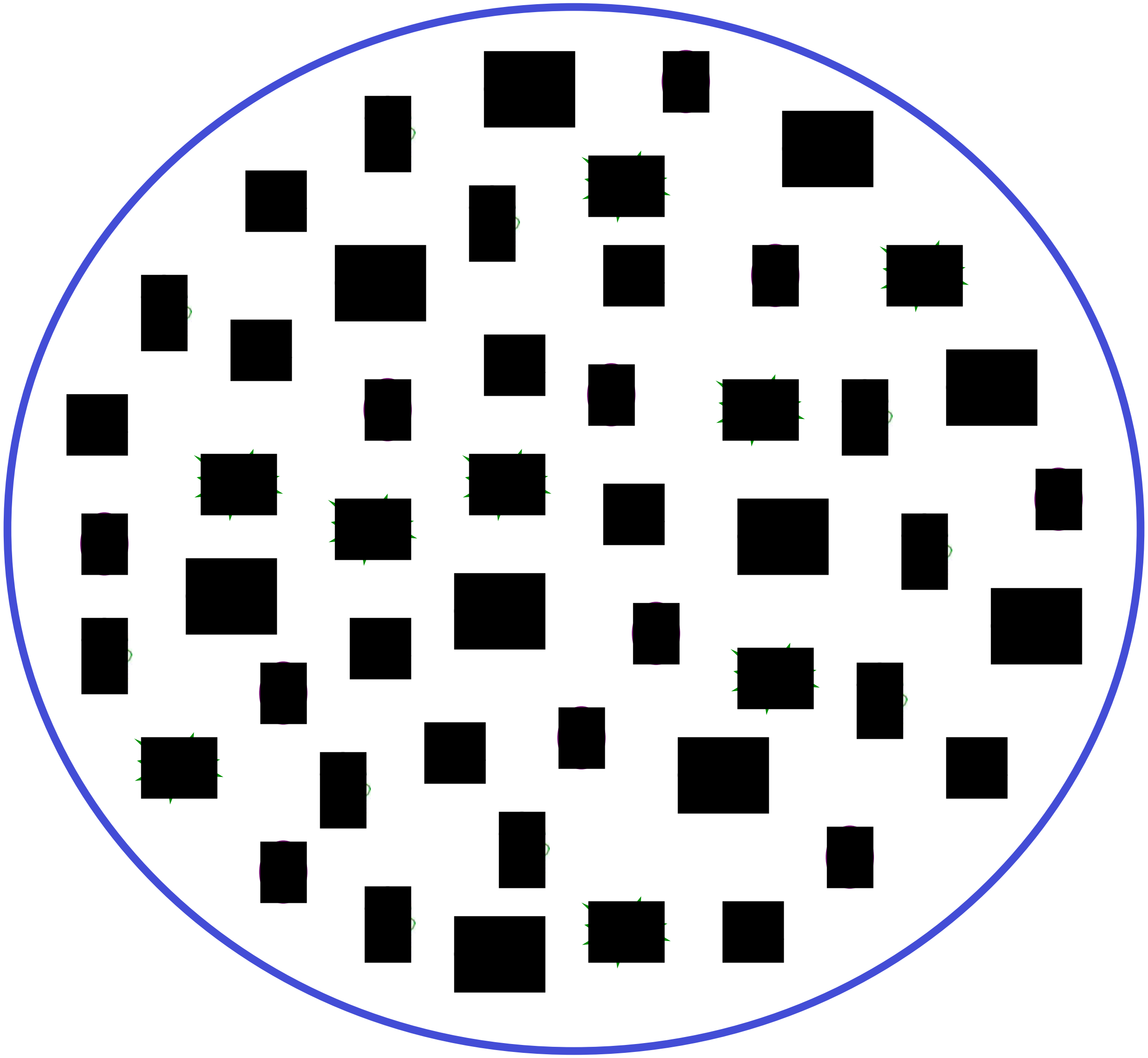
Modelling proteomes

Ram Samudrala
University of Washington

How does the genome of an organism specify its behaviour and characteristics?

Proteome – all proteins of a particular system



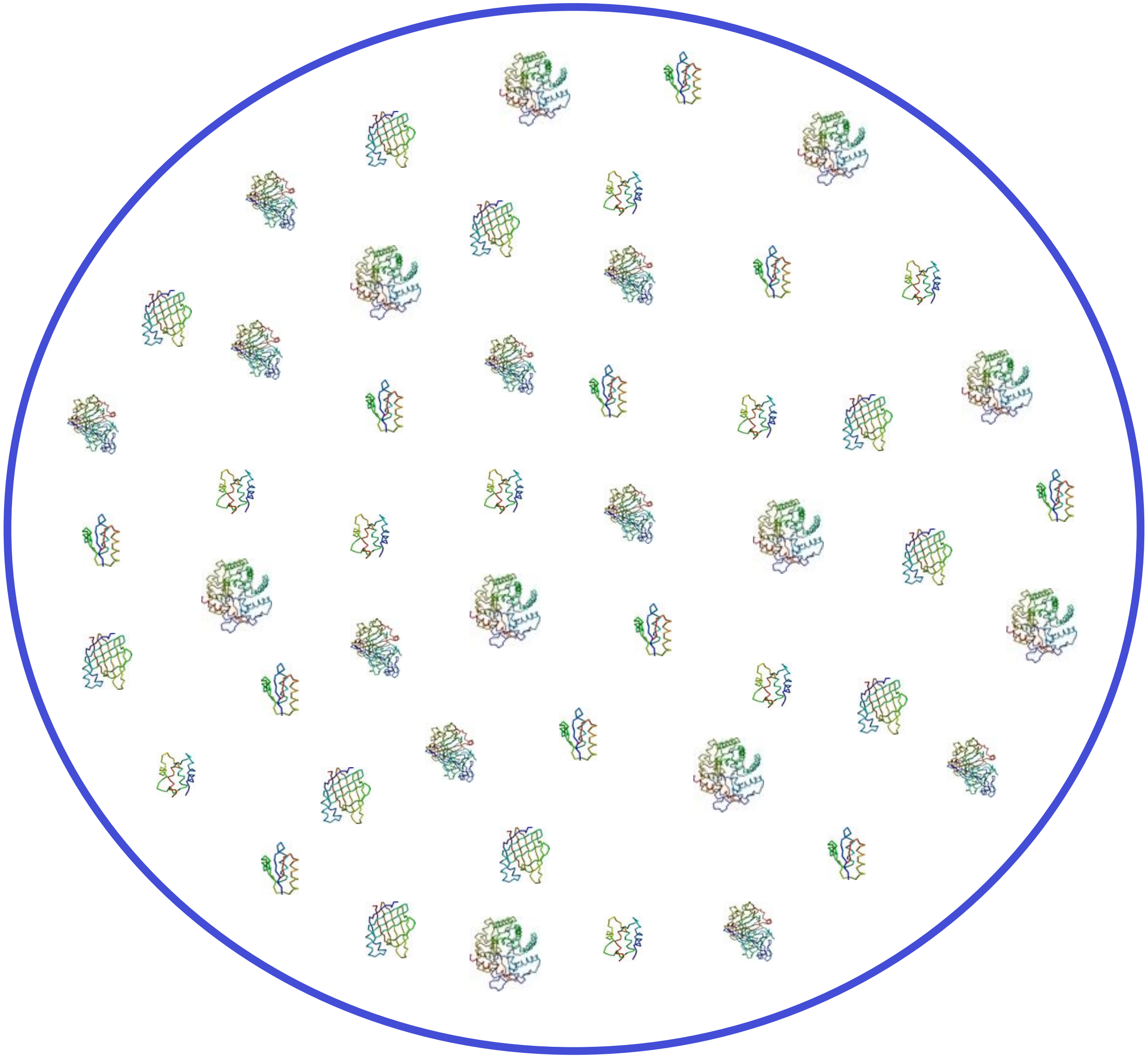
~60,000 in human

~60,000 in rice

~4500 in bacteria like Salmonella and E. coli

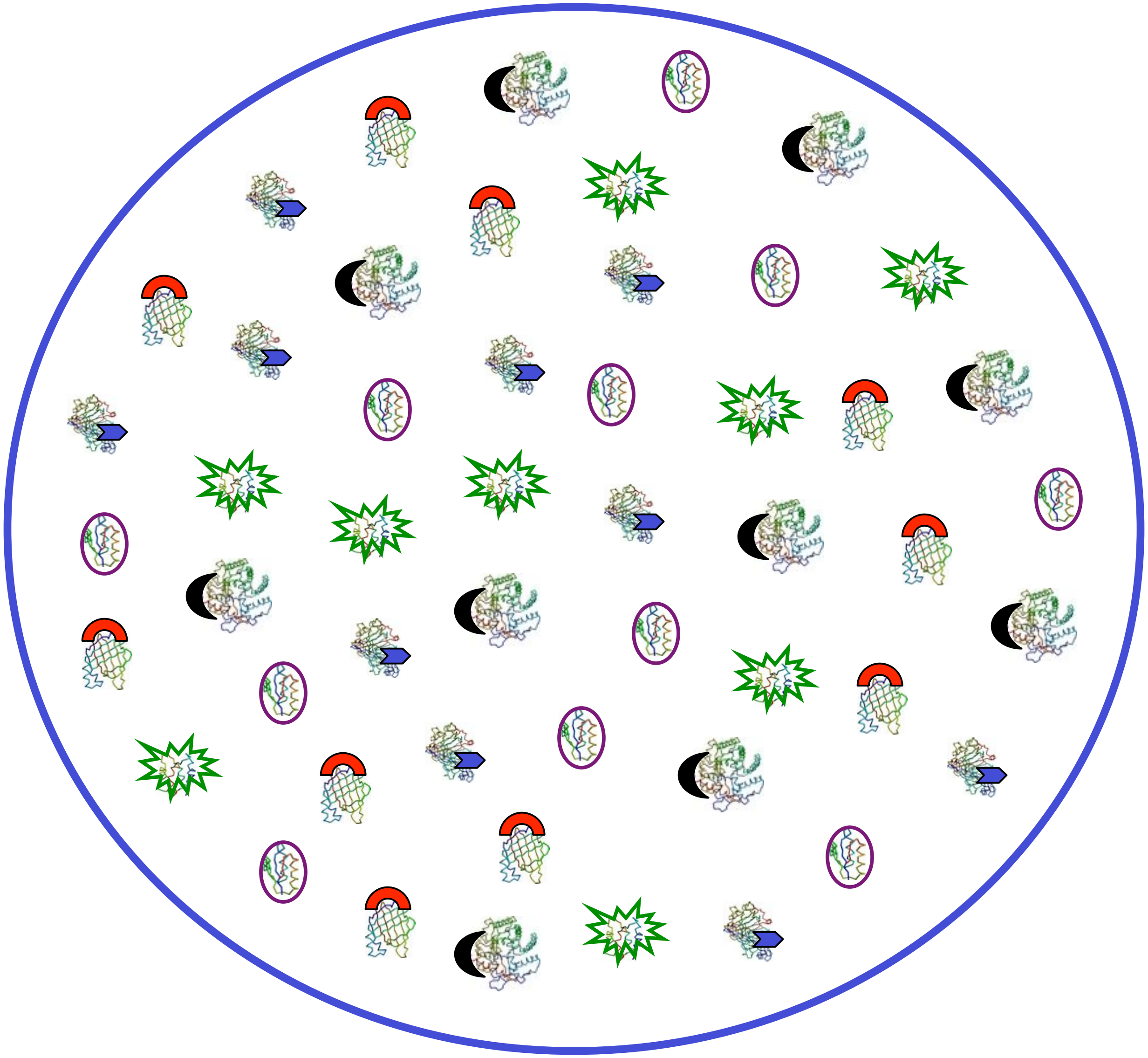
Several thousand distinct sequence families

Modelling proteomes – understand the structure of individual proteins



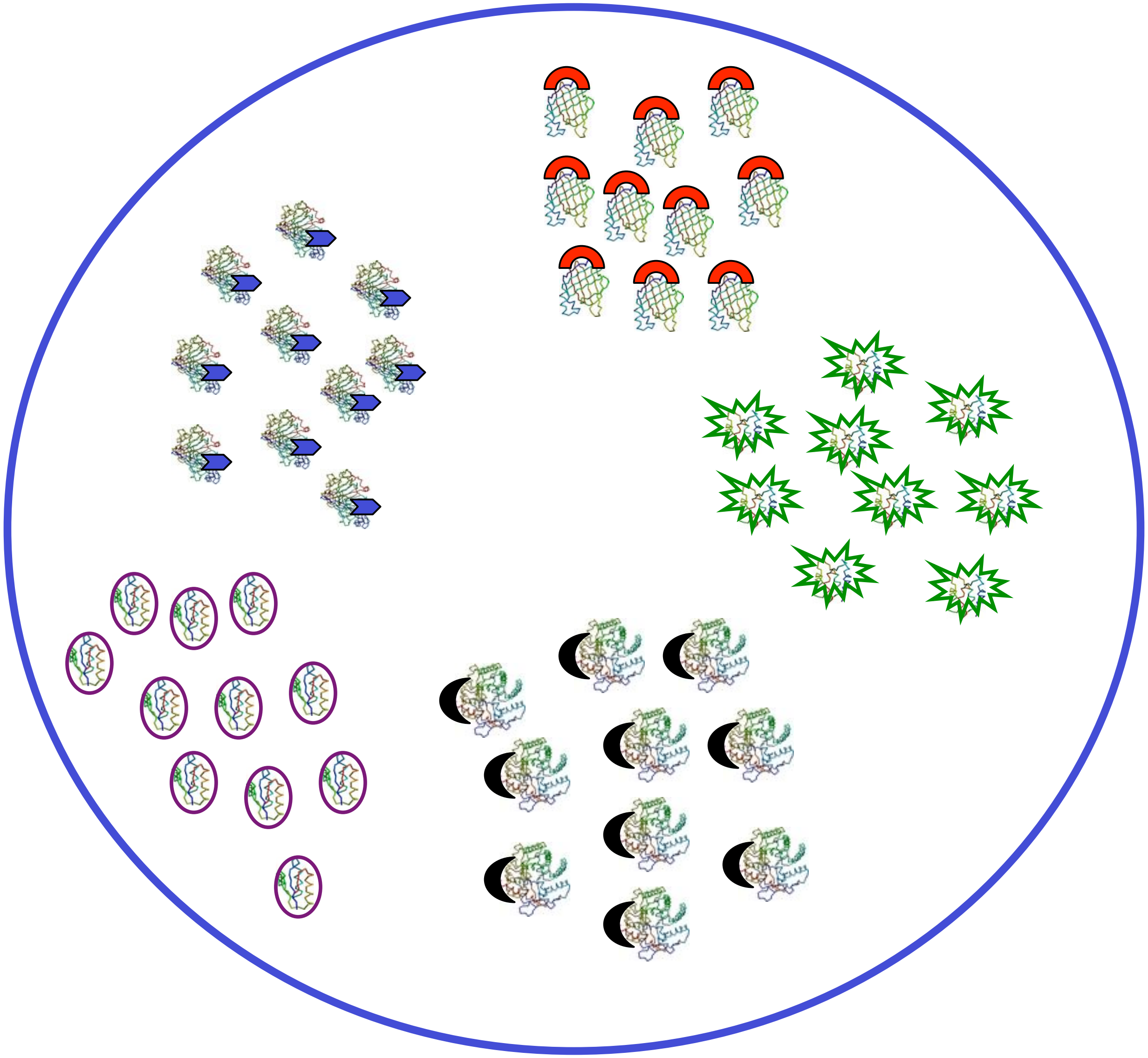
A few thousand distinct structural folds

Modelling proteomes – understand their individual functions



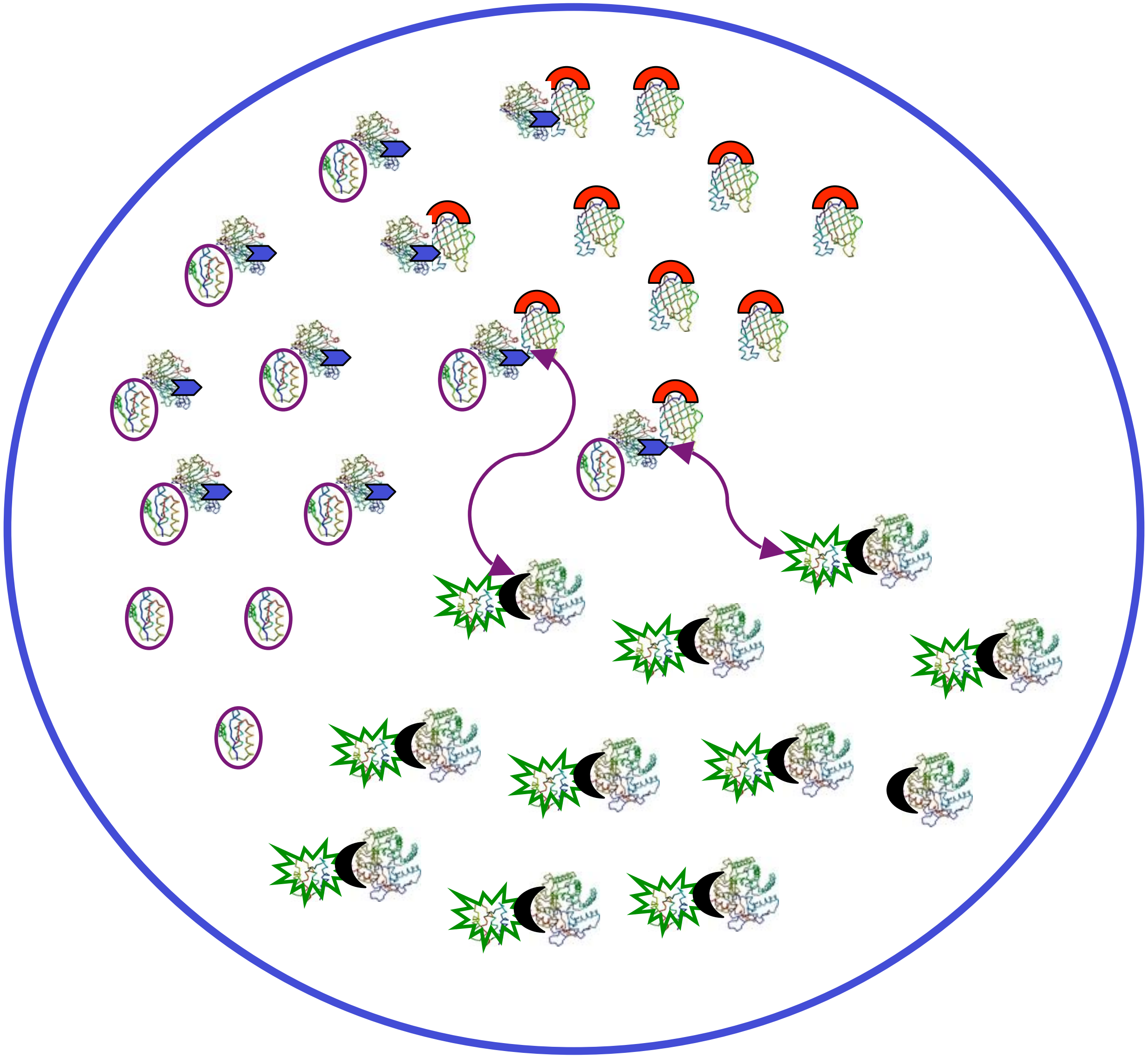
Thousands of possible functions

Modelling proteomes – understand their expression



Different expression
patterns based on
time and location

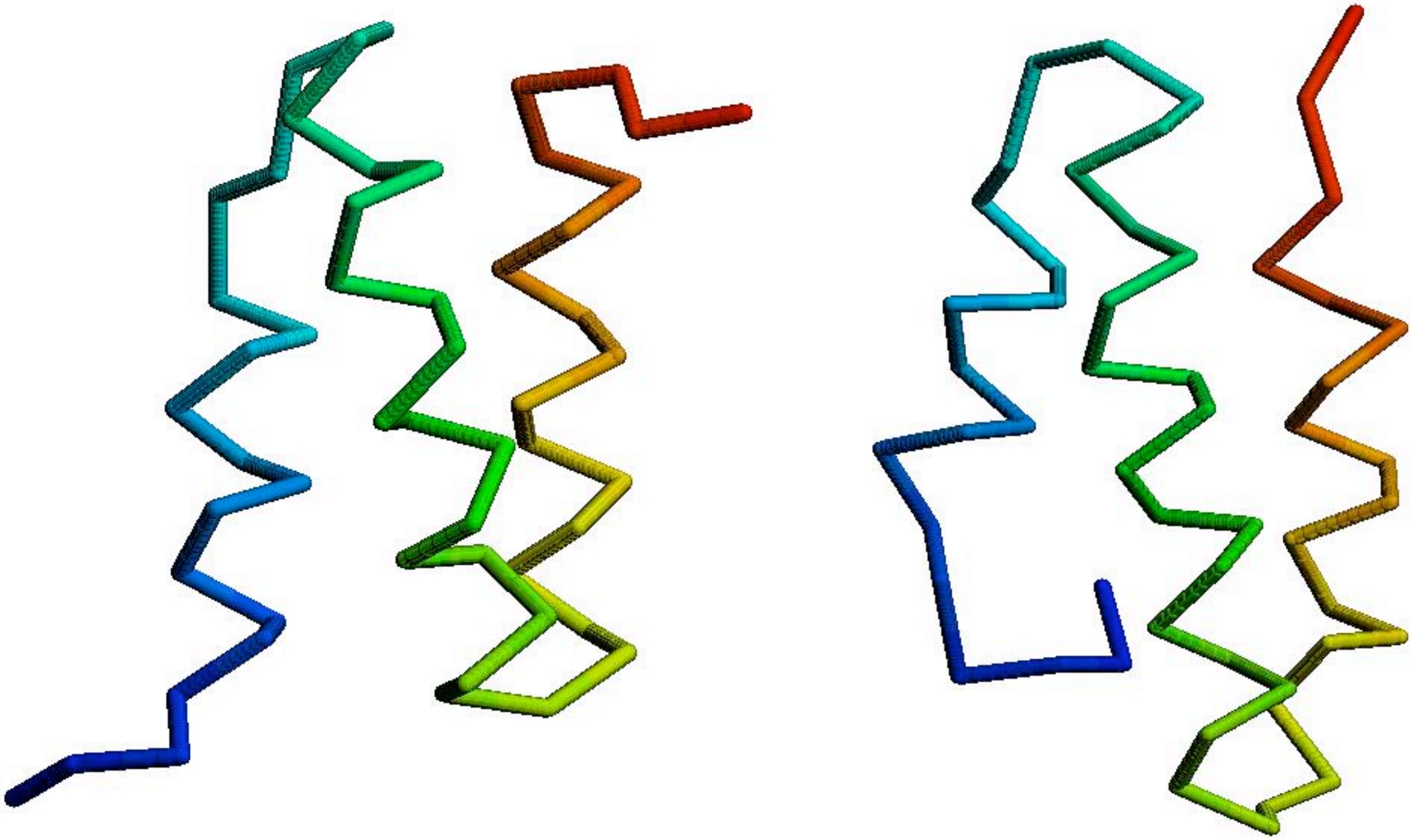
Modelling proteomes – understand their interactions



Interactions and expression patterns are interdependent with structure and function

CASP6 prediction (model1) for T0215

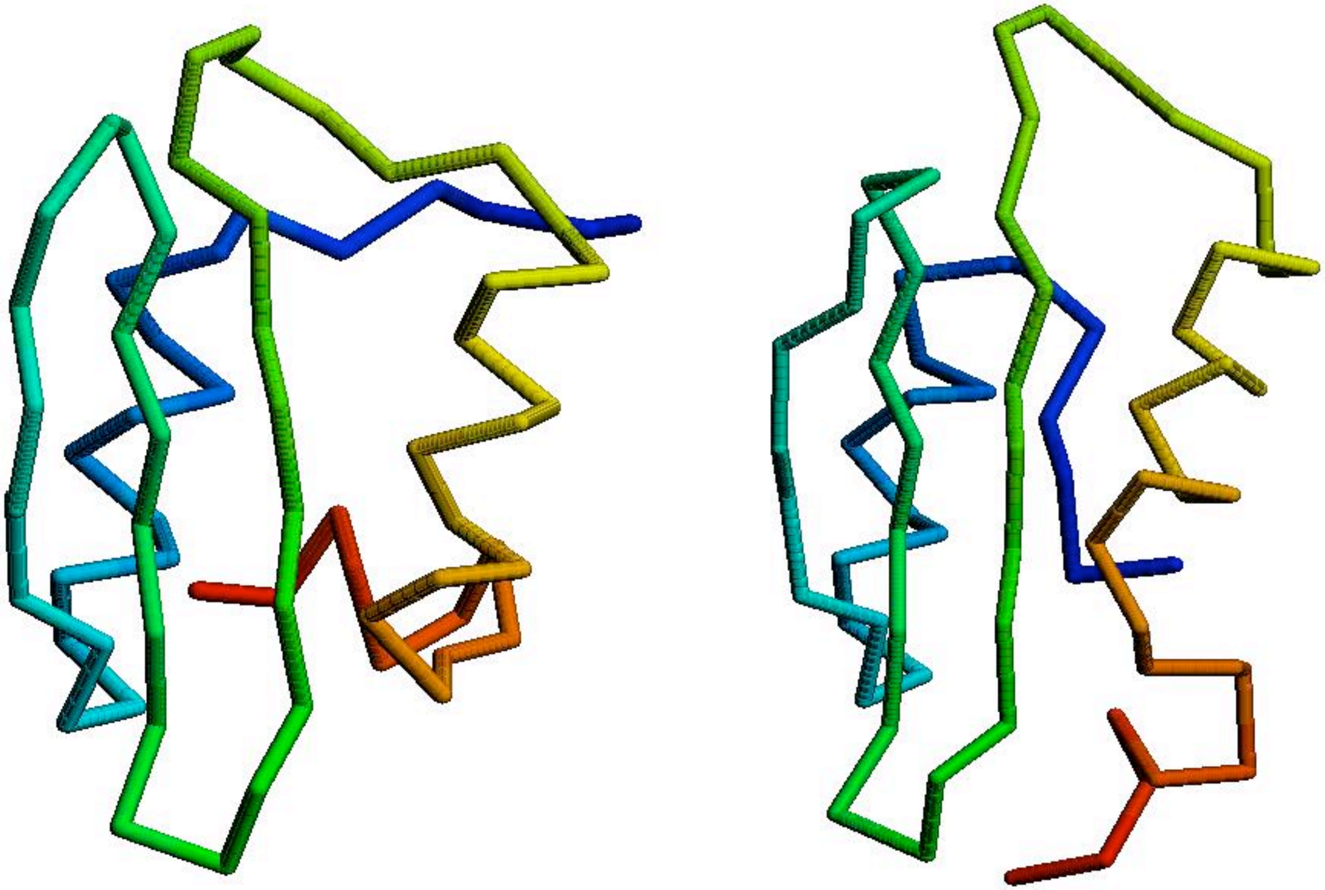
5.0 Å C_α RMSD for all 53 residues



Ling-Hong Hung/Shing-Chung Ngan

CASP6 prediction (model1) for T0281

4.3 Å C_α RMSD for all 70 residues



Ling-Hong Hung/Shing-Chung Ngan

CASP6 prediction (model1) for T0231

1.3 Å C_α RMSD for all 137 residues (80% ID)

Tianyun Liu

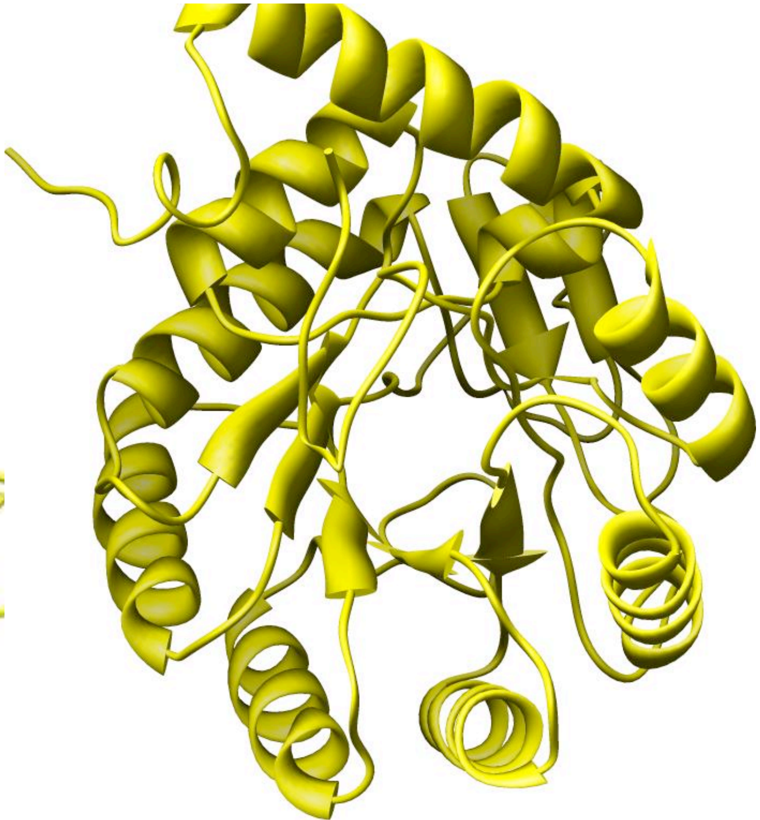
CASP6 prediction (model1) for T0271

2.4 Å C_α RMSD for all 142 residues (46% ID)

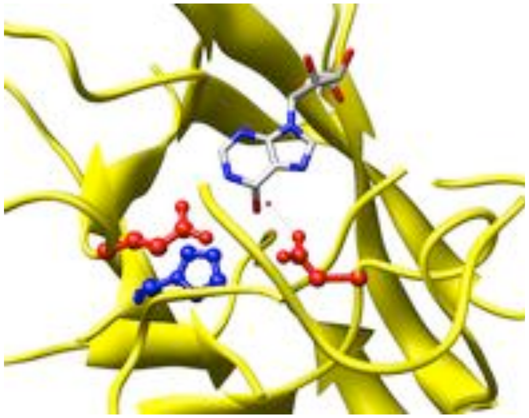
Tianyun Liu

Similar global sequence or structure does not imply similar function

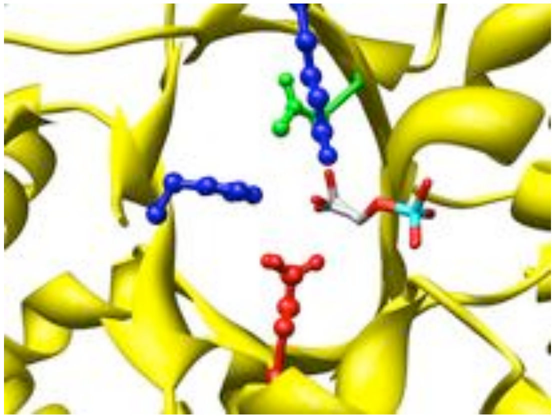
TIM barrel
proteins



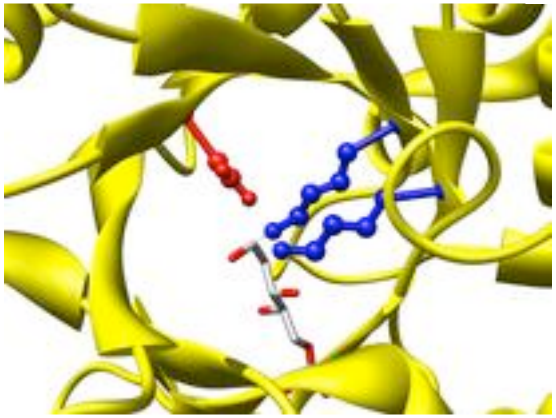
2246 with
known structure



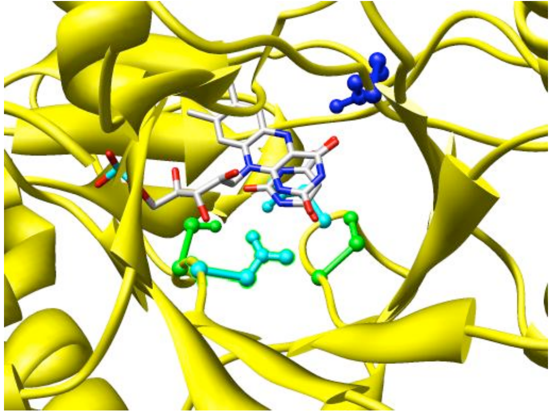
hydrolase



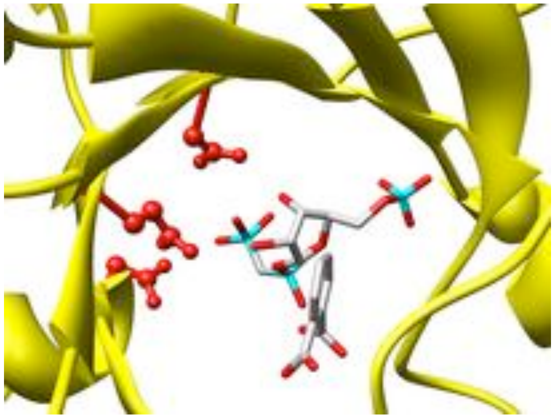
ligase



lyase

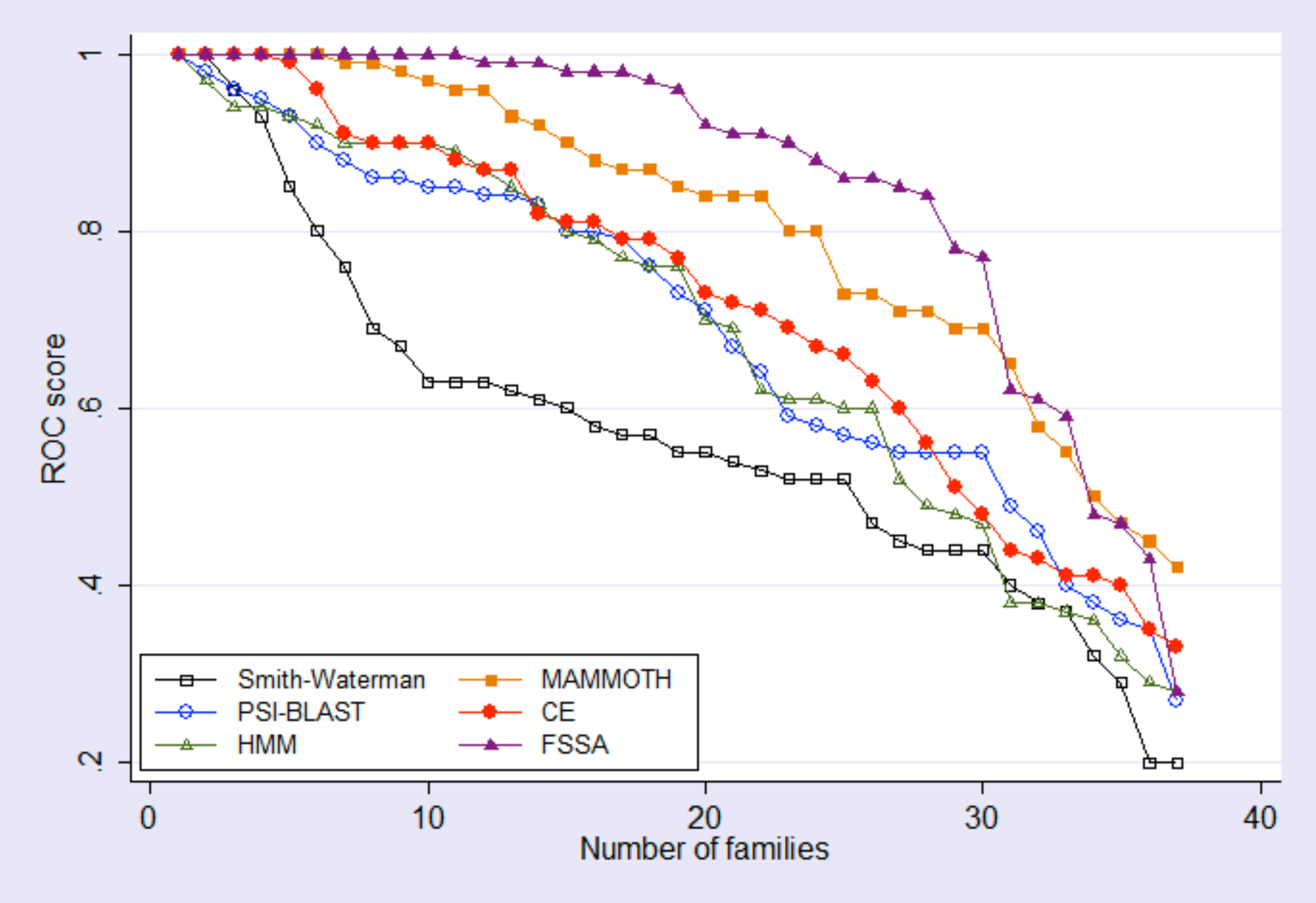


oxidoreductase



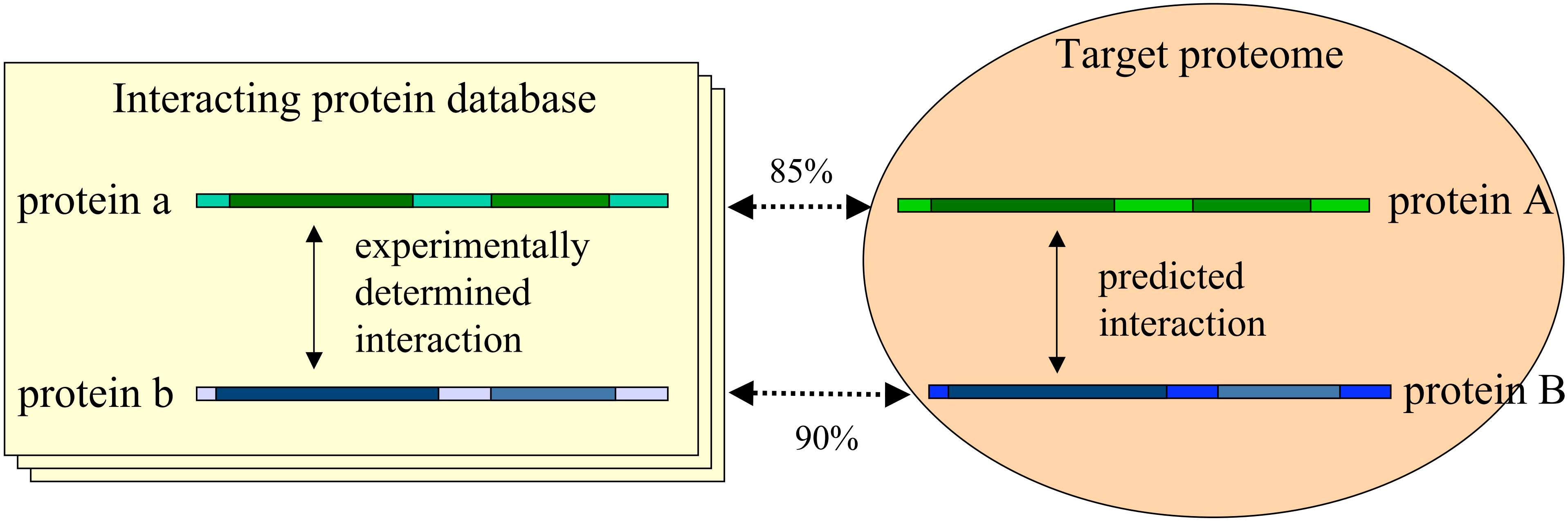
transferase

Function prediction from structure



Kai Wang

Prediction of protein interaction networks



Assign confidence based on similarity and strength of interaction

Key paradigm is the use of homology to transfer information across organisms; not limited to yeast, fly, and worm

Consensus of interactions helps with confidence assignments

Jason McDermott

E. coli predicted protein interaction network

Interpro Annotation

□ □ Annotated/Unannotated Cluster

○ ○ Annotated/Unannotated Node

Edge Confidence

0.0

0.5

1.0

KEGG Pathway Annotation

□ No KEGG Annotation

🟧 Sugar and Lipid Metabolism

▧ Nucleotide Metabolism

🟨 Amino Acid Metabolism

🟥 Energy Metabolism

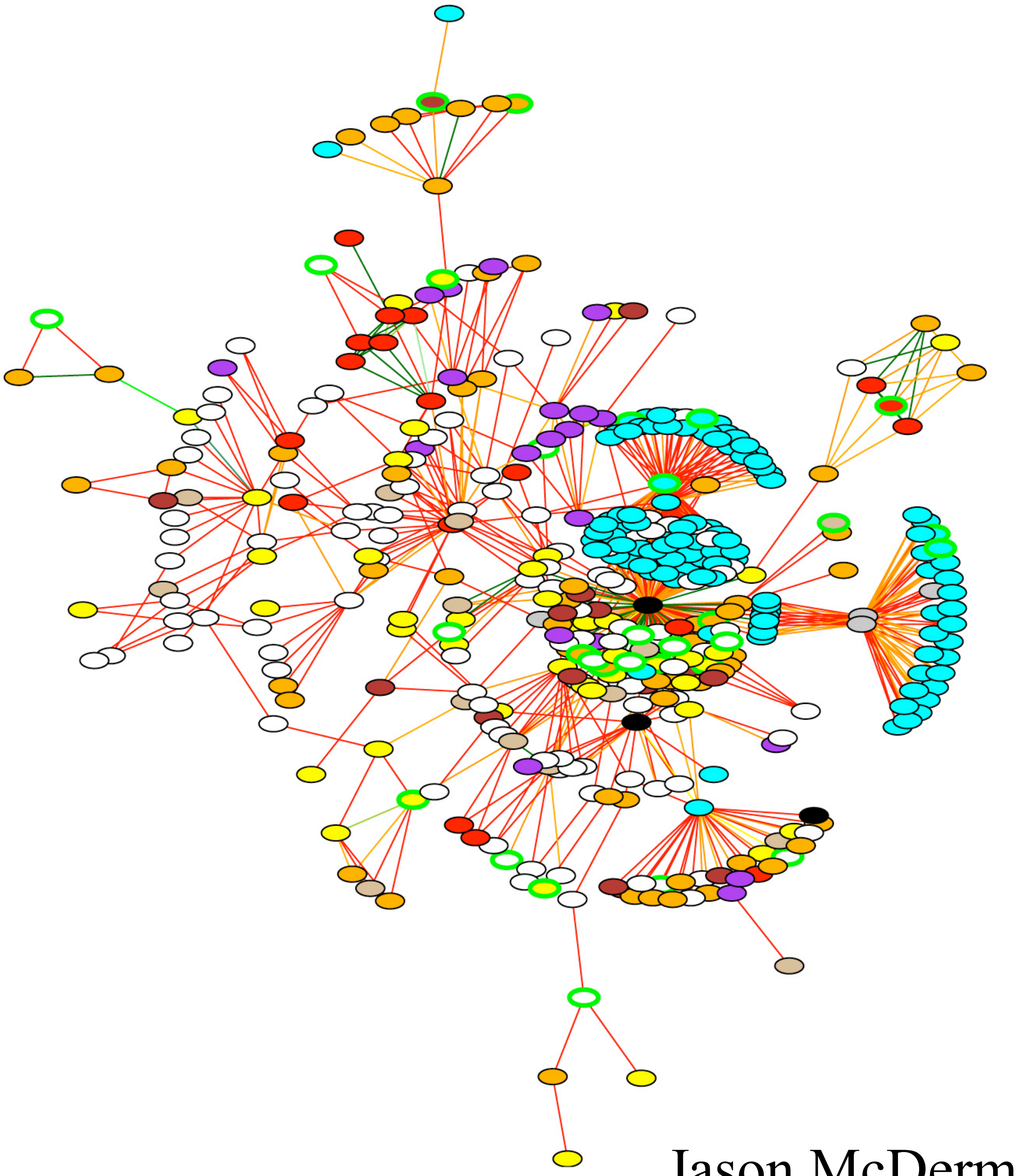
🟫 Other Metabolism

🟪 Genetic Information Processing

🟦 Environmental Information Processing

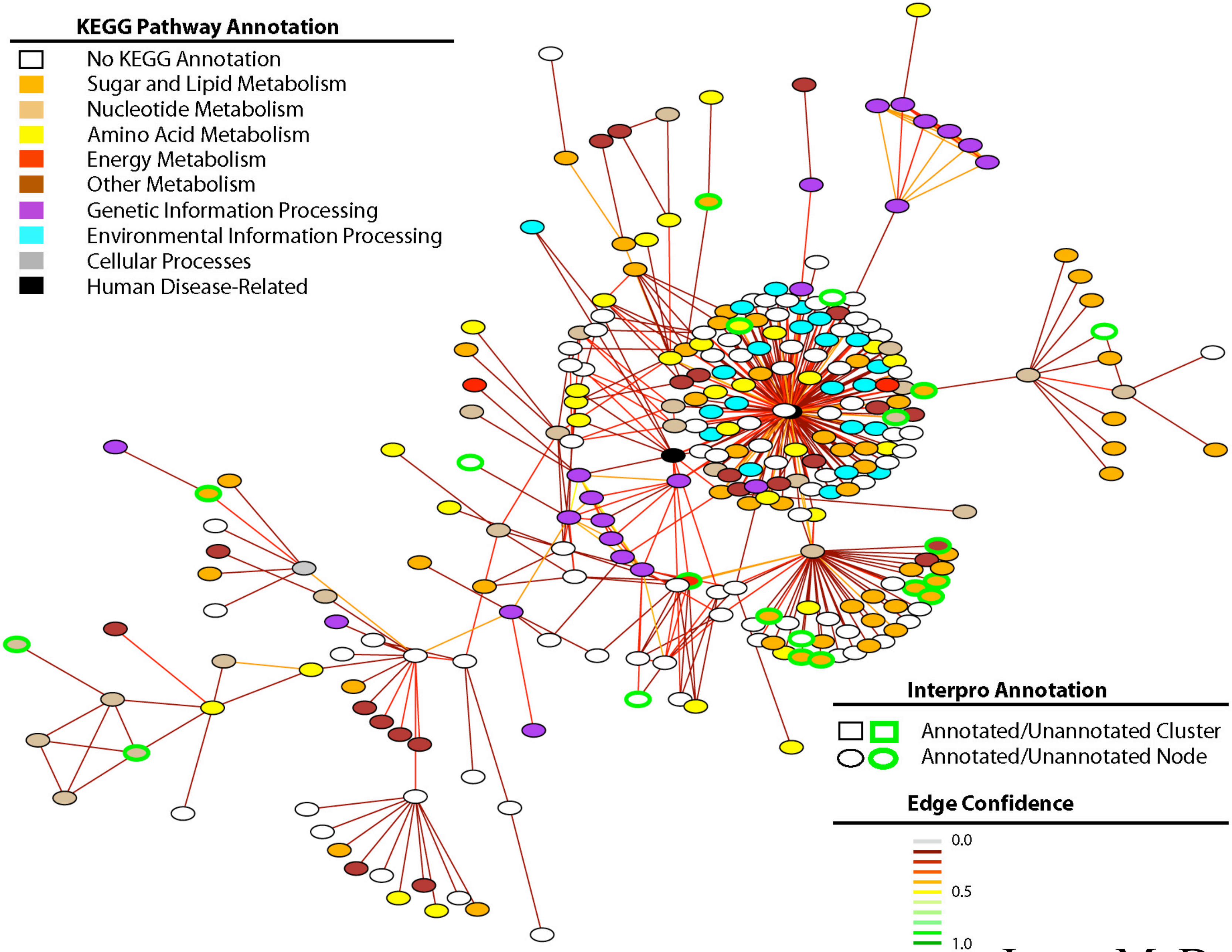
▨ Cellular Processes

⬛ Human Disease-Related



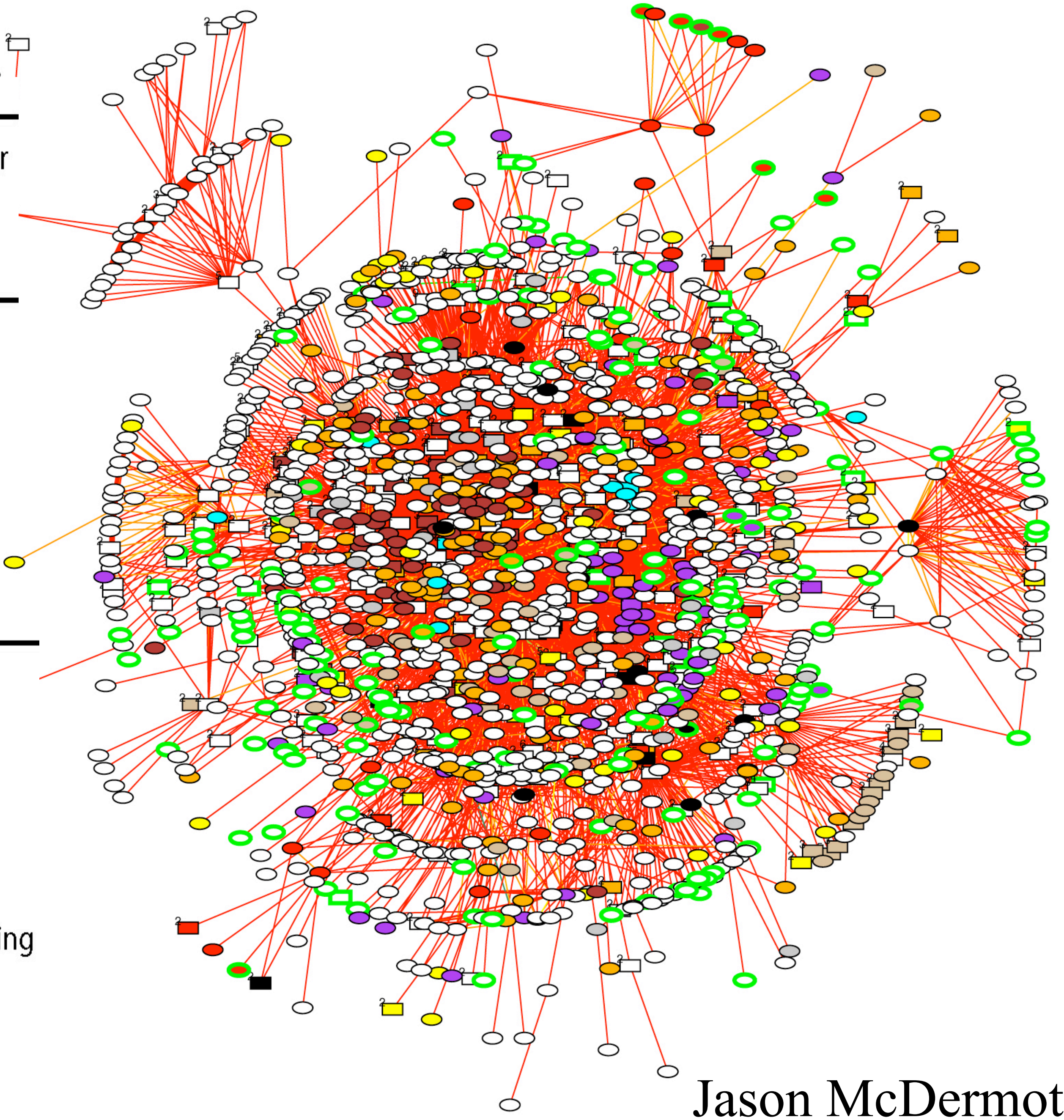
Jason McDermott

M. tuberculosis predicted protein interaction network



Jason McDermott

C. elegans predicted protein interaction network



Interpro Annotation

☐ ☐ Annotated/Unannotated Cluster
○ ○ Annotated/Unannotated Node

Edge Confidence

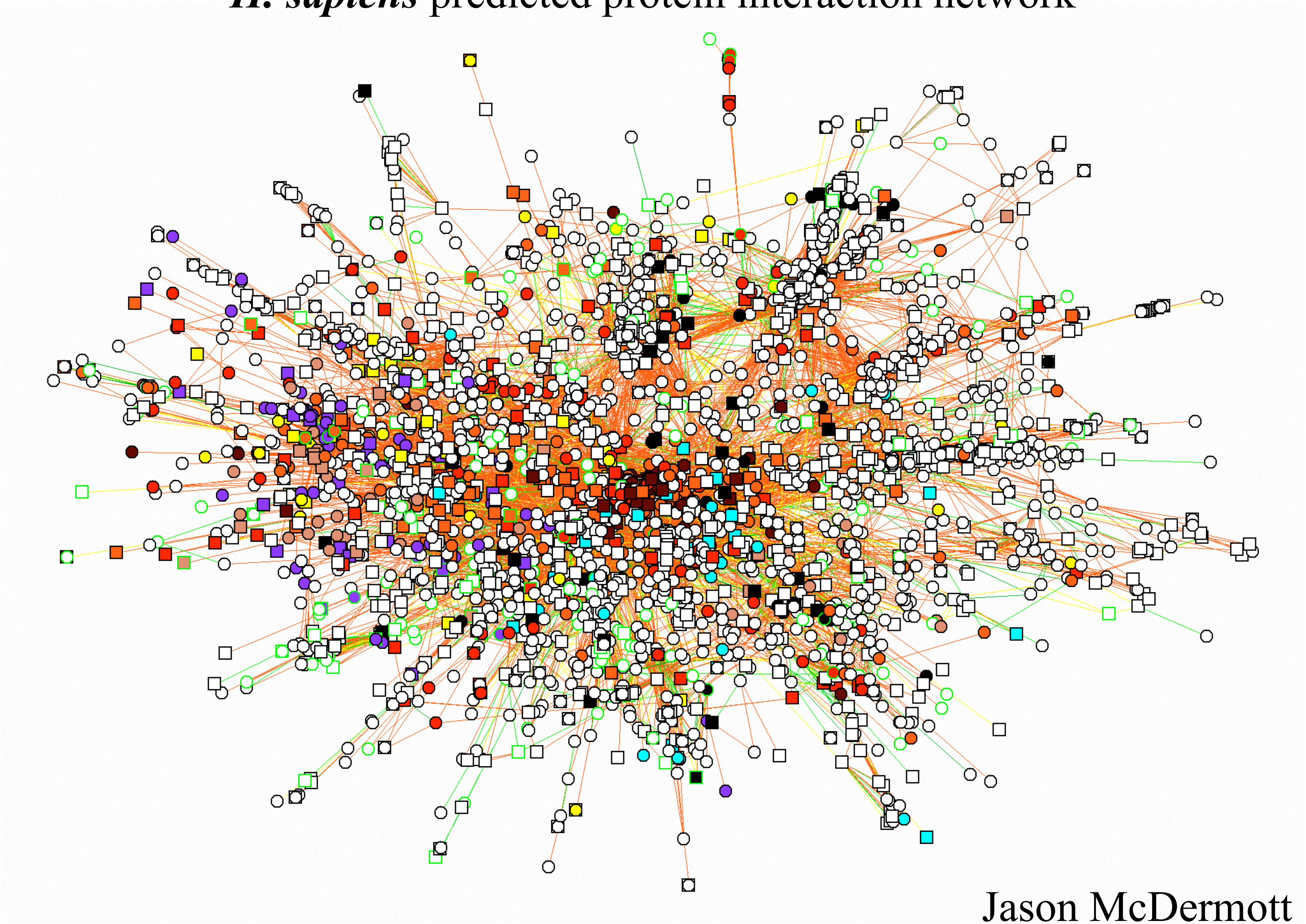
0.0
0.5
1.0

KEGG Pathway Annotation

- ☐ No KEGG Annotation
- 🟧 Sugar and Lipid Metabolism
- ▨ Nucleotide Metabolism
- 🟨 Amino Acid Metabolism
- 🟥 Energy Metabolism
- 🟫 Other Metabolism
- 🟪 Genetic Information Processing
- 🟦 Environmental Information Processing
- ▩ Cellular Processes
- ⬛ Human Disease-Related

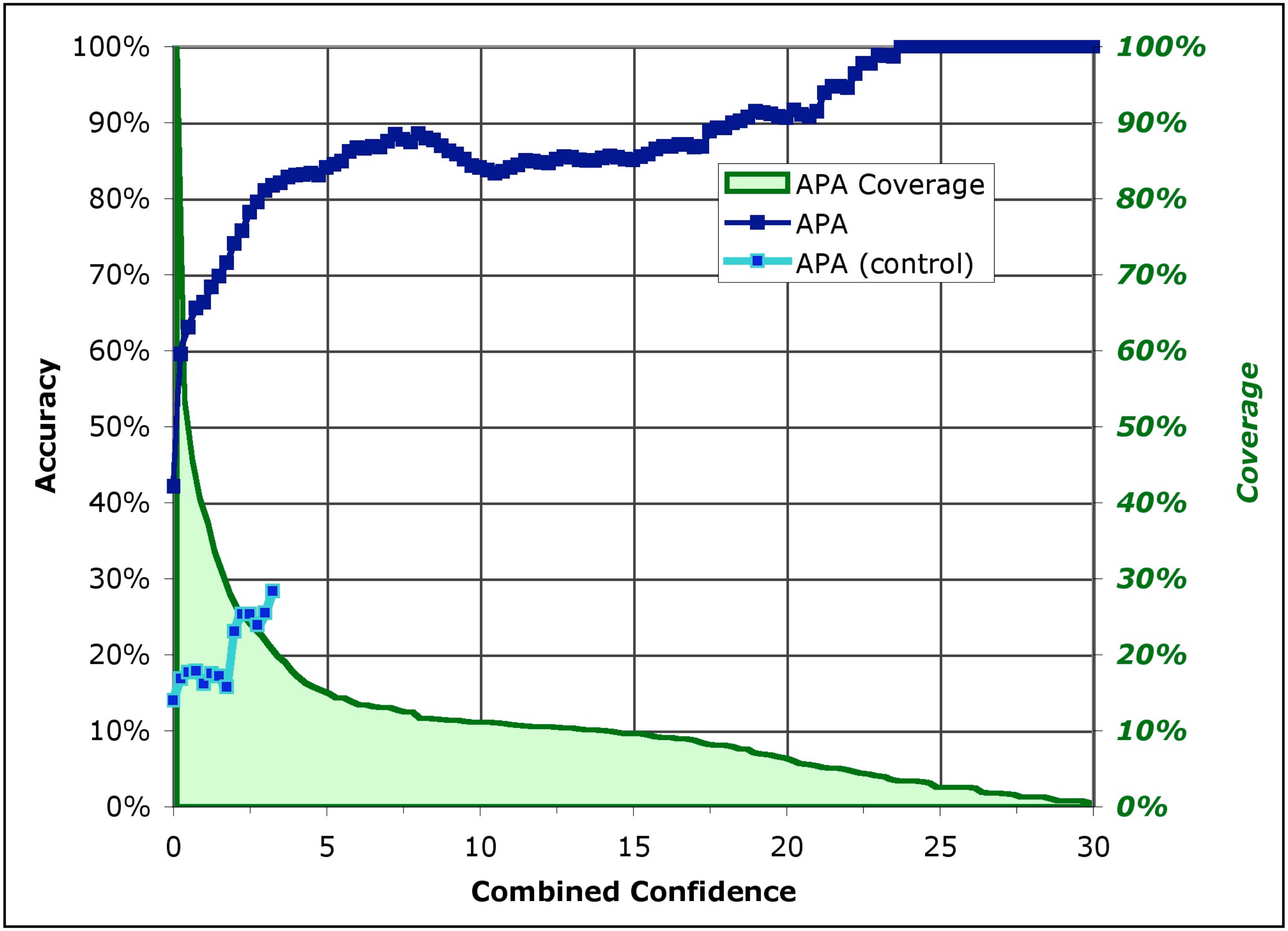
Jason McDermott

H. sapiens predicted protein interaction network



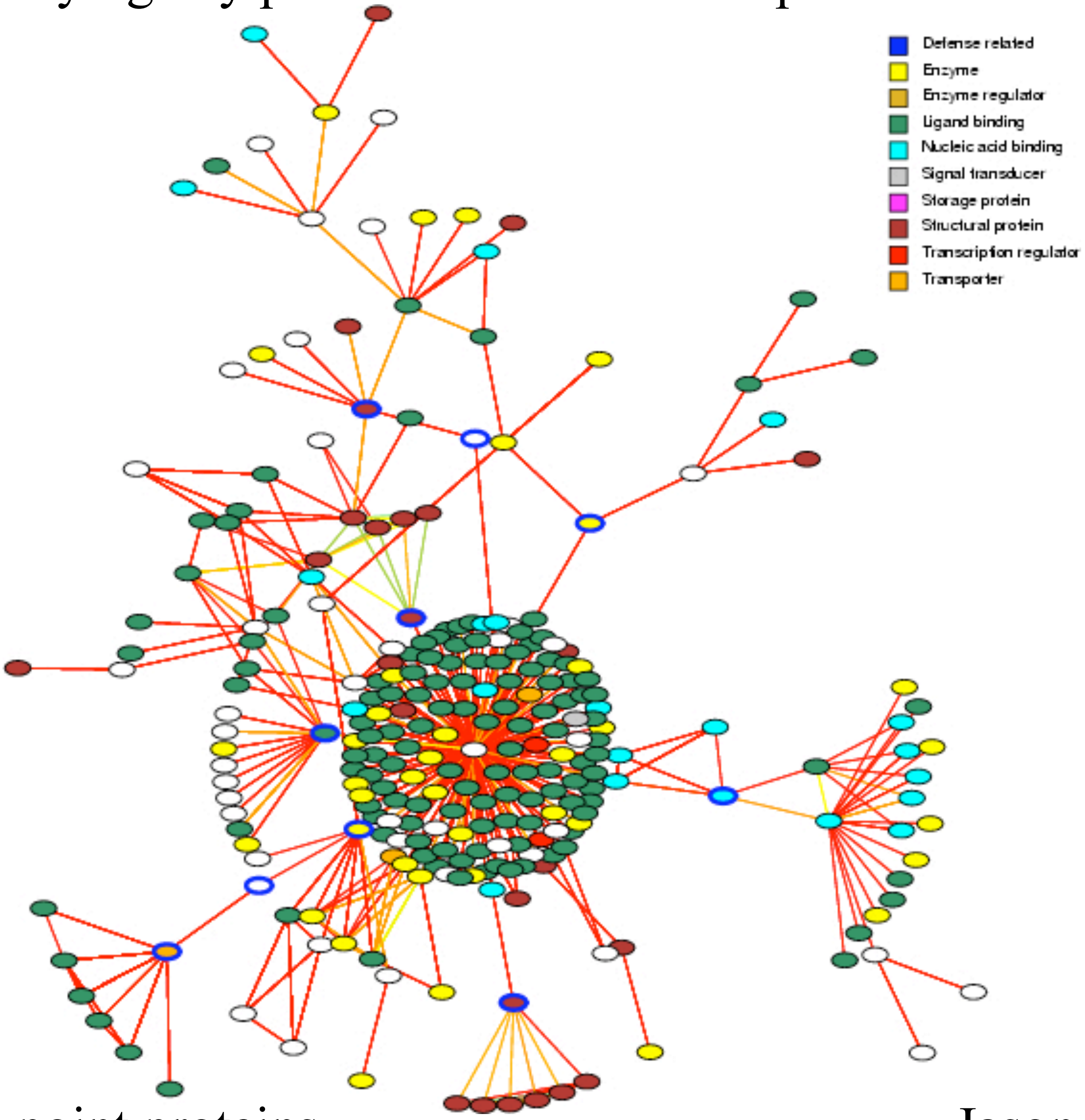
Jason McDermott

Network-based annotation for ***D. melanogaster***



Jason McDermott

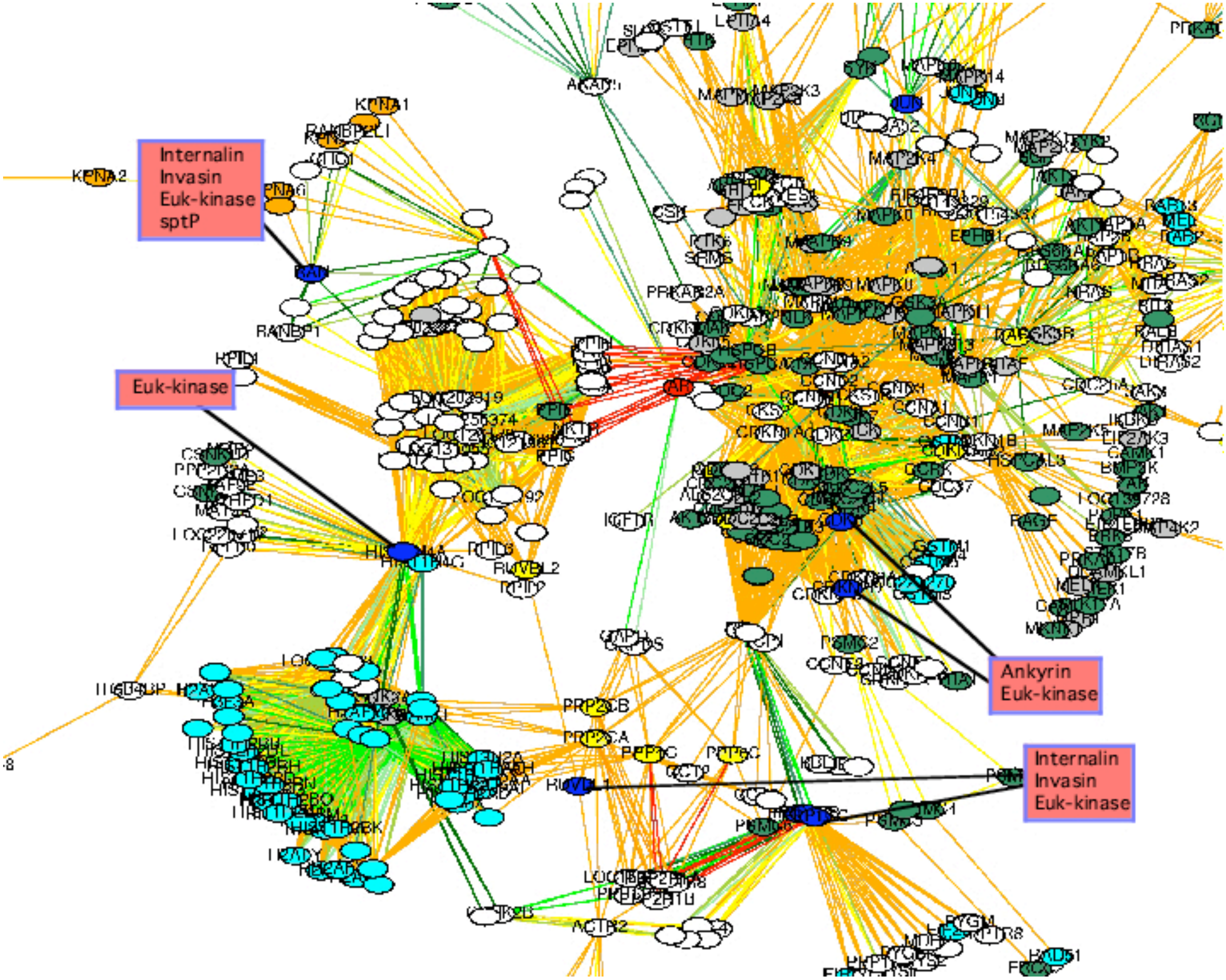
Identifying key proteins on the anthrax predicted network



Articulation point proteins

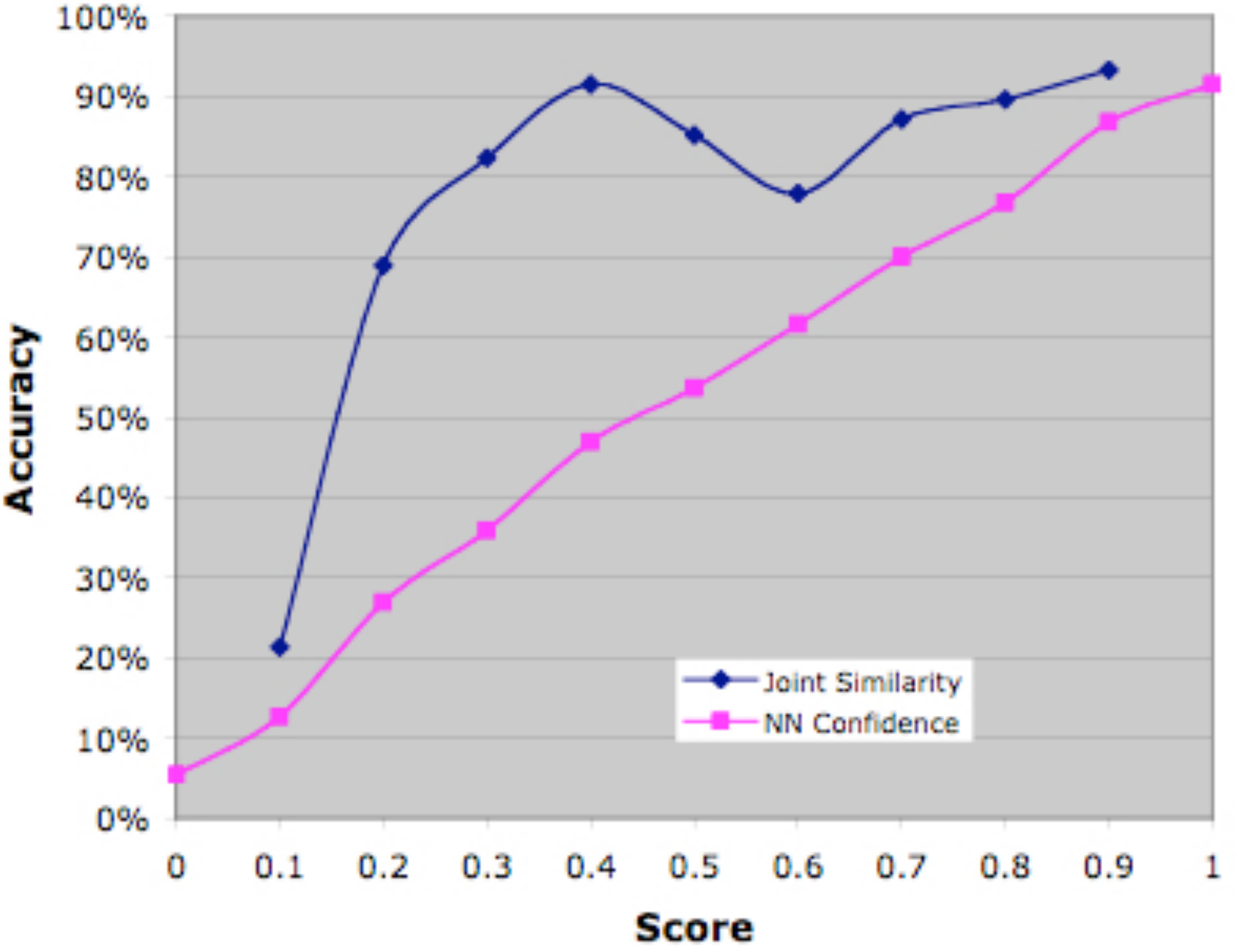
Jason McDermott

Identification of virulence factors



Jason McDermott

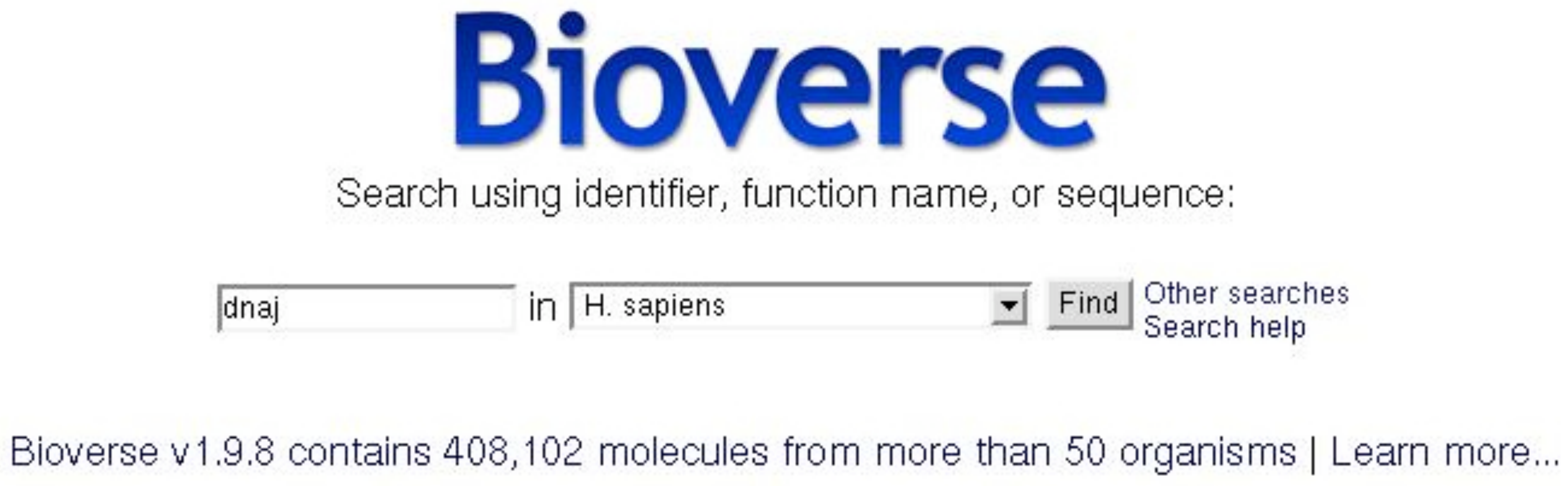
Integration of heterogeneous data with robust confidences



Jason McDermott

Bioverse – explore relationships among molecules and systems

http://bioverse.compbio.washington.edu



Search using identifier, function name, or sequence:

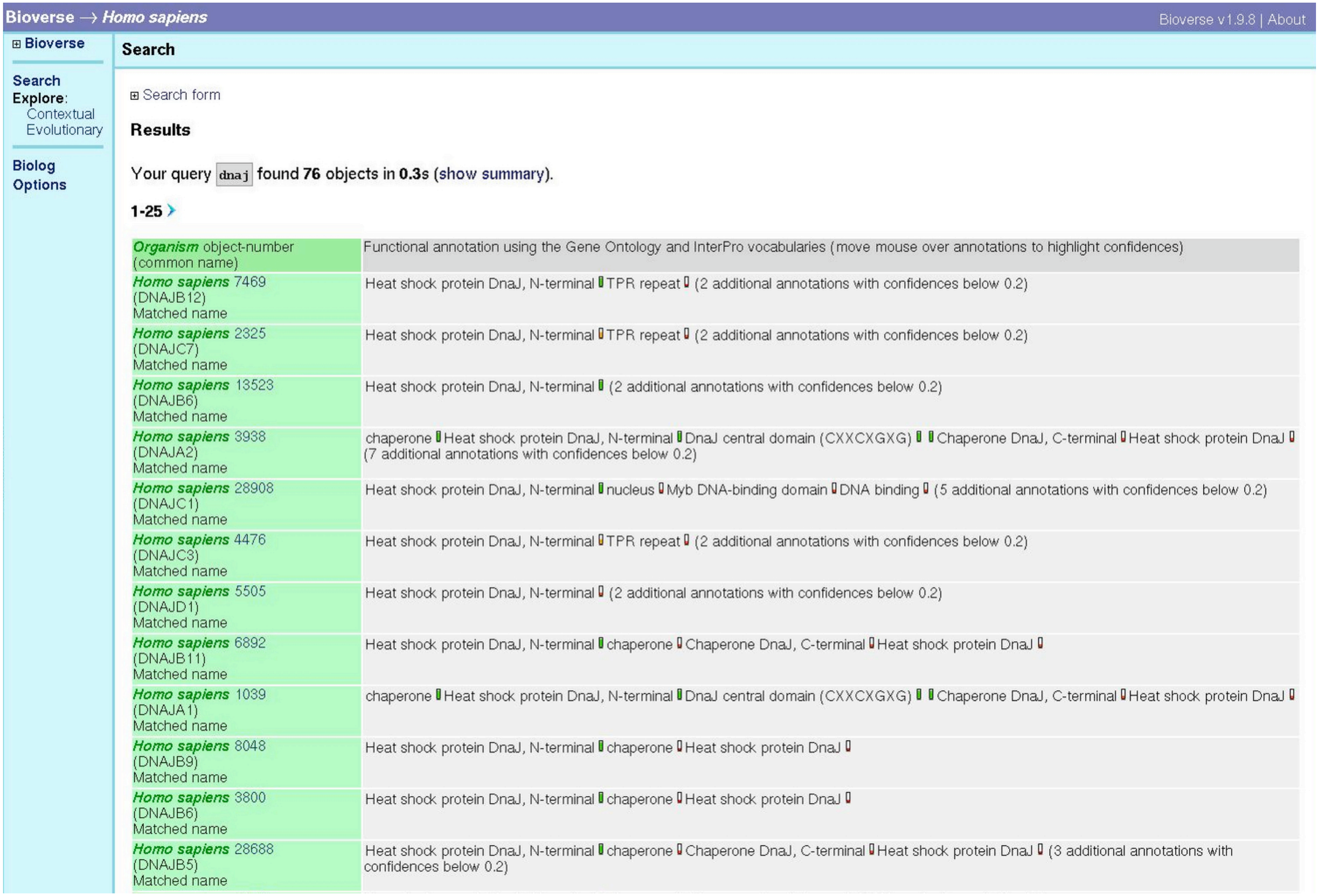
dnaj in H. sapiens

Bioverse v1.9.8 contains 408,102 molecules from more than 50 organisms | Learn more...

Jason McDermott/Michal Guerquin/Zach Frazier

Bioverse – explore relationships among molecules and systems

http://bioverse.compbio.washington.edu



Search

⊞ Search form

Results

Your query `dnaj` found **76** objects in **0.3s** (show summary).

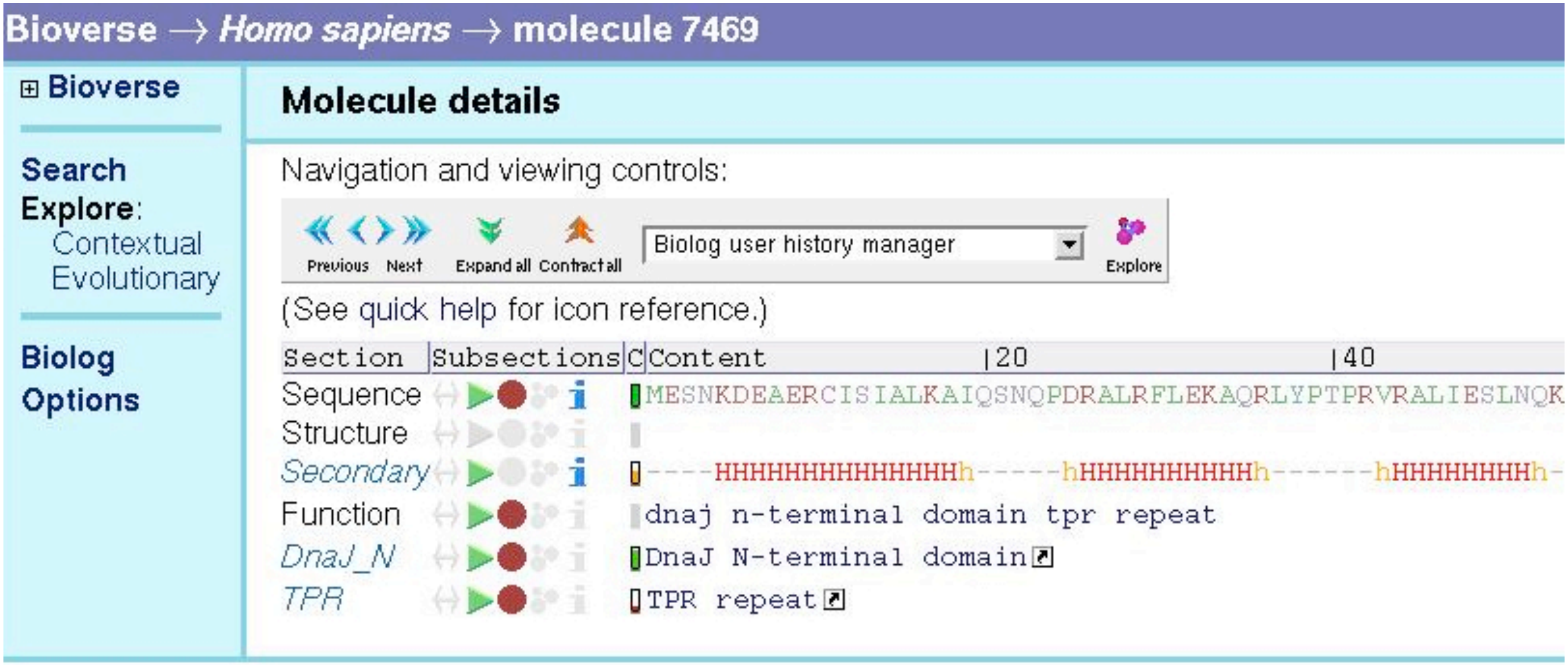
1-25 ›

Organism object-number (common name)	Functional annotation using the Gene Ontology and InterPro vocabularies (move mouse over annotations to highlight confidences)
Homo sapiens 7469 (DNAJB12) Matched name	Heat shock protein DnaJ, N-terminal TPR repeat (2 additional annotations with confidences below 0.2)
Homo sapiens 2325 (DNAJC7) Matched name	Heat shock protein DnaJ, N-terminal TPR repeat (2 additional annotations with confidences below 0.2)
Homo sapiens 13523 (DNAJB6) Matched name	Heat shock protein DnaJ, N-terminal (2 additional annotations with confidences below 0.2)
Homo sapiens 3938 (DNAJA2) Matched name	chaperone Heat shock protein DnaJ, N-terminal DnaJ central domain (CXXCXGXG) Chaperone DnaJ, C-terminal Heat shock protein DnaJ (7 additional annotations with confidences below 0.2)
Homo sapiens 28908 (DNAJC1) Matched name	Heat shock protein DnaJ, N-terminal nucleus Myb DNA-binding domain DNA binding (5 additional annotations with confidences below 0.2)
Homo sapiens 4476 (DNAJC3) Matched name	Heat shock protein DnaJ, N-terminal TPR repeat (2 additional annotations with confidences below 0.2)
Homo sapiens 5505 (DNAJD1) Matched name	Heat shock protein DnaJ, N-terminal (2 additional annotations with confidences below 0.2)
Homo sapiens 6892 (DNAJB11) Matched name	Heat shock protein DnaJ, N-terminal chaperone Chaperone DnaJ, C-terminal Heat shock protein DnaJ
Homo sapiens 1039 (DNAJA1) Matched name	chaperone Heat shock protein DnaJ, N-terminal DnaJ central domain (CXXCXGXG) Chaperone DnaJ, C-terminal Heat shock protein DnaJ
Homo sapiens 8048 (DNAJB9) Matched name	Heat shock protein DnaJ, N-terminal chaperone Heat shock protein DnaJ
Homo sapiens 3800 (DNAJB6) Matched name	Heat shock protein DnaJ, N-terminal chaperone Heat shock protein DnaJ
Homo sapiens 28688 (DNAJB5) Matched name	Heat shock protein DnaJ, N-terminal chaperone Chaperone DnaJ, C-terminal Heat shock protein DnaJ (3 additional annotations with confidences below 0.2)

Jason McDermott/Michal Guerquin/Zach Frazier

Bioverse – explore relationships among molecules and systems

http://bioverse.compbio.washington.edu



Molecule details

Navigation and viewing controls:

(See quick help for icon reference.)

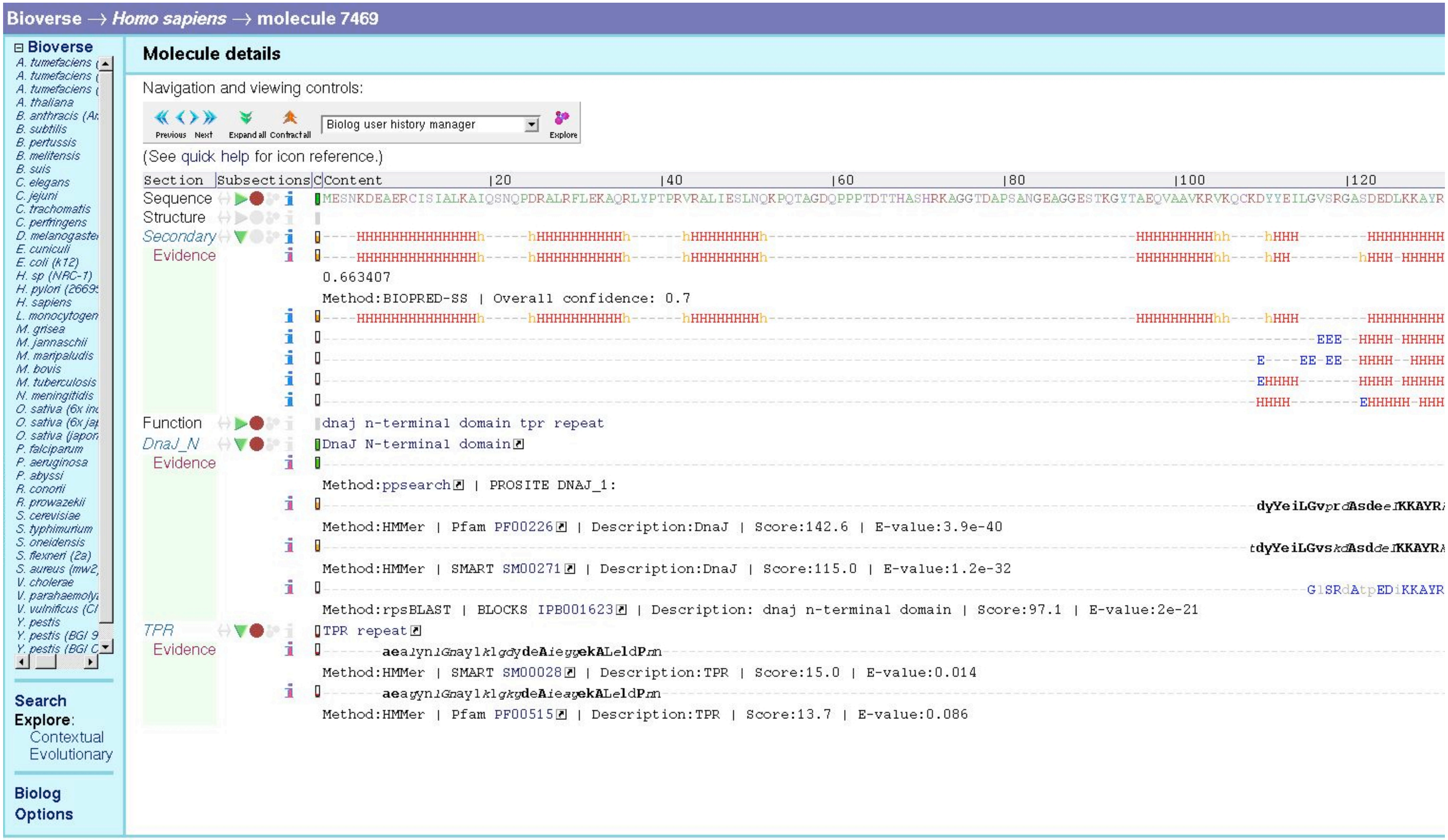
Section	Subsections	C	Content |20 |40
Sequence	↔ ▶ ● i	▮	MESNKDEAERCISIALKAIQSNQPDRALRFLEKAQRLYPTPRVRALIESLNQK
Structure	↔ ▶ ● i	▮	
Secondary	↔ ▶ ● i	▮	----HHHHHHHHHHHHHHh-----hHHHHHHHHHHh------hHHHHHHHHh-
Function	↔ ▶ ● i	▮	dnaj n-terminal domain tpr repeat
DnaJ_N	↔ ▶ ● i	▮	DnaJ N-terminal domain ↗
TPR	↔ ▶ ● i	▯	TPR repeat ↗

▼ ▶	show/hide **evidence** (ex: bioinformatics techniques, predictions, and experimental evidence)
i i	show/hide **properties** (ex: observations and calculations about the object)
● ●	show/hide **similary relationships** (similarity with other objects within and across proteomes)
⁂ ⁂	show/hide **contextual relationships** (ex: protein-protein interactions, co-expression, co-translation)
↔ ↔	show/hide **variants** (ex: splice variants, mutants, alternate alleles, polymorphisms - currently unimplemented)
⇊ ⇈	show/hide **all**
« »	**previous/next** record
⌄	show **more**
▯▯▯▯▯▯▯▯▮▮▮	**confidence** bars representing values from 0-1.0 in 0.1 increments

Jason McDermott/Michal Guerquin/Zach Frazier

Bioverse – explore relationships among molecules and systems

http://bioverse.compbio.washington.edu



Molecule details

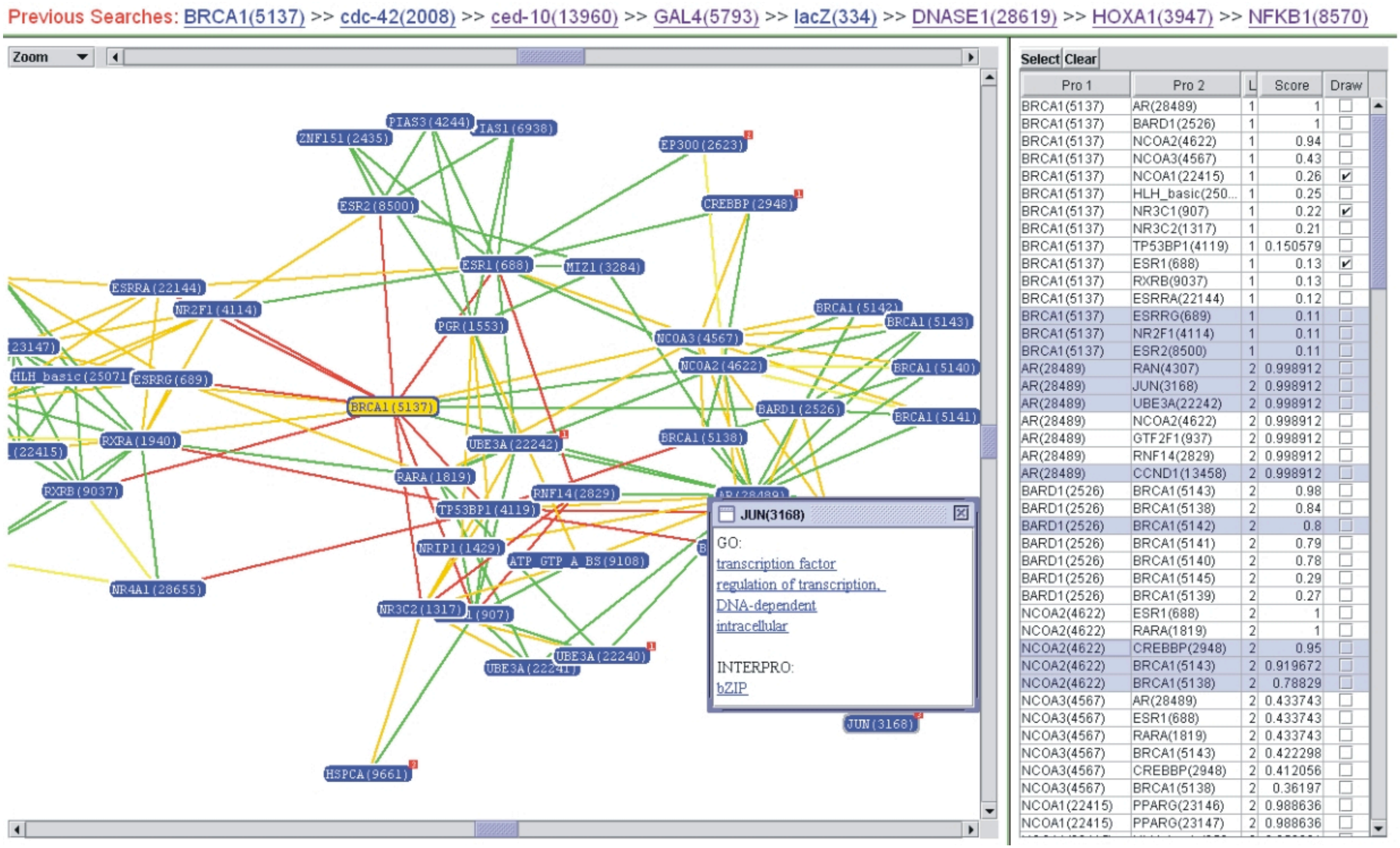
Navigation and viewing controls:

(See quick help for icon reference.)

```
Section   Subsections C Content          |20        |40        |60        |80        |100       |120
Sequence                  MESNKDEAERCISIALKAIQSNQPDRALRFLEKAQRLYPTPRVRALIESLNQKPQTAGDQPPPTDTTHASHRKAGGTDAPSANGEAGGESTKGYTAEQVAAVKRVKQCKDYYEILGVSRGASDEDLKKAYR
Structure
Secondary                 ----HHHHHHHHHHHHHHHh-----hHHHHHHHHHHh------hHHHHHHHHh-------------------------------------HHHHHHHHHhh---hHHH--------HHHHHHHHH
  Evidence                ----HHHHHHHHHHHHHHHh-----hHHHHHHHHHHh------hHHHHHHHHh-------------------------------------HHHHHHHHHhh---hHH--------hHHH-HHHHH
                          0.663407
                          Method:BIOPRED-SS | Overall confidence: 0.7
                          ----HHHHHHHHHHHHHHHh-----hHHHHHHHHHHh------hHHHHHHHHh-------------------------------------HHHHHHHHHhh---hHHH--------HHHHHHHHH
                          -------------------------------------------------------------------------------------------------------------------EEE--HHHH-HHHHH
                          --------------------------------------------------------------------------------------------------------E----EE-EE--HHHH--HHHH
                          --------------------------------------------------------------------------------------------------------EHHHH------HHHH-HHHHH
                          --------------------------------------------------------------------------------------------------------HHHH-------EHHHHH-HHH
Function                  dnaj n-terminal domain tpr repeat
DnaJ_N                    DnaJ N-terminal domain
  Evidence
                          Method:ppsearch | PROSITE DNAJ_1:
                                                                                                                  dyYeiLGvprdAsdeeIKKAYR
                          Method:HMMer | Pfam PF00226 | Description:DnaJ | Score:142.6 | E-value:3.9e-40
                                                                                                                 tdyYeiLGvskdAsddeIKKAYR
                          Method:HMMer | SMART SM00271 | Description:DnaJ | Score:115.0 | E-value:1.2e-32
                                                                                                                     GlSRdAtpEDiKKAYR
                          Method:rpsBLAST | BLOCKS IPB001623 | Description: dnaj n-terminal domain | Score:97.1 | E-value:2e-21
TPR                       TPR repeat
  Evidence                      aealynlGnaylklgdydeAieggekALeldPnn
                          Method:HMMer | SMART SM00028 | Description:TPR | Score:15.0 | E-value:0.014
                                aeagynlGnaylklgkydeAieagekALeldPnn
                          Method:HMMer | Pfam PF00515 | Description:TPR | Score:13.7 | E-value:0.086
```

Jason McDermott/Michal Guerquin/Zach Frazier

Bioverse - Integrator



Previous Searches: BRCA1(5137) >> cdc-42(2008) >> ced-10(13960) >> GAL4(5793) >> lacZ(334) >> DNASE1(28619) >> HOXA1(3947) >> NFKB1(8570)

JUN(3168)

GO:

transcription factor

regulation of transcription, DNA-dependent

intracellular

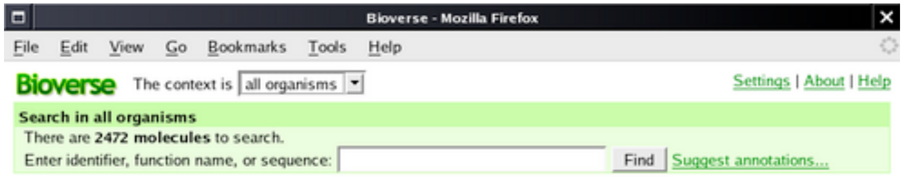
INTERPRO:

bZIP

Pro 1	Pro 2	L	Score	Draw
BRCA1(5137)	AR(28489)	1	1	☐
BRCA1(5137)	BARD1(2526)	1	1	☐
BRCA1(5137)	NCOA2(4622)	1	0.94	☐
BRCA1(5137)	NCOA3(4567)	1	0.43	☐
BRCA1(5137)	NCOA1(22415)	1	0.26	☑
BRCA1(5137)	HLH_basic(250...	1	0.25	☐
BRCA1(5137)	NR3C1(907)	1	0.22	☑
BRCA1(5137)	NR3C2(1317)	1	0.21	☐
BRCA1(5137)	TP53BP1(4119)	1	0.150579	☐
BRCA1(5137)	ESR1(688)	1	0.13	☑
BRCA1(5137)	RXRB(9037)	1	0.13	☐
BRCA1(5137)	ESRRA(22144)	1	0.12	☐
BRCA1(5137)	ESRRG(689)	1	0.11	☐
BRCA1(5137)	NR2F1(4114)	1	0.11	☐
BRCA1(5137)	ESR2(8500)	1	0.11	☐
AR(28489)	RAN(4307)	2	0.998912	☐
AR(28489)	JUN(3168)	2	0.998912	☐
AR(28489)	UBE3A(22242)	2	0.998912	☐
AR(28489)	NCOA2(4622)	2	0.998912	☐
AR(28489)	GTF2F1(937)	2	0.998912	☐
AR(28489)	RNF14(2829)	2	0.998912	☐
AR(28489)	CCND1(13458)	2	0.998912	☐
BARD1(2526)	BRCA1(5143)	2	0.98	☐
BARD1(2526)	BRCA1(5138)	2	0.84	☐
BARD1(2526)	BRCA1(5142)	2	0.8	☐
BARD1(2526)	BRCA1(5141)	2	0.79	☐
BARD1(2526)	BRCA1(5140)	2	0.78	☐
BARD1(2526)	BRCA1(5145)	2	0.29	☐
BARD1(2526)	BRCA1(5139)	2	0.27	☐
NCOA2(4622)	ESR1(688)	2	1	☐
NCOA2(4622)	RARA(1819)	2	1	☐
NCOA2(4622)	CREBBP(2948)	2	0.95	☐
NCOA2(4622)	BRCA1(5143)	2	0.919672	☐
NCOA2(4622)	BRCA1(5138)	2	0.78829	☐
NCOA3(4567)	AR(28489)	2	0.433743	☐
NCOA3(4567)	ESR1(688)	2	0.433743	☐
NCOA3(4567)	RARA(1819)	2	0.433743	☐
NCOA3(4567)	BRCA1(5143)	2	0.422298	☐
NCOA3(4567)	CREBBP(2948)	2	0.412056	☐
NCOA3(4567)	BRCA1(5138)	2	0.36197	☐
NCOA1(22415)	PPARG(23146)	2	0.988636	☐
NCOA1(22415)	PPARG(23147)	2	0.988636	☐

Aaron Chang

Bioverse – v2.0

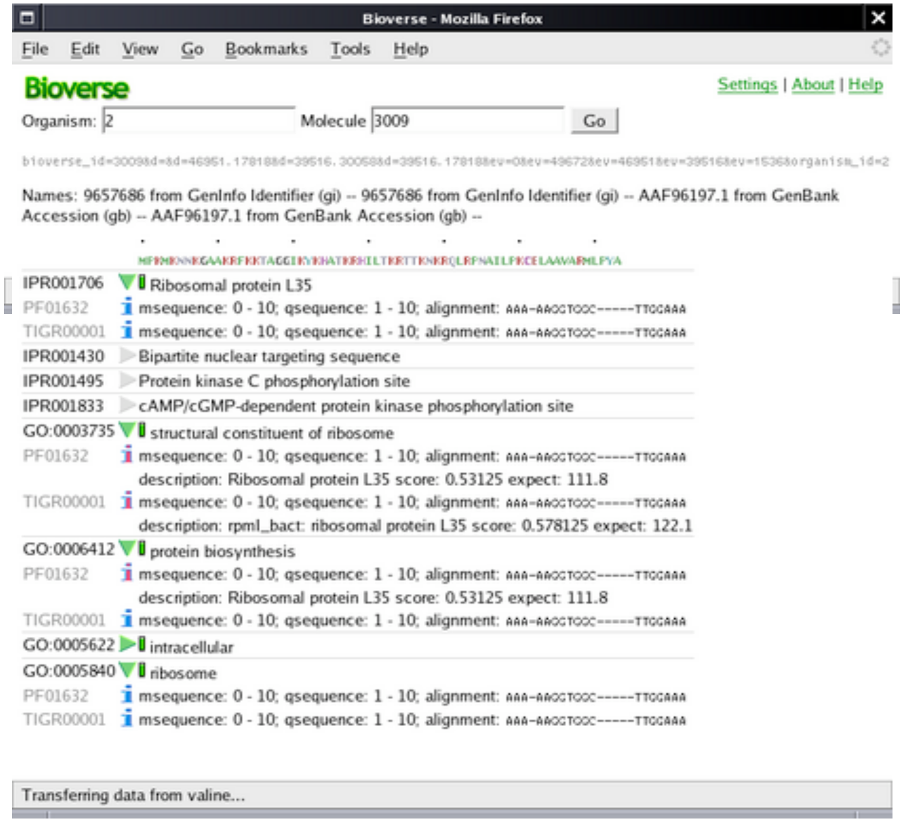


The context is all organisms

Search in all organisms

There are **2472 molecules** to search.

Enter identifier, function name, or sequence:



Organism: 2 Molecule 3009

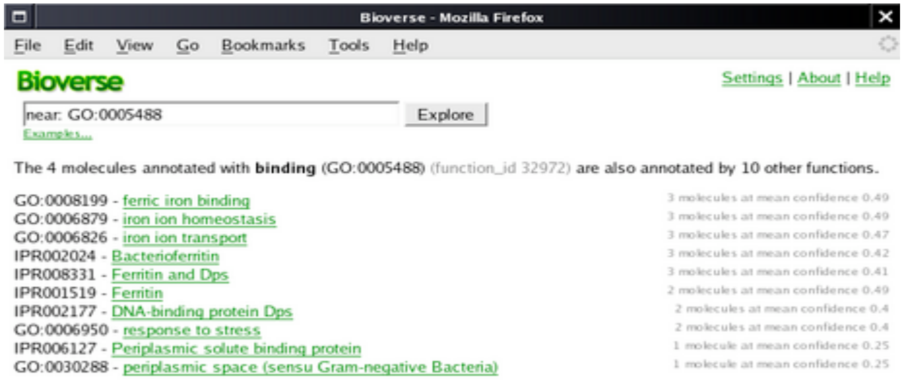
```
bioverse_id=3009&d=&d=46951.17818&d=39516.30058&d=39516.17818&ev=0&ev=49672&ev=46951&ev=39516&ev=1536&organism_id=2
```

Names: 9657686 from GenInfo Identifier (gi) -- 9657686 from GenInfo Identifier (gi) -- AAF96197.1 from GenBank Accession (gb) -- AAF96197.1 from GenBank Accession (gb) --

```
MPKMKNNKGAAKRFKKTAGGIKYKHATKRHILTKRTTKNKRQLRPNAILPKCELAAVARMLPYA
```

IPR001706 Ribosomal protein L35
PF01632 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
TIGR00001 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
IPR001430 Bipartite nuclear targeting sequence
IPR001495 Protein kinase C phosphorylation site
IPR001833 cAMP/cGMP-dependent protein kinase phosphorylation site
GO:0003735 structural constituent of ribosome
PF01632 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
description: Ribosomal protein L35 score: 0.53125 expect: 111.8
TIGR00001 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
description: rpml_bact: ribosomal protein L35 score: 0.578125 expect: 122.1
GO:0006412 protein biosynthesis
PF01632 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
description: Ribosomal protein L35 score: 0.53125 expect: 111.8
TIGR00001 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
GO:0005622 intracellular
GO:0005840 ribosome
PF01632 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA
TIGR00001 msequence: 0 - 10; qsequence: 1 - 10; alignment: AAA-AAGGTGGC-----TTGGAAA

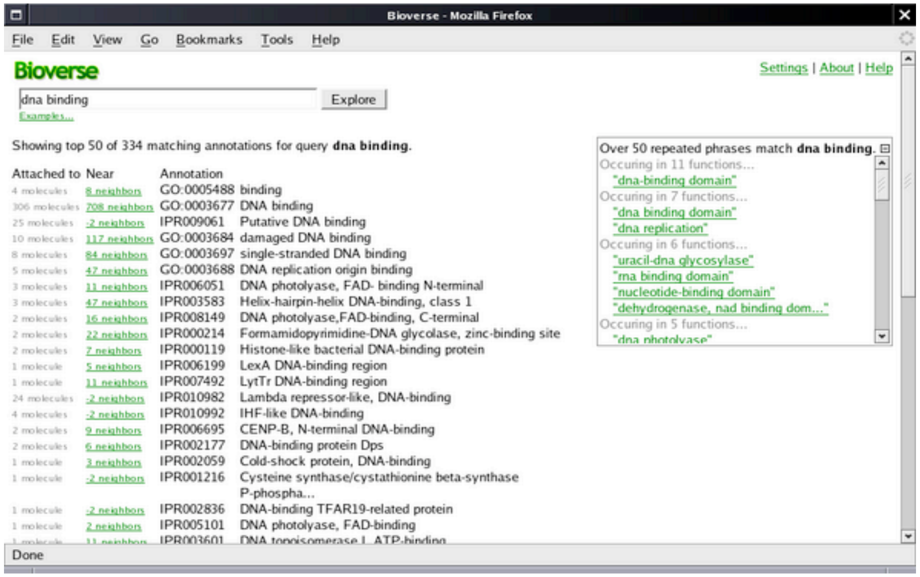
Transferring data from valine...



near: GO:0005488

The 4 molecules annotated with **binding** (GO:0005488) (function_id 32972) are also annotated by 10 other functions.

GO:0008199 - ferric iron binding	3 molecules at mean confidence 0.49
GO:0006879 - iron ion homeostasis	3 molecules at mean confidence 0.49
GO:0006826 - iron ion transport	3 molecules at mean confidence 0.47
IPR002024 - Bacterioferritin	3 molecules at mean confidence 0.42
IPR008331 - Ferritin and Dps	3 molecules at mean confidence 0.41
IPR001519 - Ferritin	2 molecules at mean confidence 0.49
IPR002177 - DNA-binding protein Dps	2 molecules at mean confidence 0.4
GO:0006950 - response to stress	2 molecules at mean confidence 0.4
IPR006127 - Periplasmic solute binding protein	1 molecule at mean confidence 0.25
GO:0030288 - periplasmic space (sensu Gram-negative Bacteria)	1 molecule at mean confidence 0.25



dna binding

Showing top 50 of 334 matching annotations for query **dna binding**.

Attached to	Near	Annotation	
4 molecules	8 neighbors	GO:0005488	binding
306 molecules	708 neighbors	GO:0003677	DNA binding
25 molecules	-2 neighbors	IPR009061	Putative DNA binding
10 molecules	117 neighbors	GO:0003684	damaged DNA binding
8 molecules	84 neighbors	GO:0003697	single-stranded DNA binding
5 molecules	47 neighbors	GO:0003688	DNA replication origin binding
3 molecules	11 neighbors	IPR006051	DNA photolyase, FAD- binding N-terminal
3 molecules	47 neighbors	IPR003583	Helix-hairpin-helix DNA-binding, class 1
2 molecules	16 neighbors	IPR008149	DNA photolyase,FAD-binding, C-terminal
2 molecules	22 neighbors	IPR000214	Formamidopyrimidine-DNA glycolase, zinc-binding site
2 molecules	7 neighbors	IPR000119	Histone-like bacterial DNA-binding protein
1 molecule	5 neighbors	IPR006199	LexA DNA-binding region
1 molecule	11 neighbors	IPR007492	LytTr DNA-binding region
24 molecules	-2 neighbors	IPR010982	Lambda repressor-like, DNA-binding
4 molecules	-2 neighbors	IPR010992	IHF-like DNA-binding
2 molecules	9 neighbors	IPR006695	CENP-B, N-terminal DNA-binding
2 molecules	6 neighbors	IPR002177	DNA-binding protein Dps
1 molecule	3 neighbors	IPR002059	Cold-shock protein, DNA-binding
1 molecule	-2 neighbors	IPR001216	Cysteine synthase/cystathionine beta-synthase P-phospha...
1 molecule	-2 neighbors	IPR002836	DNA-binding TFAR19-related protein
1 molecule	2 neighbors	IPR005101	DNA photolyase, FAD-binding

Over 50 repeated phrases match **dna binding**.

Occuring in 11 functions...
- "dna-binding domain"

Occuring in 7 functions...
- "dna binding domain"
- "dna replication"

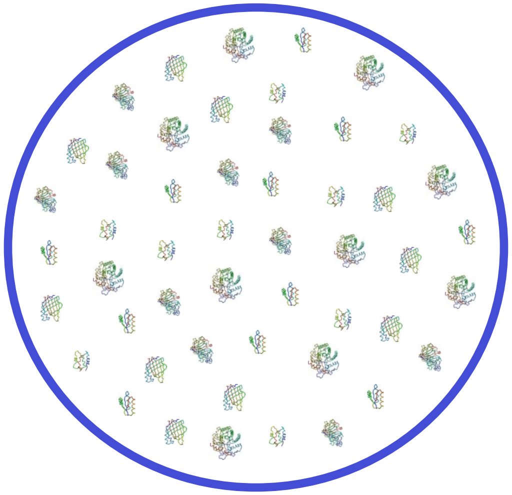
Occuring in 6 functions...
- "uracil-dna glycosylase"
- "rna binding domain"
- "nucleotide-binding domain"
- "dehydrogenase, nad binding dom..."

Occuring in 5 functions...

Done

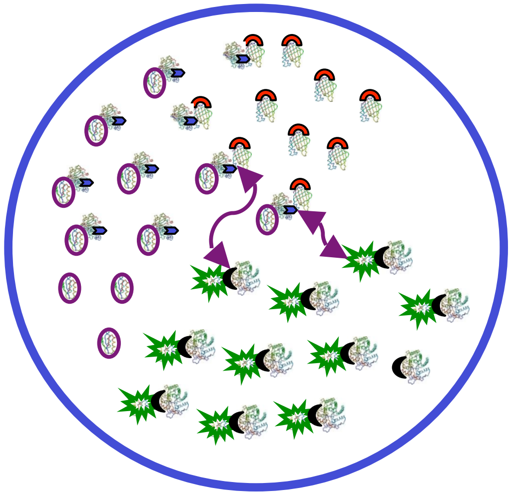
Michal Guerquin

Where is all this going?



Structural genomics

+



Functional genomics

+

Computational biology

Take home message

Prediction of protein structure, function, and networks may be used to model whole genomes to understand organismal function and evolution

Acknowledgements

Aaron Chang	Ling-Hong Hung
David Nickle	Mike Inouye
Ekachai Jenwitheesuk	Michal Guerquin
Gong Cheng	Stewart Moughon
Jason McDermott	Shing-Chung Ngan
Jeremy Horst	Tianyun Liu
Kai Wang	Zach Frazier

National Institutes of Health
National Science Foundation
Searle Scholars Program (Kinship Foundation)
UW Advanced Technology Initiative in Infectious Diseases

http://protinfo.compbio.washington.edu
http://bioverse.compbio.washington.edu