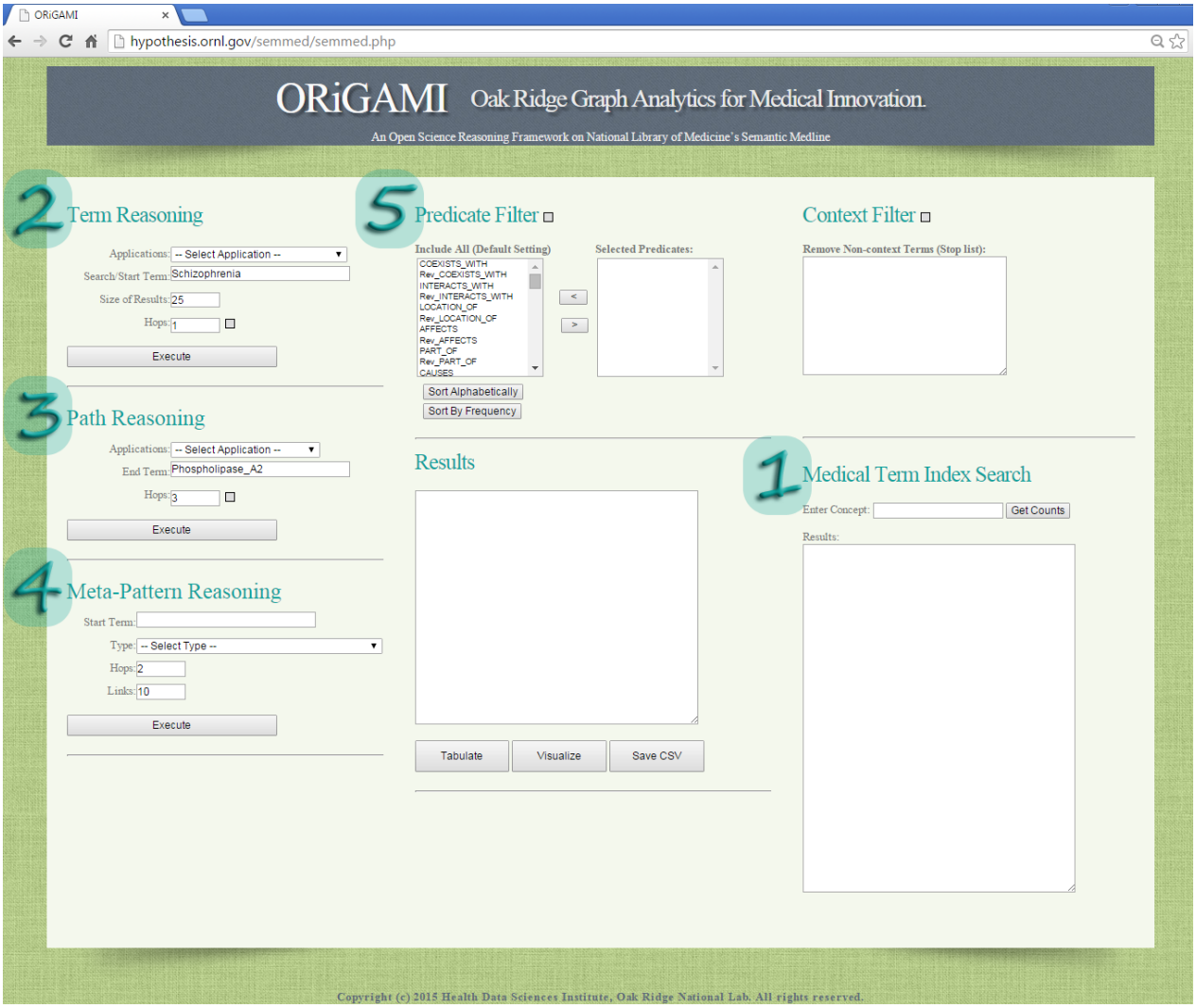
ORiGAMI: Oak Ridge Graph Analytics for Medical Innovation

An Open Science Reasoning Framework on National Library of Medicine's Semantic Medline



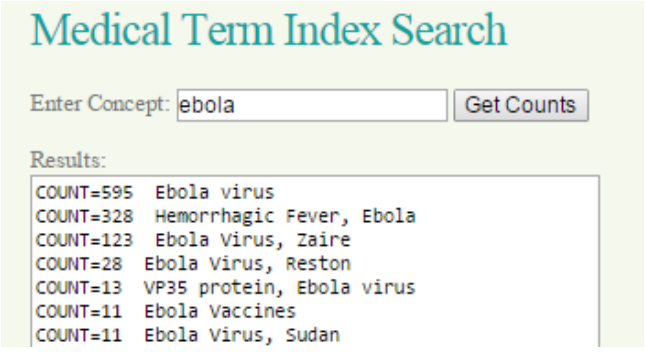
Contributors: Larry Roberts, Sangkeun (Matt) Lee, Keela C. Ainsworth (University of Tennessee), Tyler C. Brown (University of Tennessee), Katie Senter (University of Pennsylvania), Aleksandra Zakrzewska (Yale University), Seokyong Hong (North Carolina State University), Jeffrey Graves (Tennessee Tech), Yifu Zhao (Denison University), Nathaniel Bond (Cedarville University), Yang Zhang (University of Tennessee), Derek Kistler

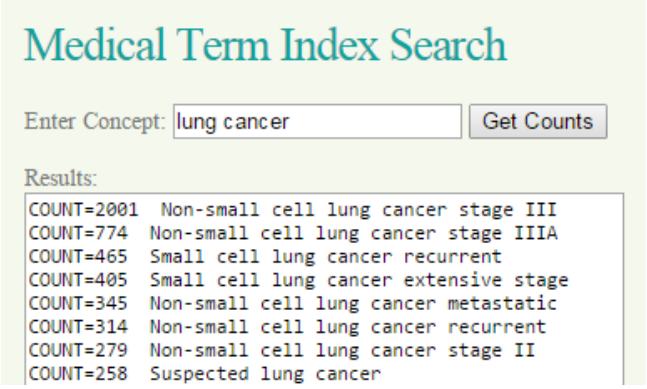
Collaborators: National Library of Medicine – Thomas Rindflesch, Mike Cairelli, Marcelo Fiszman, Federal Big Data Working Group – Brand Neiman, linical/Medical/Bio-Medical – Ed Chaum (UTHSC – Hamilton Eye Institute), Gina Tourassi (Health Data Sciences Institute)

Principal Investigator: Sreenivas Rangan Sukumar

1. Medical Term Index Search

Using the Medical Term Index Search to carefully select the start and end terms is the first step in discovering informative relationships. The Medical Term Index Search helps determine the correct spelling, the most frequently used term, and related terms. Enter a term, click **Get Counts**, and valid terms will be shown.





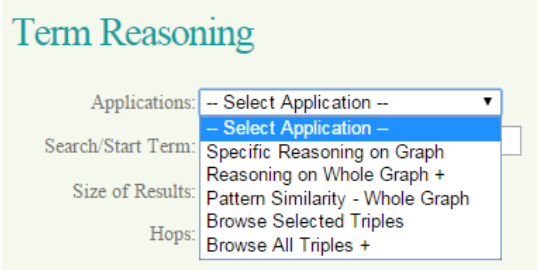
Often a specific term is required for searches rather than a general term. As an example, an application will not return any results for the term *Lung Cancer*. The Medical Term Index Search is available to assist in selecting the best term. As shown here, entering *Lung Cancer* will show various related terms and how often they appear.

Genes and proteins require the duplication of the name. As an example, *TNF protein, human* is *TNF protein, human TNF* and *TNF gene* is *TNF gene TNF*. An unusual exception is *Neoplasm* will show up in the Results as *Neoplasms*, but must be entered as *Neoplasm*.

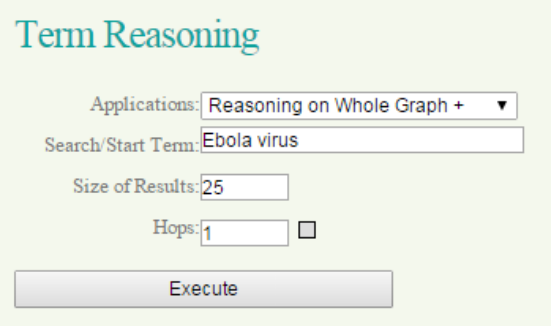
It is important to have the correct spelling, as the system is case sensitive. As an example, *Ebola virus* does not have *virus* capitalized. In addition, only alphanumeric characters are stored in the triples. The system will help convert spaces, commas, and other characters, but if an application does not produce the expected results, try replacing all non-alphanumeric characters with an underscore. *Ebola virus* is *Ebola_virus* and *Alzheimer's Disease* is *Alzheimer_s_Disease*.

2. Term Reasoning

Term Reasoning searches for terms that are similar or related to the **Search/Start Term**. Such terms give clues as to the nature and characteristics of the search term. There are five applications in Term Reasoning.



a. Reasoning on Whole Graph +



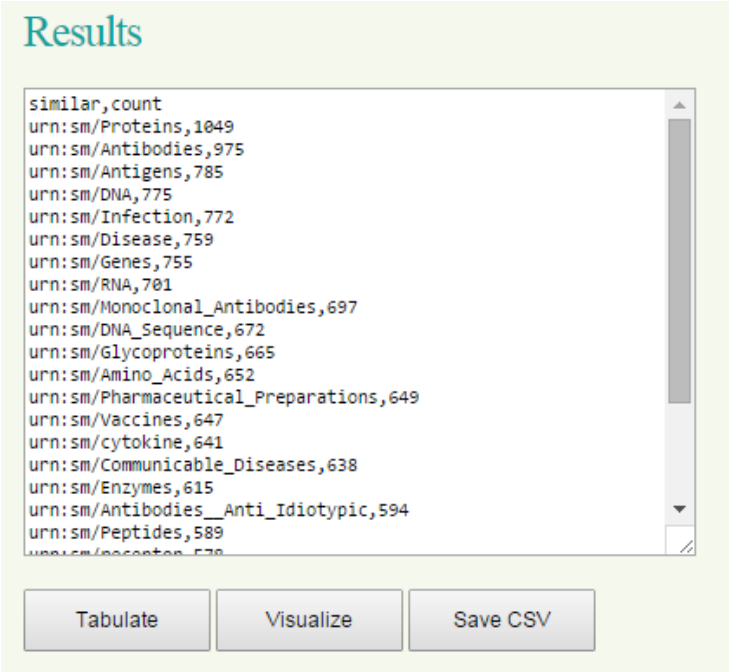
Reasoning on Whole Graph finds terms that have many connecting links to the **Search/Start Term**. A high number of connecting links is one way of determining similarity between the terms. For the whole graph, these terms tend to be descriptive of the search term.

The gray box next to Hops indicates that Hops is not an option for this application.

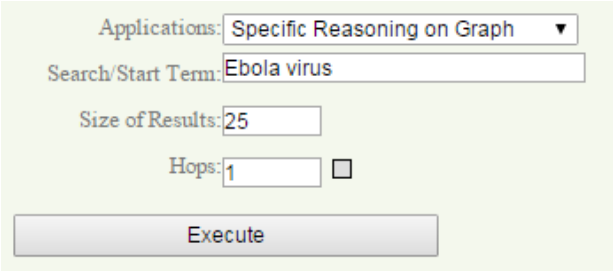
Enter a **Search/Start Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here.

As might be expected, many terms that link to Ebola virus also link to Proteins, Antibodies, and Antigens. A method of searching for more specific terms is shown on the next page.

[The results are only the terms and counts, so the Tabulate and Visualize buttons shown at the bottom of the results are not available.]



b. Specific Reasoning on Graph



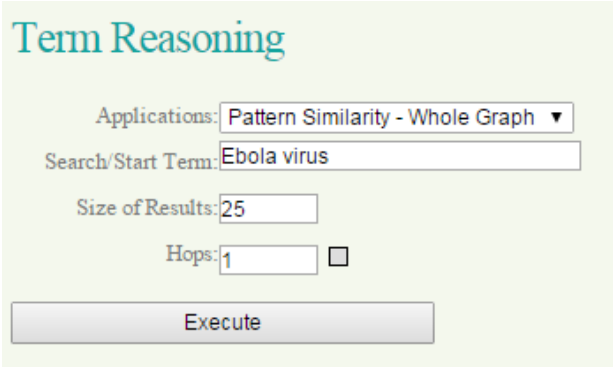
Specific Reasoning on Graph finds terms that uniquely connect to the start term. These terms are more informative than the descriptive terms shown in **Reasoning on Whole Graph**. This search can produce novel connections and show meaningful relationships, leading to new insights.

Enter a **Search/Start Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here.

Note that in comparison to **Reasoning on Whole Graph**, there are fewer connecting links but the results are more informative.

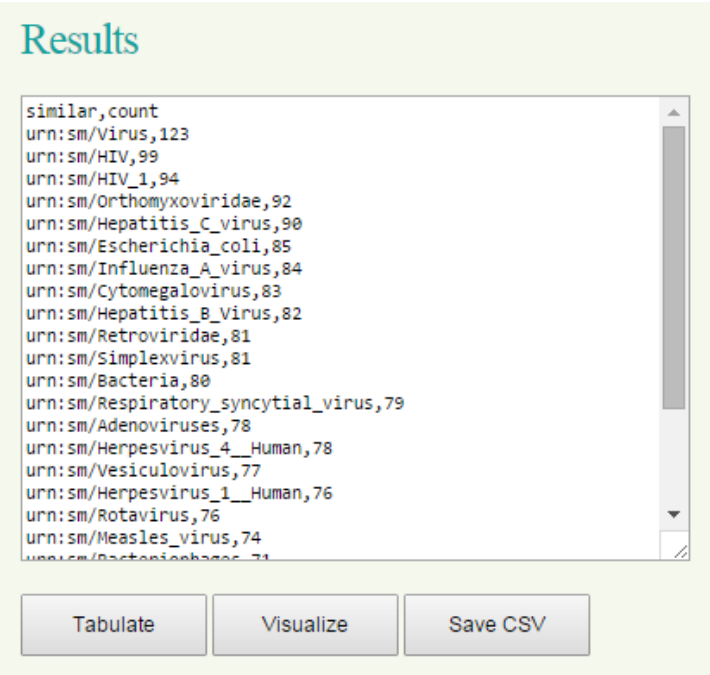
[The results are only the terms and counts, so the Tabulate and Visualize buttons shown at the bottom of the results are not available.]

c. Pattern Similarity – Whole Graph



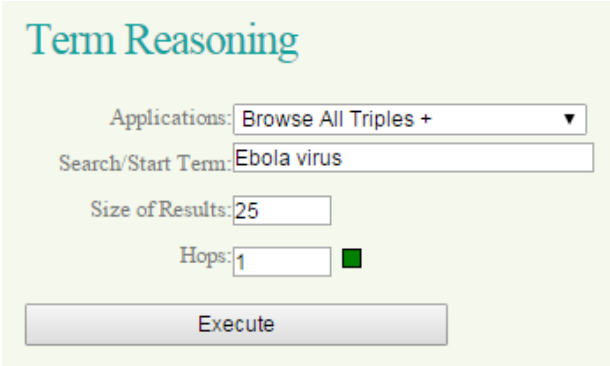
Pattern Similarity – Whole Graph finds terms that have many connecting links, similar to **Reasoning on Whole Graph**, but also accounts for how the terms are connected (predicates). New relationships can be discovered, bringing new insights and leading to serendipitous discoveries.

Enter a **Search/Start Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here. The terms are often more novel than **Reasoning on Whole Graph** results, but less unique than **Specific Reasoning on Graph** results.



[The results are only the terms and counts, so the Tabulate and Visualize buttons shown at the bottom of the results are not available.]

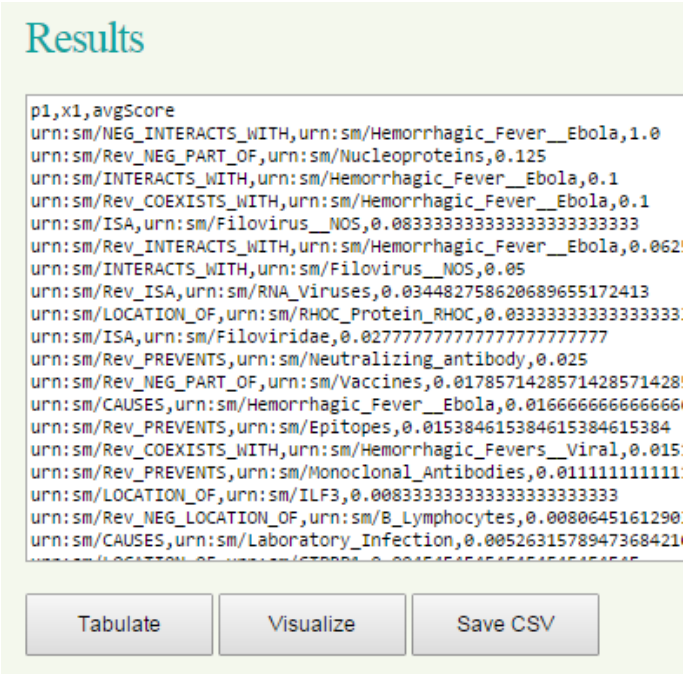
d. Browse All Triples +



Browse All Triples + shows the raw data as it is stored in the graph, and determines a score based on a measure of uniqueness of each triple. The results are shown sorted by score.

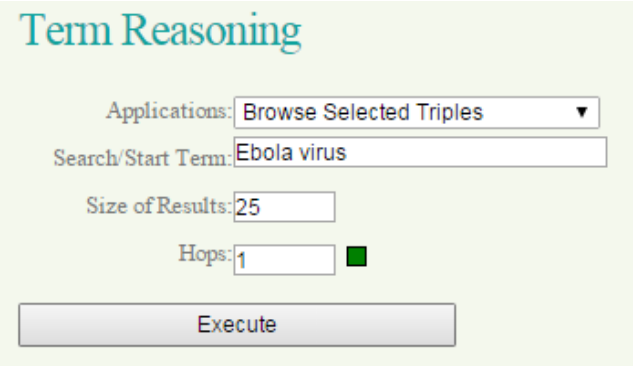
The green box next to Hops indicates that the number of hops (links) is an option for this application. It can take several minutes to get the results from two hops.

Enter a **Search/Start Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here. The raw data is often noisy and contains errors, but can be seen here unfiltered, as desired.



The results can be seen in either a table (Tabulate) or graph (Visualize), as explained on page 8.

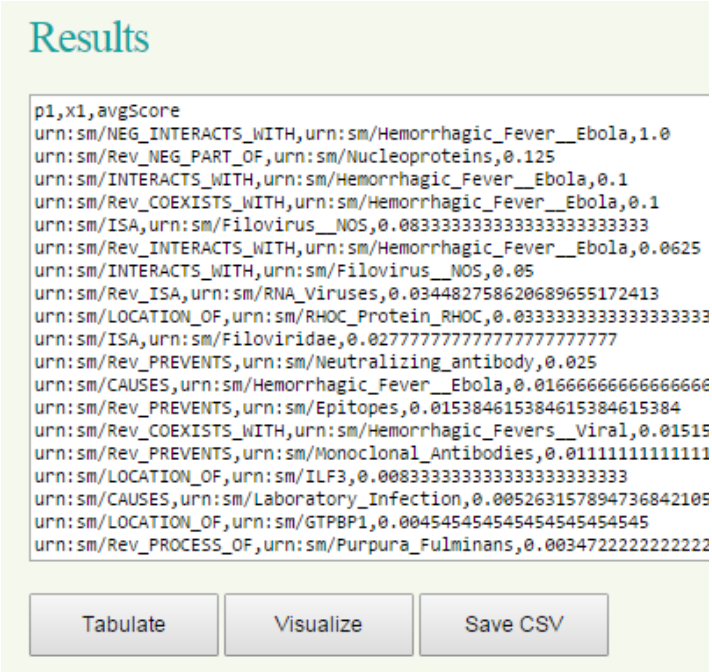
e. Browse Selected Triples



Browse Selected Triples shows the data as it is stored in the informative graph, and determines a score based on a measure of uniqueness of the triple. The informative graph is about 10% of the whole graph, and contains relationships that are descriptive of the terms.

The green box next to **Hops** indicates that **Hops** is an option for this application.

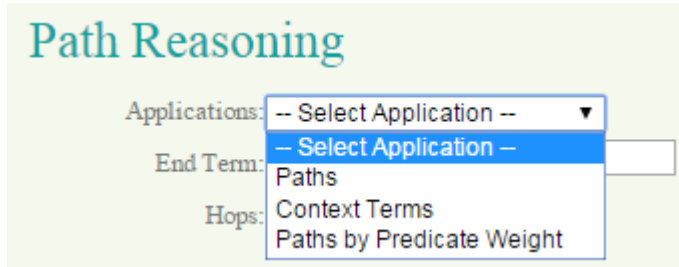
Enter a **Search/Start Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here. Although some relationships are unique simply because they are invalid, hidden relationships that are valid can also be discovered.



The results can be seen in either a table (Tabulate) or graph (Visualize), as explained on the next page.

3. Path Reasoning

Path Reasoning shows the relationships between two terms. This can reveal unobvious and hidden relationships and can lead to novel discoveries. There are three applications in Term Reasoning.

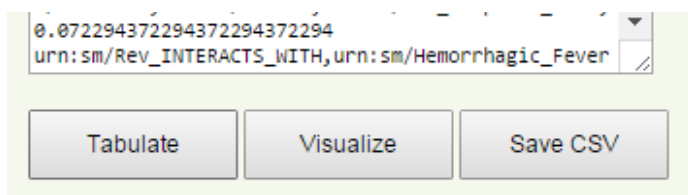


a. Paths

Paths shows the connections between two terms in the informative graph. This information is useful in discovering relationships between close as well as distant terms.

The number of **Hops** determines the number of connections between the two terms. A high number of hops could require several minutes to find.

Enter a **Search/Start Term**, and **End Term**, select the **Size of Results** and **Hops**, and click **Execute**. The results are best seen in either a table (Tabulate) or graph (Visualize).

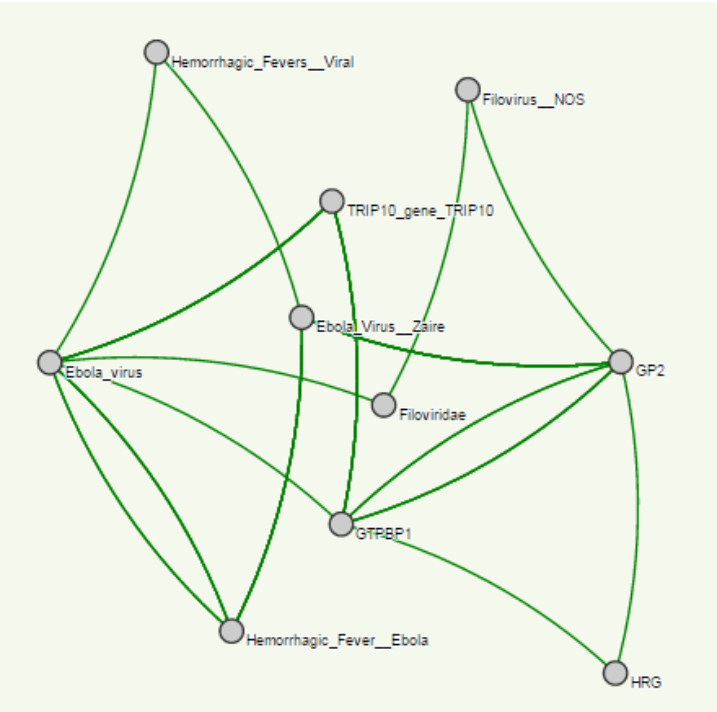


Tabulate

start	p1	x1	p2	x2	p3	end	score
Ebola_virus	NEG_INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.375617
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	compared_with	GP2	0.139754
Ebola_virus	INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.075617
Ebola_virus	Rev_COEXISTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.075617
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_compared_with	GP2	0.072294
Ebola_virus	Rev_INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.063117
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	COEXISTS_WITH	GP2	0.058736
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_COEXISTS_WITH	GP2	0.058203
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_INTERACTS_WITH	GP2	0.058016
Ebola_virus	CAUSES	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.047839
Ebola_virus	ISA	Filoviridae	Rev_ISA	Filovirus__NOS	Rev_PART_OF	GP2	0.023544
Ebola_virus	LOCATION_OF	GTPBP1	STIMULATES	HRG	Rev_STIMULATES	GP2	0.022882
Ebola_virus	Rev_COEXISTS_WITH	Hemorrhagic_Fevers__Viral	PROCESS_OF	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.006522

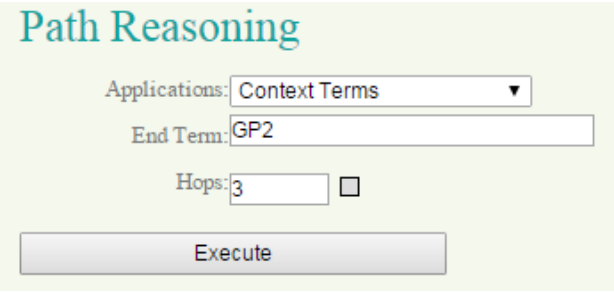
The resulting paths are shown here in a table ordered by descending score. The score is based on the uniqueness of the path, with 1 being the most unique.

Visualize



Visualization shows how terms are interconnected in the results. Following paths, exploring close relationships, observing clusters, and looking for patterns are all easier with visualization.

b. Context Terms

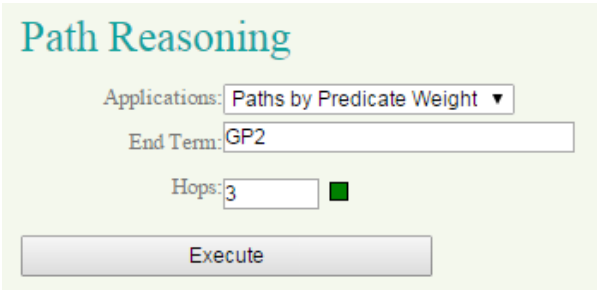


Context Terms looks for the terms that connect to both the **Start Term** and the **End Term**. The common links are an indication of a relationship, and could be used to select terms for additional path searches.

Enter a **Search/Start Term** and **End Term**, select the **Size of Results**, and click **Execute**. The Results will appear as shown here. The terms shown will have a direct link to both the **Search/Start Term** and the **End Term**.

[The results are only the terms and counts, so the Tabulate and Visualize buttons shown at the bottom of the results are not available.]

c. Paths by Predicate Weight



Paths by Predicate Weight is similar to **Paths**, except the score is based on how important the predicates have shown to be upon review. This method can produce results that are more reliable.

The number of **Hops** determines the number of connections between the two terms. A high number of hops could require several minutes to find.

Enter a **Search/Start Term**, and **End Term**, select the **Size of Results** and **Hops**, and click **Execute**. The results are best seen in either a table (Tabulate) or graph (Visualize).

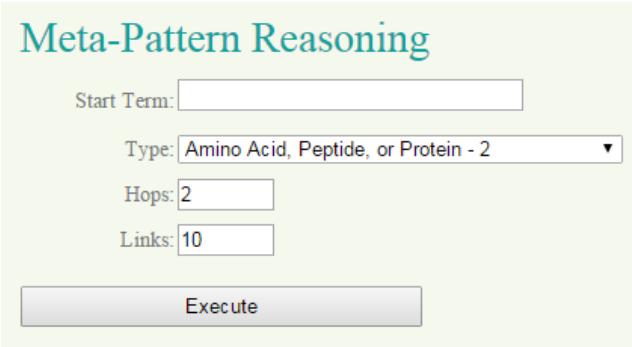
Tabulate

start	p1	x1	p2	x2	p3	end	score
Ebola_virus	NEG_INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	325017
Ebola_virus	Rev_INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	265923
Ebola_virus	ISA	Filoviridae	Rev_ISA	Filovirus__NOS	Rev_PART_OF	GP2	248570
Ebola_virus	Rev_COEXISTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	240597
Ebola_virus	CAUSES	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	185724
Ebola_virus	INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	177282
Ebola_virus	LOCATION_OF	GTPBP1	STIMULATES	HRG	Rev_STIMULATES	GP2	145266
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_INTERACTS_WITH	GP2	92862
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_COEXISTS_WITH	GP2	84018
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	COEXISTS_WITH	GP2	69278
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	compared_with	GP2	56012
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_compared_with	GP2	48642
Ebola_virus	Rev_COEXISTS_WITH	Hemorrhagic_Fevers__Viral	PROCESS_OF	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0

Based on an analysis of prior results, each predicate was scored. The scores are multiplied together to determine the score for the triple and then the triples are presented in descending order. Some predicates have a score of zero, indicating they are not helpful at all. The scores are maintained in a table that can be updated to provide more informative paths.

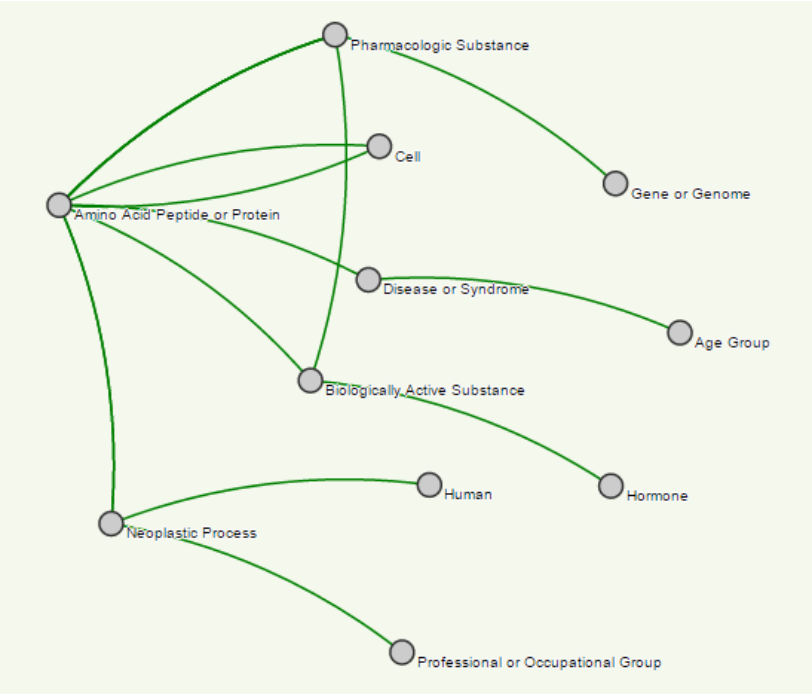
4. Meta-Pattern Reasoning

Meta-Pattern Reasoning generates paths between types based on the frequency of occurrence. This provides a view of typical paths that can be found in the graph. Each type-predicate-type is selected at random based on how often a type is connected through a predicate to another type. The triples are then connected to form a path.

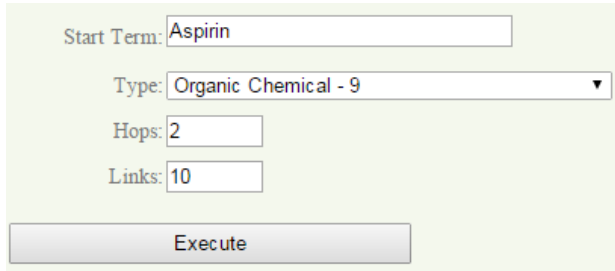


Select a starting **Type**, **Hops**, and **Links**; and click Execute. The results can be shown in a table or graph.

Amino Acid Peptide or Protein	INTERACTS_WITH	Amino Acid Peptide or Protein	STIMULATES	Amino Acid Peptide or Protein	273485
Amino Acid Peptide or Protein	INTERACTS_WITH	Amino Acid Peptide or Protein	COEXISTS_WITH	Pharmacologic Substance	205257
Amino Acid Peptide or Protein	PART_OF	Cell	LOCATION_OF	Amino Acid Peptide or Protein	180763
Amino Acid Peptide or Protein	COEXISTS_WITH	Amino Acid Peptide or Protein	INHIBITS	Pharmacologic Substance	119852
Amino Acid Peptide or Protein	INTERACTS_WITH	Pharmacologic Substance	compared_with	Gene or Genome	75146
Amino Acid Peptide or Protein	ASSOCIATED_WITH	Neoplastic Process	OCCURS_IN	Human	61869
Amino Acid Peptide or Protein	ASSOCIATED_WITH	Neoplastic Process	PROCESS_OF	Professional or Occupational Group	60480
Amino Acid Peptide or Protein	COEXISTS_WITH	Biologically Active Substance	STIMULATES	Hormone	46070
Amino Acid Peptide or Protein	compared_with	Pharmacologic Substance	INTERACTS_WITH	Biologically Active Substance	43704
Amino Acid Peptide or Protein	CAUSES	Disease or Syndrome	PROCESS_OF	Age Group	37158



A **Start Term** may also be selected, and the **Type** will be determined. The pattern will be based on the **Start Term**.

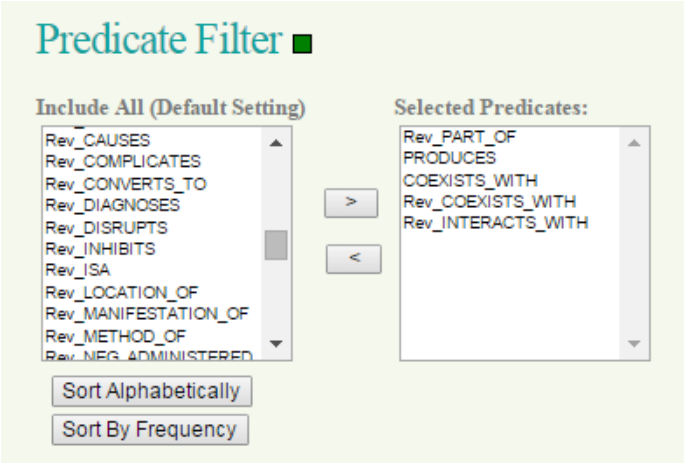


Aspirin	CAUSES	Pathologic Function	COEXISTS_WITH	Disease or Syndrome	21446
Aspirin	PREVENTS	Sign or Symptom	ASSOCIATED_WITH	Disease or Syndrome	13287
Aspirin	Rev_higher_than	Pharmacologic Substance	AFFECTS	Organ or Tissue Function	12558
Aspirin	NEG_AFFECTS	Disease or Syndrome	CAUSES	Sign or Symptom	6906
Aspirin	PREVENTS	Neoplastic Process	PROCESS_OF	Population Group	6209
Aspirin	INHIBITS	Gene or Genome	PART_OF	Family Group	4571
Aspirin	AFFECTS	Physiologic Function	ASSOCIATED_WITH	Pharmacologic Substance	2045
Aspirin	STIMULATES	Eicosanoid	INTERACTS_WITH	Amino Acid Peptide or Protein	1527
Aspirin	Rev_compared_with	Organic Chemical	STIMULATES	Steroid	1291
Aspirin	NEG_PREDISPOSES	Pathologic Function	PRODUCES	Biologically Active Substance	841

5. The Predicate Filter and Context Filter

Sometimes the results shown include information that is not helpful. It is possible to filter the results by either showing only selected predicates, or excluding selected terms.

The Predicate Filter and Context Filter are available when the box next to them is green. In this example, only triples that contain the predicates selected will now appear in the results.



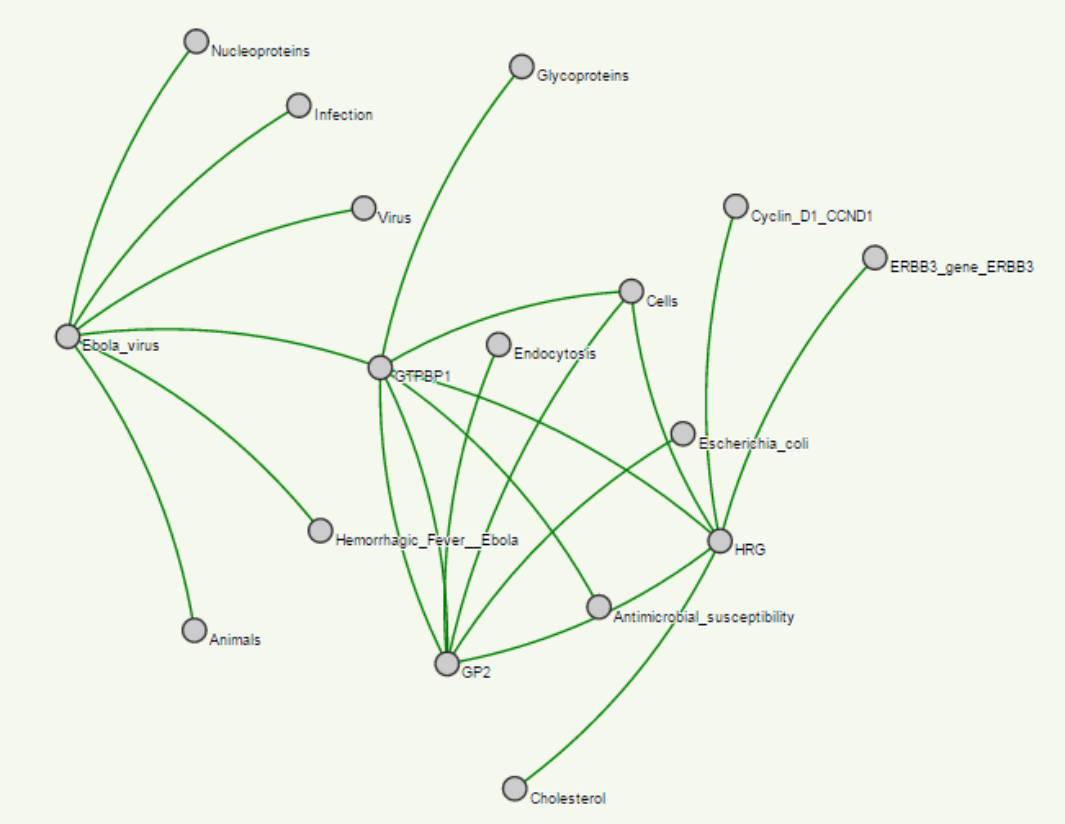
start	p1	x1	p2	x2	p3	end	score
Ebola_virus	Rev_COEXISTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.075617
Ebola_virus	Rev_INTERACTS_WITH	Hemorrhagic_Fever__Ebola	Rev_INTERACTS_WITH	Ebola_Virus__Zaire	Rev_PART_OF	GP2	0.063117
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	COEXISTS_WITH	GP2	0.058736
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_COEXISTS_WITH	GP2	0.058203
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_INTERACTS_WITH	GP2	0.058016

The Context Filter will remove all results containing the terms entered. As an example, all results that contain *Ebola_Virus__Zaire* will be eliminated if entered in the Context Filter as shown. Multiple terms may be entered.

start	p1	x1	p2	x2	p3	end	score
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	compared_with	GP2	0.139754
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_compared_with	GP2	0.072294
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	COEXISTS_WITH	GP2	0.058736
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_COEXISTS_WITH	GP2	0.058203
Ebola_virus	Rev_PART_OF	TRIP10_gene_TRIP10	PRODUCES	GTPBP1	Rev_INTERACTS_WITH	GP2	0.058016
Ebola_virus	ISA	Filoviridae	Rev_ISA	Filovirus__NOS	Rev_PART_OF	GP2	0.023544
Ebola_virus	LOCATION_OF	GTPBP1	STIMULATES	HRG	Rev_STIMULATES	GP2	0.022882

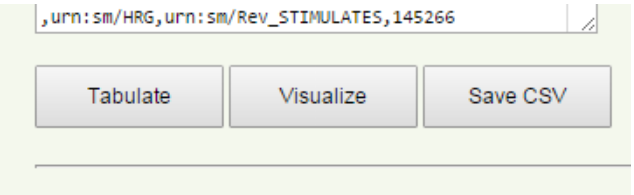
6. Rows

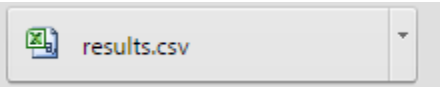
Clicking on a row in a table will show terms that are descriptive of the terms in the row clicked. The results are returned as pairs of related terms. Click **Visualize** to see a graph of the descriptive terms:



7. Save CSV

Save CSV will download the results shown in the Results textbox to a file in CSV format. Depending on your browser settings the file will automatically be saved on your computer in a file named results.csv.





The file can be directly loaded into Excel, or other software programs.

```
p1,x1,p2,x2,p3,avgScore
urn:sm/NEG_INTERACTS_WITH,urn:sm/Hemorrhagic_Fever__Ebola,urn
urn:sm/Rev_INTERACTS_WITH,urn:sm/Hemorrhagic_Fever__Ebola,urn
urn:sm/ISA,urn:sm/Filoviridae,urn:sm/Rev_ISA,urn:sm/Filovirus
urn:sm/Rev_COEXISTS_WITH,urn:sm/Hemorrhagic_Fever__Ebola,urn:
urn:sm/CAUSES,urn:sm/Hemorrhagic_Fever__Ebola,urn:sm/Rev_INTE
urn:sm/INTERACTS_WITH,urn:sm/Hemorrhagic_Fever__Ebola,urn:sm/
urn:sm/LOCATION_OF,urn:sm/GTPBP1,urn:sm/STIMULATES,urn:sm/HRG
urn:sm/Rev_PART_OF,urn:sm/TRIP10_gene_TRIP10,urn:sm/PRODUCES,
urn:sm/Rev_PART_OF,urn:sm/TRIP10_gene_TRIP10,urn:sm/PRODUCES,
urn:sm/Rev_PART_OF,urn:sm/TRIP10_gene_TRIP10,urn:sm/PRODUCES,
urn:sm/Rev_PART_OF,urn:sm/TRIP10_gene_TRIP10,urn:sm/PRODUCES,
urn:sm/Rev_PART_OF,urn:sm/TRIP10_gene_TRIP10,urn:sm/PRODUCES,
urn:sm/Rev_COEXISTS_WITH,urn:sm/Hemorrhagic_Fevers__Viral,urn
```