

MINING MULTIPLE PATHOGEN–HOST PROTEIN INTERACTOMES FOR THE DETECTION OF SHARED PATTERNS

A DIVE INTO PROTEIN–PROTEIN INTERACTION DATABASES

Pieter Moris, Pieter Meysman & Kris Laukens

December 14th, 2017





HOST-PATHOGEN INTERACTIONS

The arms race between pathogens and their hosts is governed by molecular interactions.

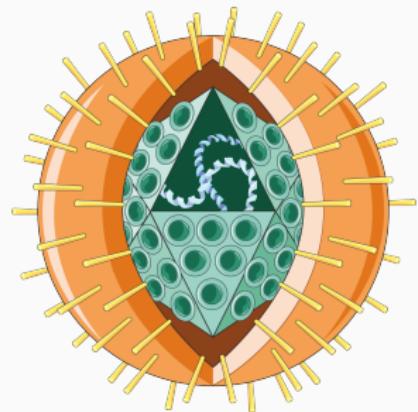
The arms race between pathogens and their hosts is governed by molecular interactions.

- Data collection efforts & public repositories (e.g. IntAct)
- Predominantly protein–protein interactions (PPIs)
- Various detection methodologies: experimental and computational

CAN WE DISTILL NEW INSIGHTS FROM
HOST-PATHOGEN PROTEIN-PROTEIN INTERACTION
NETWORKS AND THEIR PUBLIC ANNOTATION DATA?

A CASE STUDY ON THE *HERPESVIRIDAE*

- DNA viruses
- Latent and lytic phases
- Human (9) and animal hosts (~130)
- Examples:
 - herpes simplex viruses 1 & 2
 - varicella-zoster virus
 - Epstein-Barr virus



A PRIMER ON FREQUENT ITEMSET MINING

TRANSACTIONS, ITEMSETS AND FREQUENCIES

Imagine you are the owner of an online store...

TRANSACTIONS, ITEMSETS AND FREQUENCIES

~~Imagine you are the owner of an online store...~~

To avoid an overused example, instead imagine you are **Santa** and you've got access to the wishing lists of thousands of bioinformaticians.

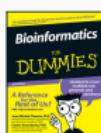
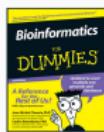
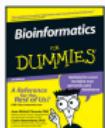
TRANSACTIONS, ITEMSETS AND FREQUENCIES

~~Imagine you are the owner of an online store...~~

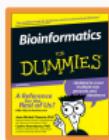
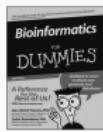
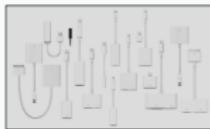
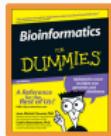
To avoid an overused example, instead imagine you are **Santa** and you've got access to the wishing lists of thousands of bioinformaticians.

Are there specific patterns or sets of items that co-occur frequently in the wishing lists?

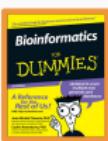
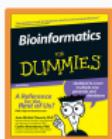
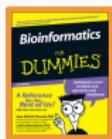
FREQUENT ITEMSETS PASS A CERTAIN SUPPORT THRESHOLD



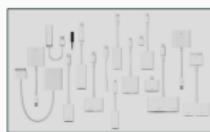
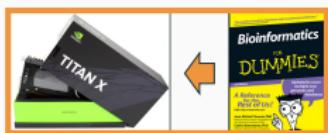
FREQUENT ITEMSETS PASS A CERTAIN SUPPORT THRESHOLD



ASSOCIATION RULES DESCRIBE HOW ITEMS RELATE TO ONE ANOTHER



ASSOCIATION RULES DESCRIBE HOW ITEMS RELATE TO ONE ANOTHER



THE NITTY-GRITTY: DATA WRANGLING

COLLECTION OF PPI NETWORKS AND ANNOTATIONS

Protein–protein interaction resources

- IntAct
- Host–Pathogen Interaction Database 2.0
- VirHostNet 2.0
- PHISTO

COLLECTION OF PPI NETWORKS AND ANNOTATIONS

Protein–protein interaction resources

- IntAct
- Host–Pathogen Interaction Database 2.0
- VirHostNet 2.0
- PHISTO

→ 47 herpes viruses, 19 hosts, 3422 proteins and 8248 interactions

COLLECTION OF PPI NETWORKS AND ANNOTATIONS

Protein–protein interaction resources

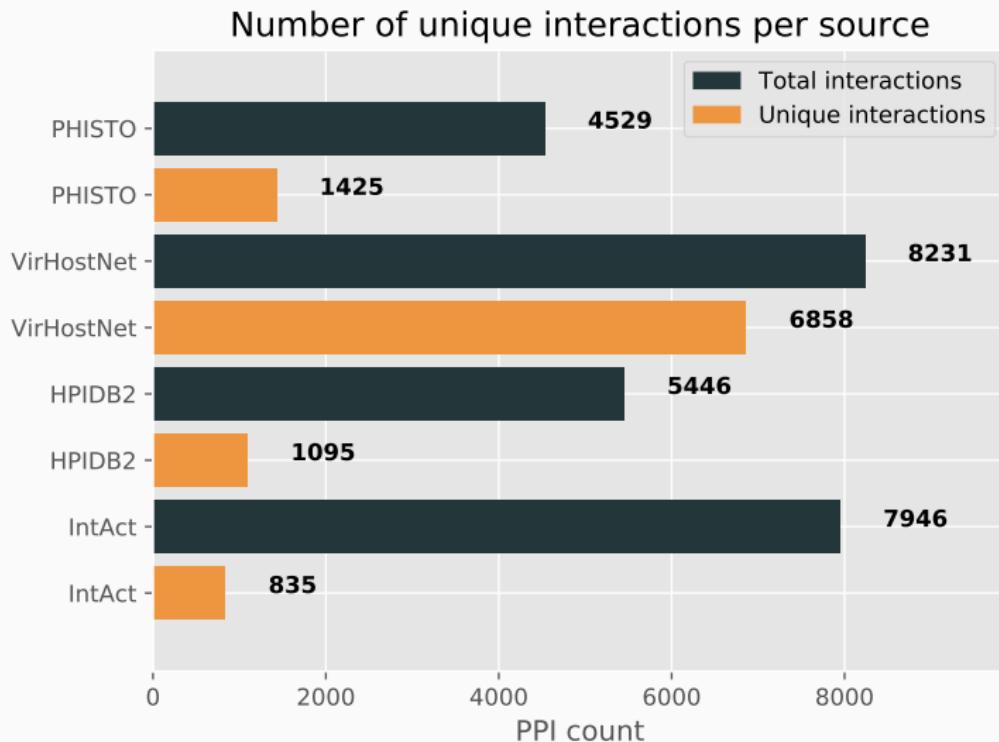
- IntAct
- Host–Pathogen Interaction Database 2.0
- VirHostNet 2.0
- PHISTO

→ 47 herpes viruses, 19 hosts, 3422 proteins and 8248 interactions

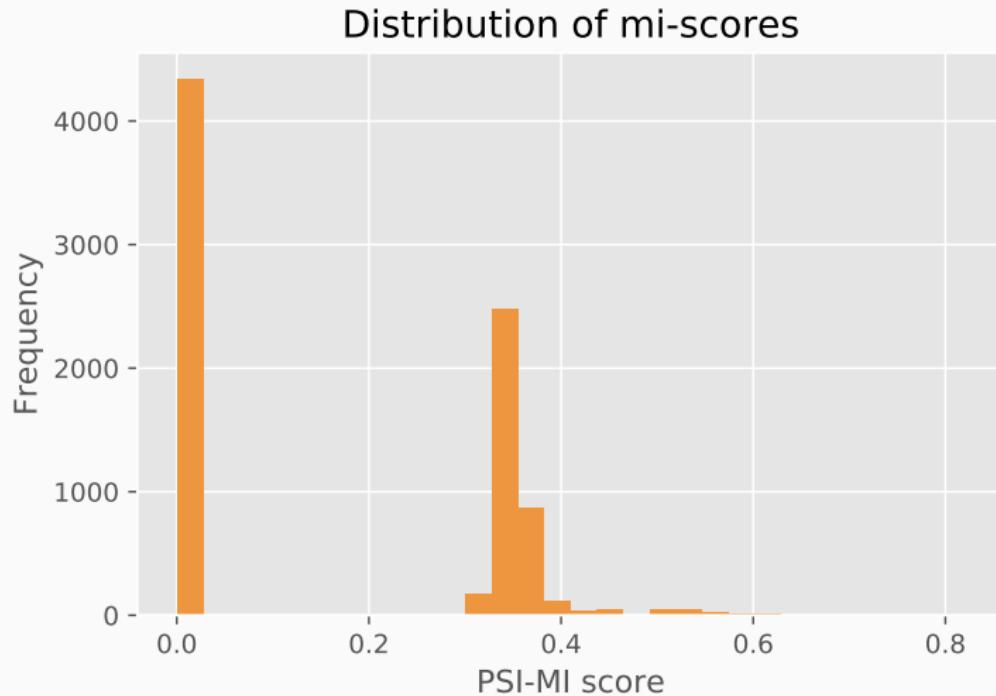
Annotation labels

- Gene Ontology: *biological process, molecular function and cellular component*
- InterPro: *homologous superfamilies, families, domains, repeats and sites*

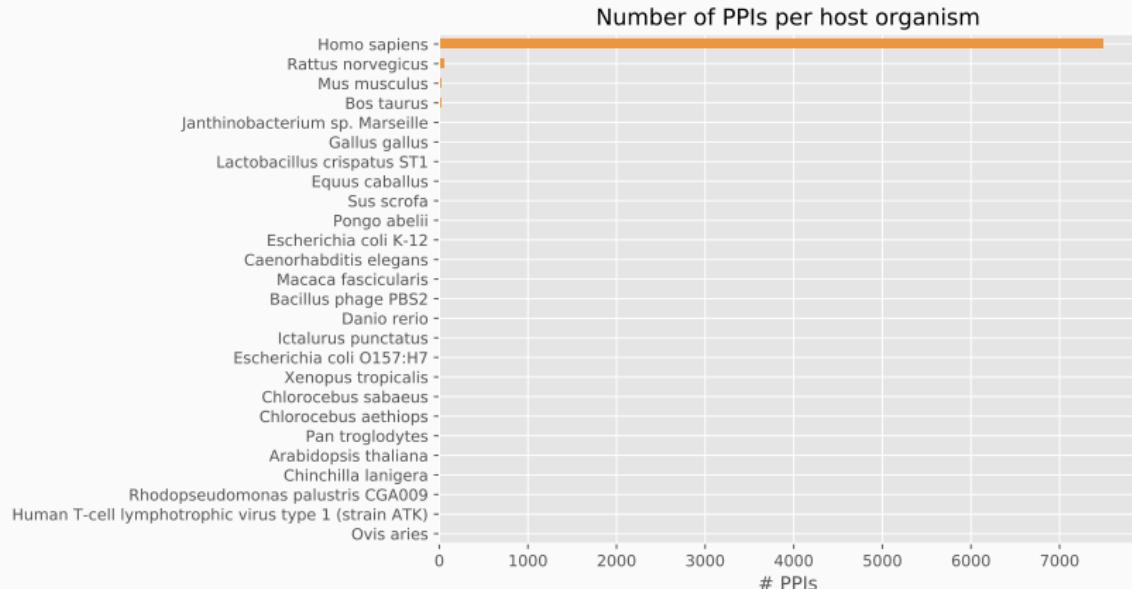
DESPITE EFFORTS TO STREAMLINE PPI DATABASES, THEY EACH STILL CONTAIN UNIQUE PPIs



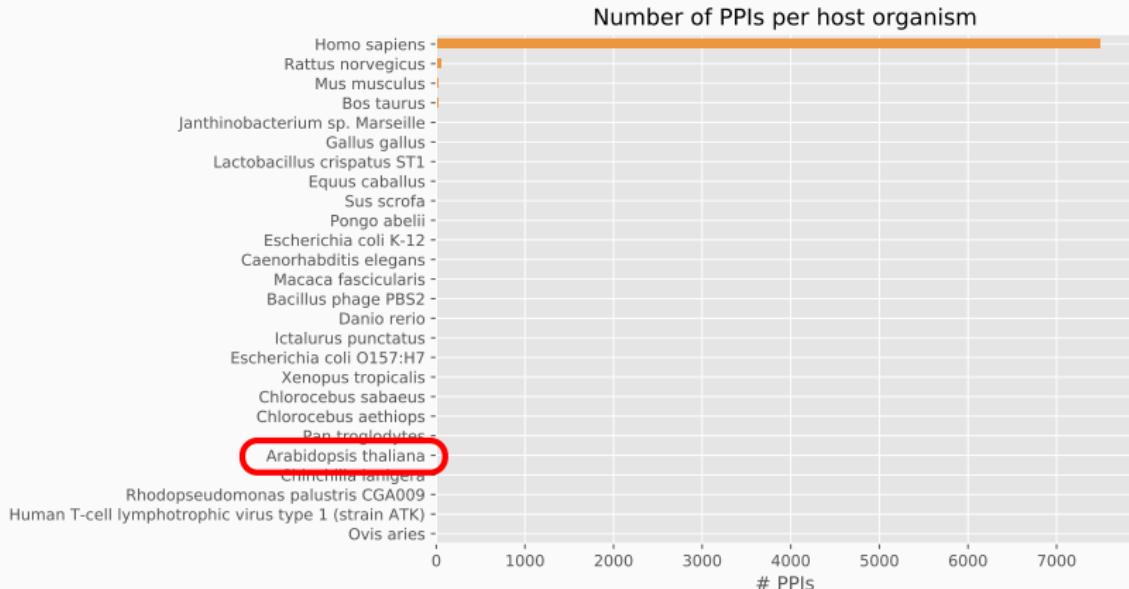
ONLY 40% OF INTERACTIONS WERE ASSIGNED A CONFIDENCE SCORE



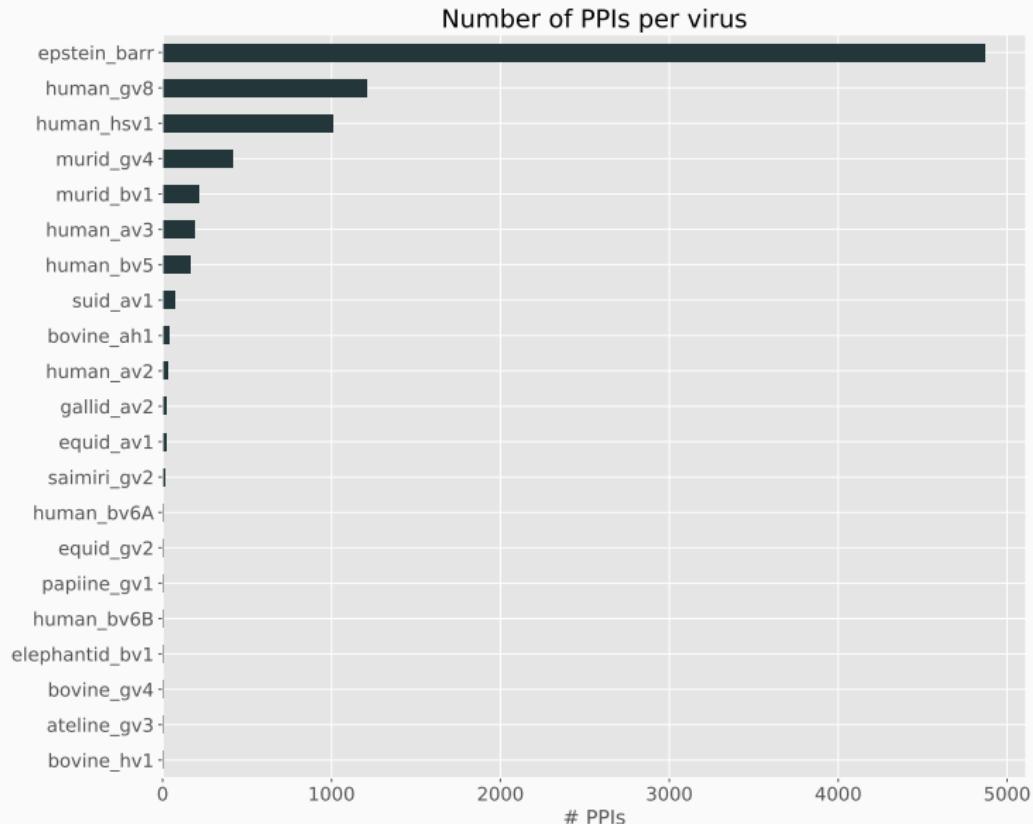
HUMAN INTERACTIONS ARE OVER-REPRESENTED



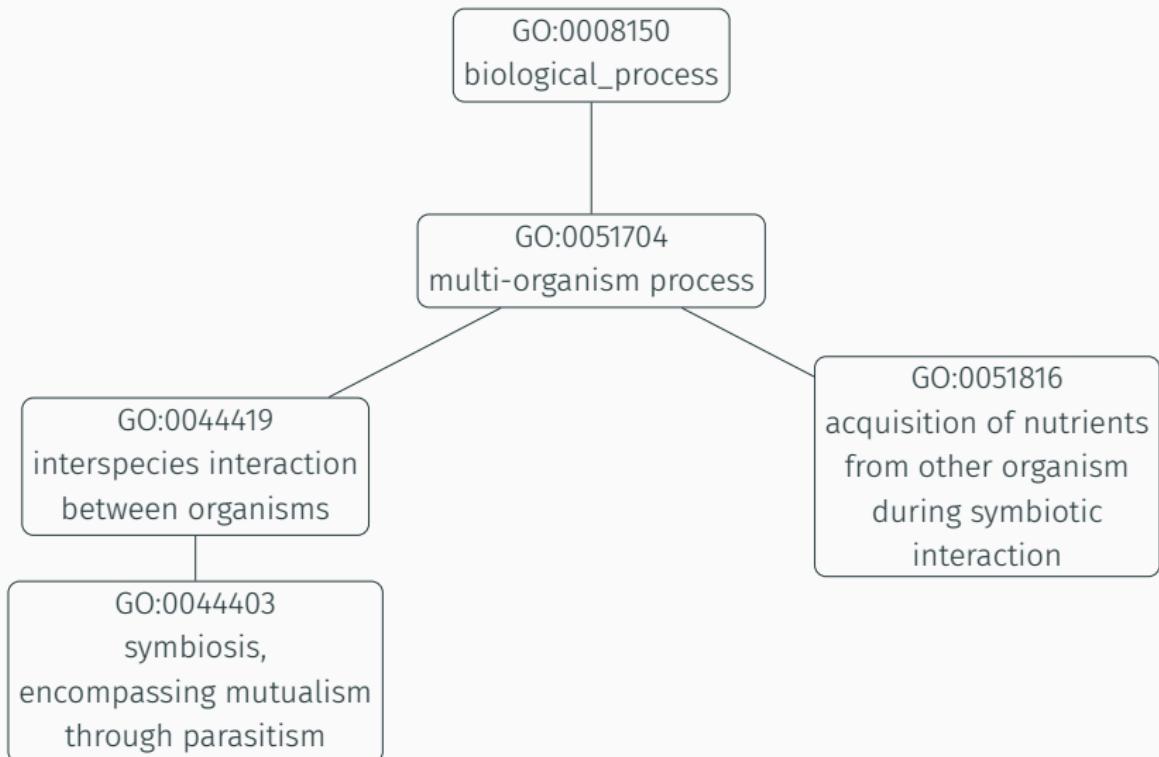
HUMAN INTERACTIONS ARE OVER-REPRESENTED



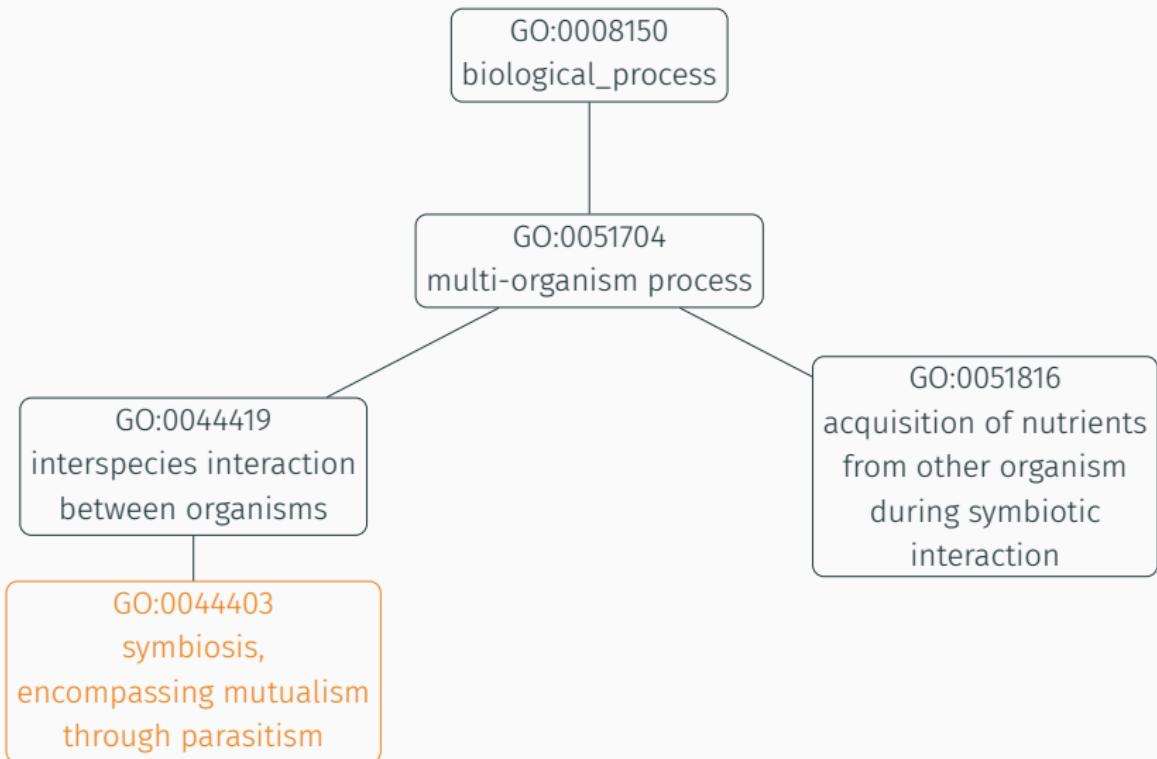
HUMAN VIRUSES HAVE BEEN STUDIED MOST THOROUGHLY



REMAPPING OF GO TERMS TO NORMALIZE INFORMATION CONTENT



REMAPPING OF GO TERMS TO NORMALIZE INFORMATION CONTENT



THE TRANSACTION DATASET: INTERACTION PAIRS & ANNOTATION ITEMS

host protein	viral protein	annotations
uniprotkb:P24941	uniprotkb:Q77Q36	h@GO:0019222,h@GO:0003824,v@GO:0044763,h@IPR011009, ...
uniprotkb:P11802	uniprotkb:Q77Q36	h@GO:0019222,h@GO:0003824,h@GO:0097159
uniprotkb:P63244	uniprotkb:P03206	h@GO:0040012,h@GO:0044444,v@GO:0019046,h@IPR015943, ...
uniprotkb:Q9UBU9	uniprotkb:P10238	h@GO:0043229, v@GO:0048519,v@GO:0030430, v@IPR008648, ...

MINING AND VISUALIZING ASSOCIATION RULES

RETRIEVED ASSOCIATION RULES

- Directional rules (host ↔ virus)
- Remapped to pairwise rules

antecedent	consequent	count	mean confidence	mean lift
V-viral budding	H-membrane-bounded organelle	10	0.7224	0.5843
v@G00019042	h@G00006807	9	0.7053	0.3542
v@G00019083	h@G00006807	1	0.7011	0.3522
v@G00044094	h@G00006807	23	0.7441	0.3735

FROM THIS...



...TO THIS



PAIRWISE SUB-RULES COVER THE BROAD STAGES OF VIRAL INFECTION



PAIRWISE SUB-RULES COVER THE BROAD STAGES OF VIRAL INFECTION



TOP LEFT CLUSTER

- INVASION OF HOST CELLS
- VIRAL SHEDDING
- HIJACKING OF HOST APOPTOSIS APPARATUS

PAIRWISE SUB-RULES COVER THE BROAD STAGES OF VIRAL INFECTION



TOP LEFT CLUSTER

- INVASION OF HOST CELLS
- VIRAL SHEDDING
- HIJACKING OF HOST APOPTOSIS APPARATUS

BOTTOM CLUSTERS

- VIRAL REPLICATION
- LATENT PHASE

WRAPPING UP

FUTURE PROSPECTS

- Incorporating additional **information sources**: e.g. experimental 'omics infection data.
- Formalising the approach to GO remapping to achieve a balanced information content.
- Applying methodology to **other host-pathogens pairs**.

IN CONCLUSION

- Itemset mining can distill prominent patterns from protein–protein interaction networks by describing frequently occurring patterns in annotation labels.
- Translation of expert knowledge into a visual summary.
- Various challenges remain with respect to heterogeneity and the subjectivity of data and annotation filtering.

ACKNOWLEDGMENTS



ADREM
DATA LAB

pmoris
pi_moris



Universiteit
Antwerpen

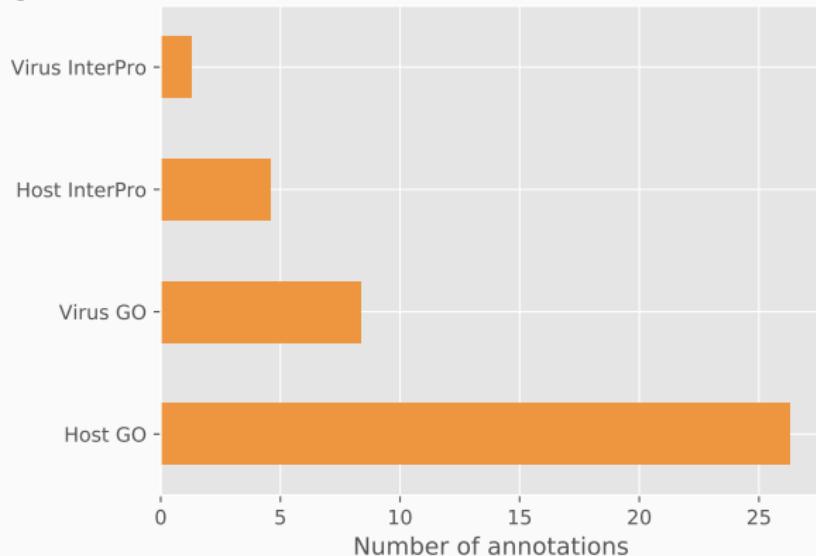


biomina

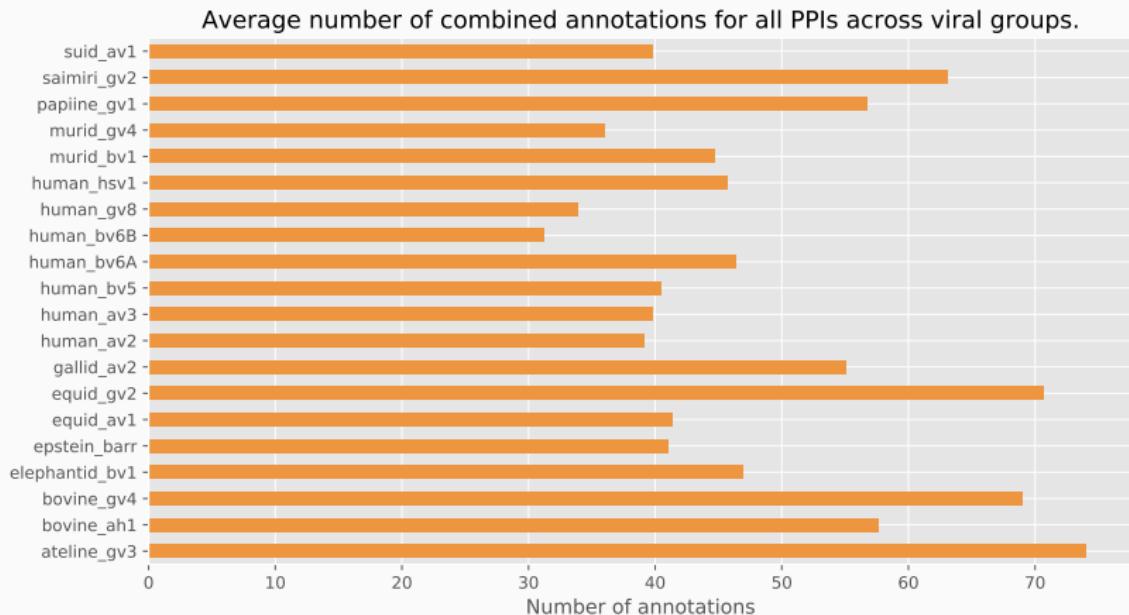
two Opening new horizons

COVERAGE OF THE 4704 PROTEIN ANNOTATIONS

Average number of GO and InterPro annotations for human and viral proteins.



COVERAGE OF THE 4704 PROTEIN ANNOTATIONS

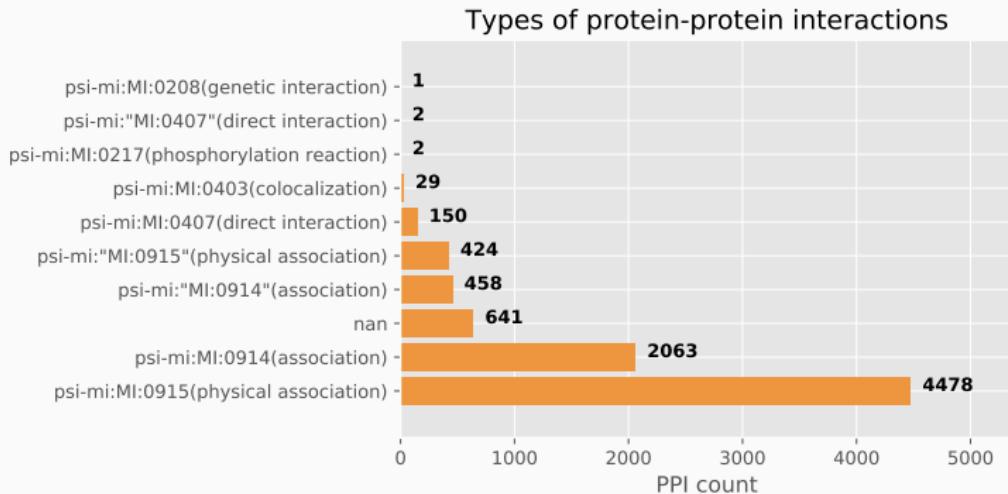


COLLECTION OF PPI NETWORKS

Problems:

- 1:n and 1:0 identifier mappings
- Conflicting file formats
- Incomplete records
- Resolution of taxonomic ID levels
- ...

NO CLEAR WAY TO COMBINE DIFFERENT INTERACTION TYPES



FREQUENT ITEM SET MINING

- In-house algorithm (A Priori-based)
- Support threshold: 0.1
- Confidence threshold: 0.75
- Filtering of uninformative terms
- Creation of pairwise sub-rules while tracking mean and maximum confidence, lift and count of occurrence.