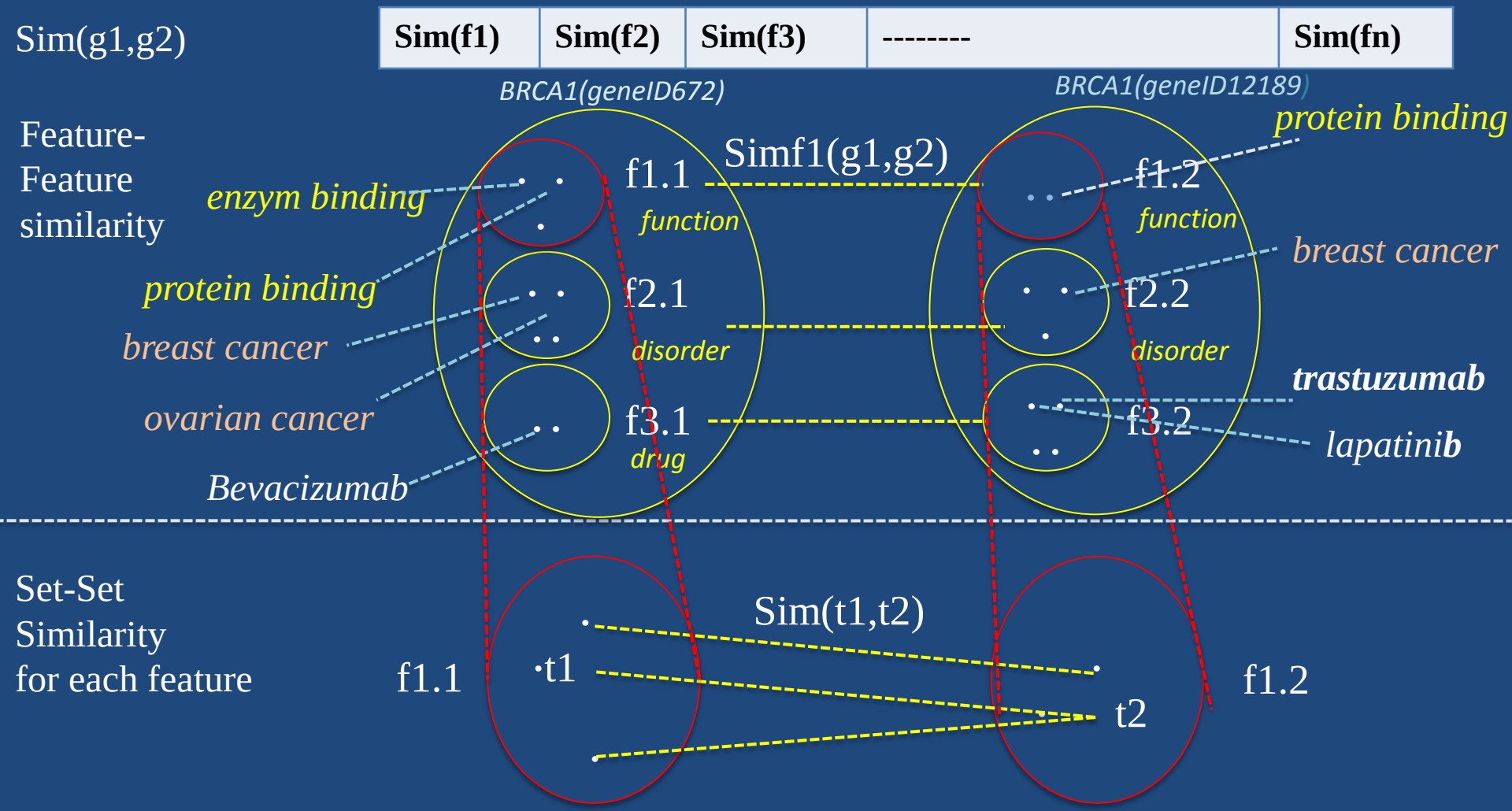
Define similarity function between 2 genes

Sim(f1)	Sim(f2)	Sim(f3)	--------	Sim(fn)

Feature-Feature similarity

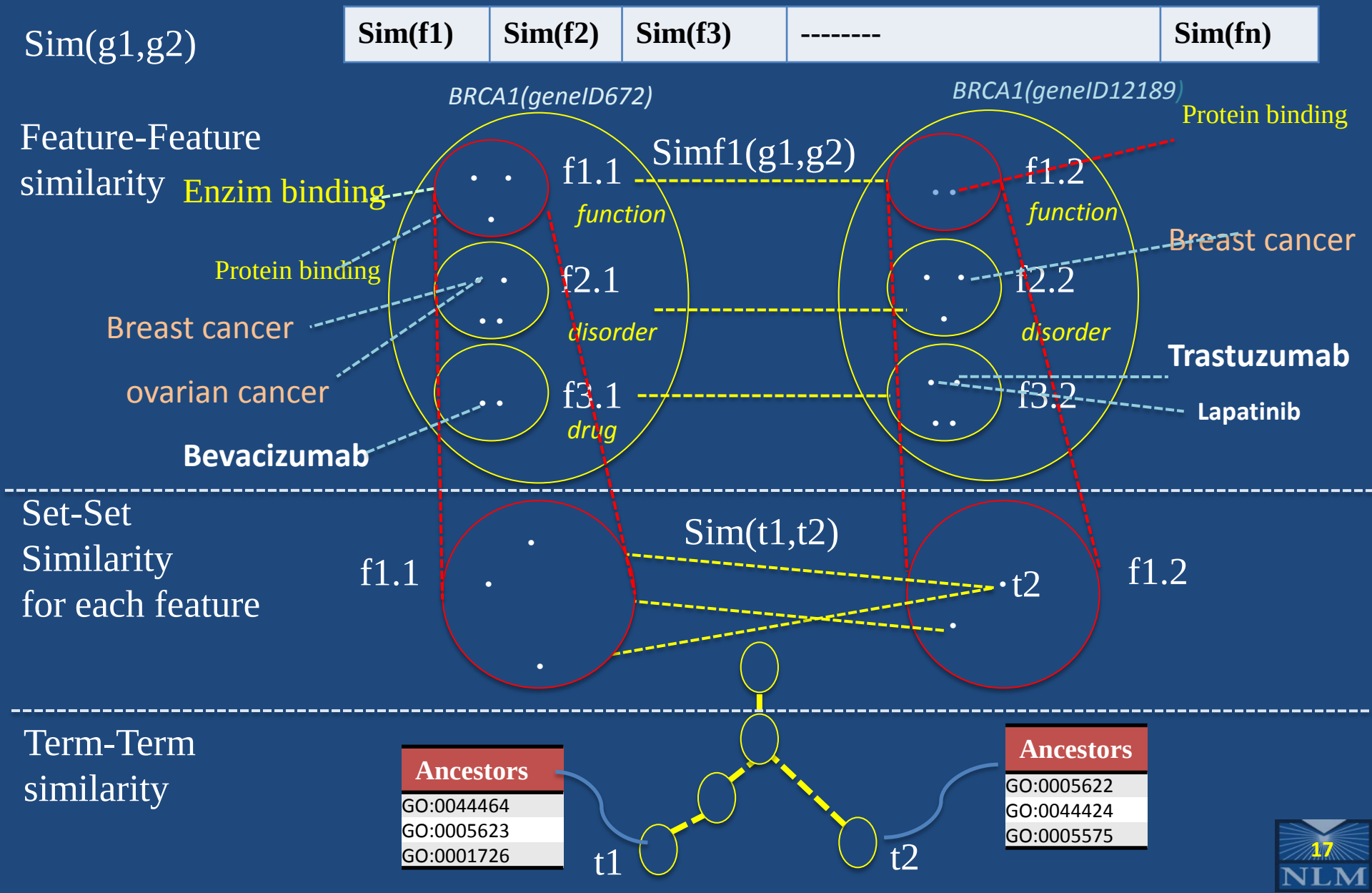


Define similarity function between 2 genes



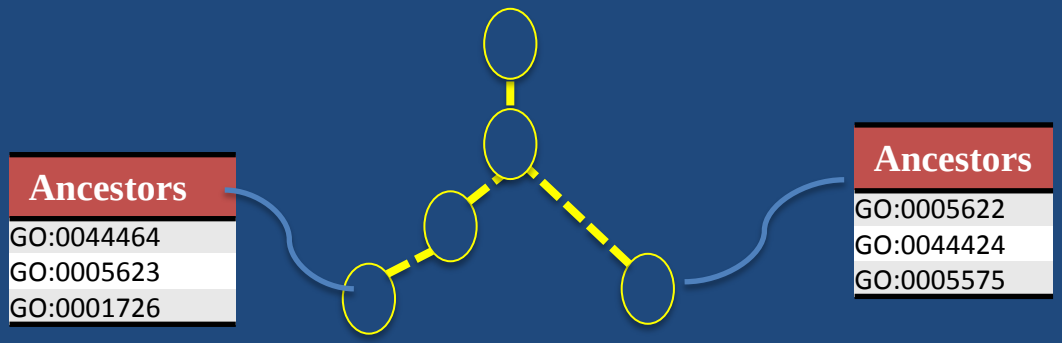
$$sim_i(g_1,g_2) = \frac{1}{m+n} \times \left(\sum_k \max(sim(t_k,t_p)) + \sum_p \max(sim(t_k,t_p))\right)$$

Define similarity functions *H.Sapiens(672) vs. M.musculus(12189)*



Similarity formula

Term-Term similarity



$$sim(t_1,t_2) = -\log \frac{\left|T(t_1) \cup T(t_2)\right| - \left|T(t_1) \cap T(t_2)\right|}{\left|T(t_1) \cup T(t_2)\right|}$$

Francisco Azuaje, Haiying Wang and Olivier Bodenreider 2005

Similarity formula

Aggregation of Feature-Feature similarity

$$sim(g_1, g_2) = \sum_{i=1}^{n} w_i sim_i(g_1, g_2)$$

$$where \quad w_1 + w_2 + ...w_i = 1$$

Set-Set similarity for each feature *i*

$$sim_i(g_1, g_2) = \frac{1}{m+n} \times (\sum_k \max(sim(t_k, t_p)) + \sum_p \max(sim(t_k, t_p)))$$

Term-Term similarity

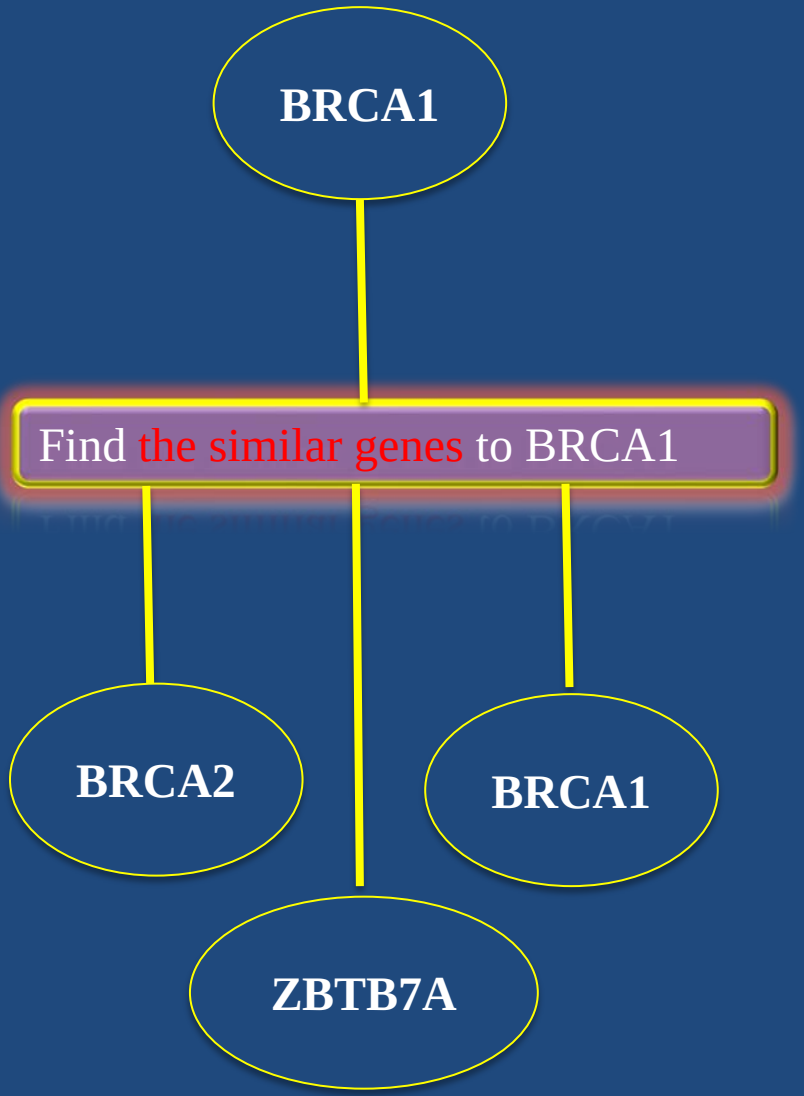
$$sim(t_1, t_2) = -\log \frac{|T(t_1) \cup T(t_2)| - |T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|}$$

Francisco Azuaje, Haiying Wang and Olivier Bodenreider 2005

NLM

Use Case

S----P----O(?)



Intermediate results

BRCA1(Gene ID: 672) in Homo sapiens

Species	GeneID	Sequence similarity		Muddis similarity**		
		Protein*	DNA*	Function	Disorder	Aggregation
vs. M.musculus	12189	58.1	74.4	62.5	45.4	53.95
vs. C.lupus	403437	74.8	84.2	67.3	58.3	62.8
vs. R.norvegicus	24227	58.3	75.2	41.2	22.4	31.8
vs. B.taurus	353120	72.6	83.8	65.8	33.4	49.6
vs. G.gallus	373983	34.3	50.0	50.0	43.4	46.7
vs. P.troglodytes	449497	98.2	99.3	73.4	73.4	62.84
vs. M.mulatta	712634	93.1	96.1	69.5	60.8	65.15

*Pairwise Alignment Scores from HomoloGene

** Results from defined similarity functions

Evaluation frame work

The hypothesis is:
"The results of our similarity functions go with the control similarity"

- Control : Similarity score from

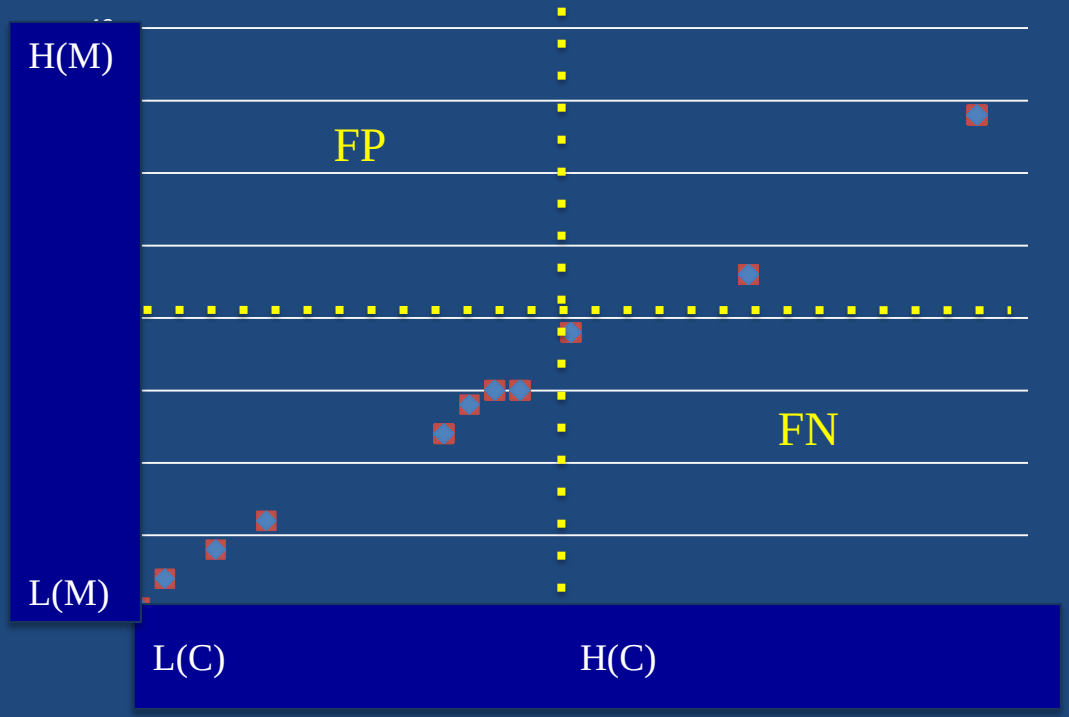
Find the Correlation between feature based similarity to control

- Control : Similarity from Curated data such as

Choose control from **orthologous** genes

Evaluation frame work

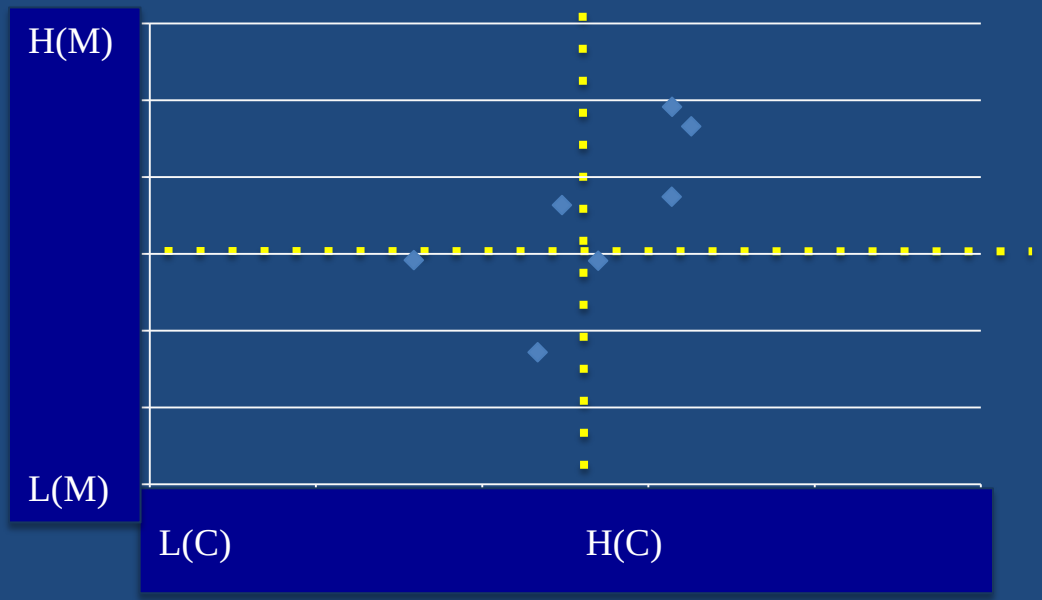
Find the correlation between score of Muddis similarity and the score of HomoloGene



- High control similarity & low Muddis similarity: False negative
- Low control similarity & High Muddis similarity: False positive

Evaluation frame work

Find the correlation between score of Muddis similarity and the score of HomoloGene



•High control similarity & low Muddis similarity: False negative

•Low control similarity & High Muddis similarity: False positive

Correlation {score(muddis), score(homologene)} = 0.67

Discussion

Significance of findings:

- False positives: Candidates for missed gene pairs with high similarity
- False negatives: No correlation between sequence and functional similarity

The results of this research are useful in:

- Model organism research
- Drug target discovery research in human

Future work

- Apply framework in different use cases and evaluate them.
- Use the results for knowledge discovery.
- Add or remove dimensions for similarity comparison.
- Develop an interactive user interface.

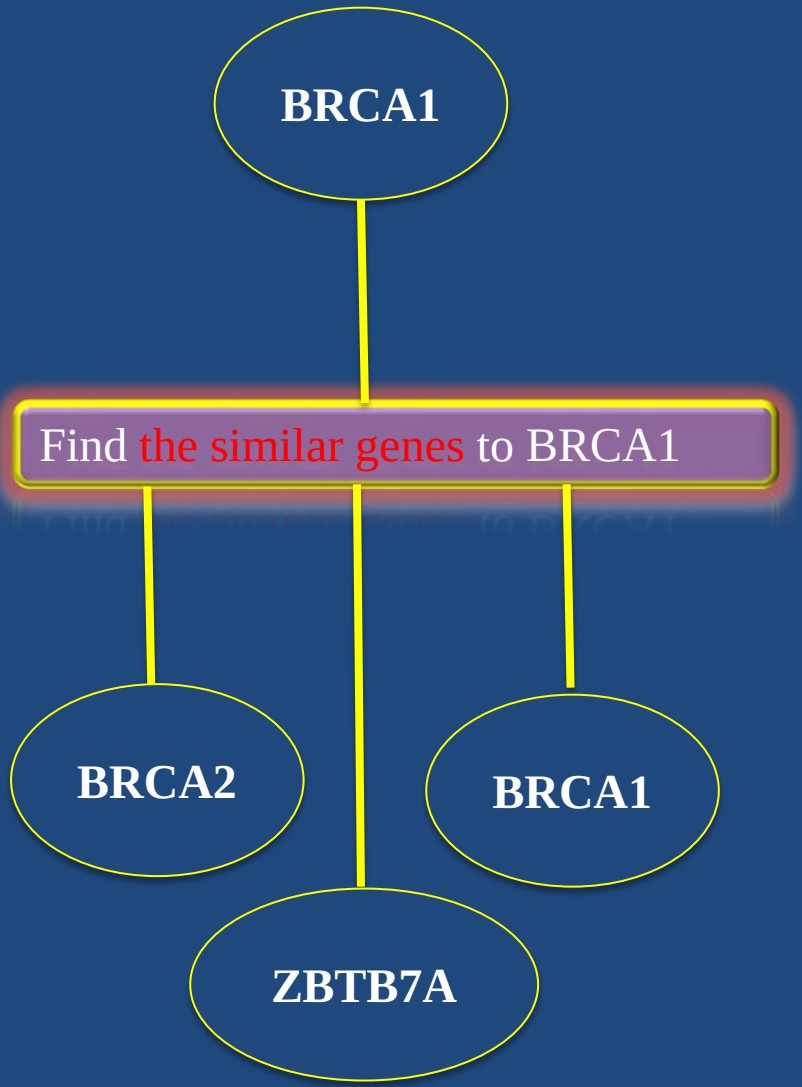
Acknowledgment

Mentor at knoesis: Dr. Amit Sheth
Mentor at NIH: Dr. Olivier Bodenreider

Dr. Bastien Rance
Dr. Rainer Winnenburg
Dr. Marcelo Fiszman
May Cheh

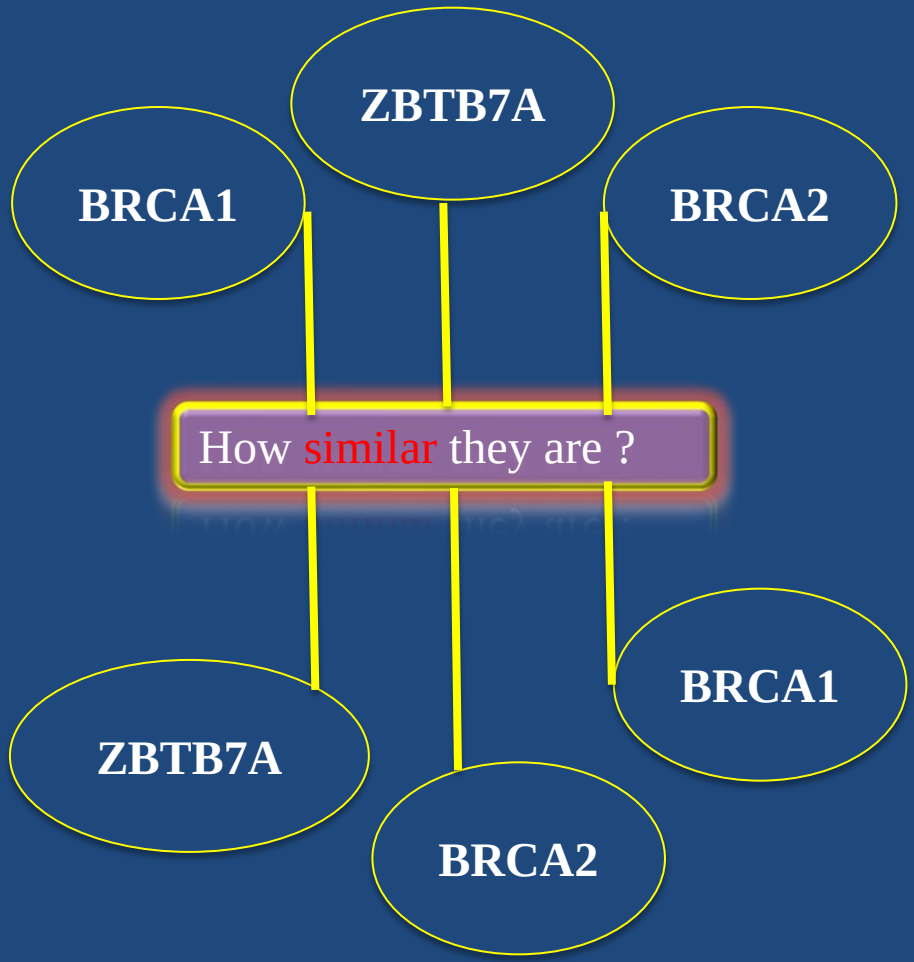
Use Case A

S----P----O(?)



Use Case B

S----P(?)----O



Thanks for your attention

Questions?

Similarity Formula

(C)Feature-feature similarity

$$sim(g_1, g_2) = \sum_{i=1}^{n} w_i sim_i(g_1, g_2)$$

(B)Set-set similarity for each feature *i*

$$sim_i(g_1, g_2) = \frac{1}{m+n} \times \left(\sum\nolimits_k \max(sim(t_k, t_p)) + \sum\nolimits_p \max(sim(t_k, t_p))\right)$$

(A)Term-term similarity

$$sim(t_1, t_2) = -\log \frac{|T(t_1) \cup T(t_2)| - |T(t_1) \cap T(t_2)|}{|T(t_1) \cup T(t_2)|}$$