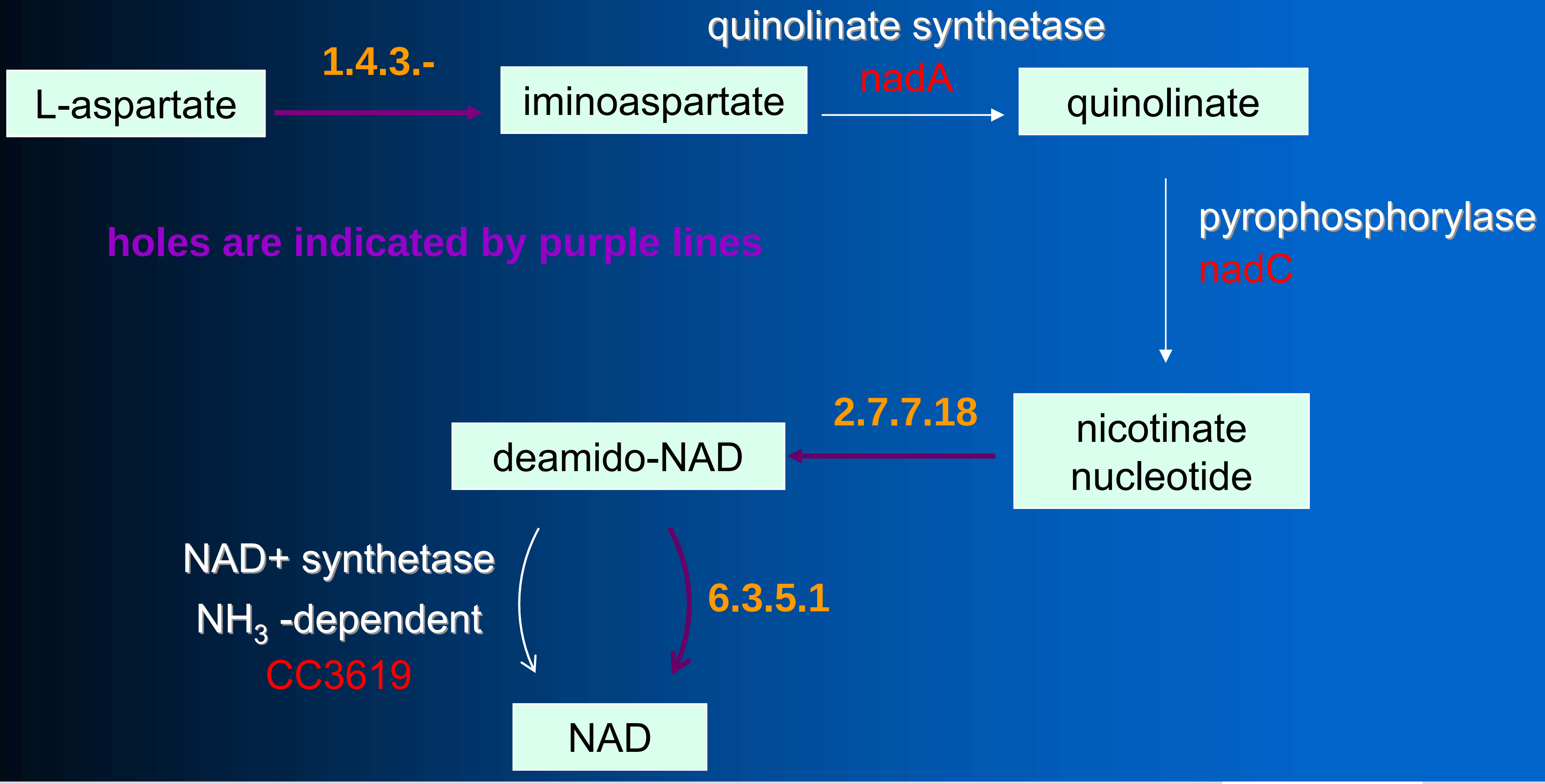
Using genome context data to identify missing enzymes in PGDBs

Outline

- Motivation
- Genome context methods used
- Principle behind PHFiller-GC
- Algorithm – Bayesian classifier
- Validation
  - Gold-standard dataset from EcoCyc
  - Results for EcoCyc
  - Results for other PGDBs
- Results summary
- Implementation as part of Pathway Tools

What is a pathway hole?

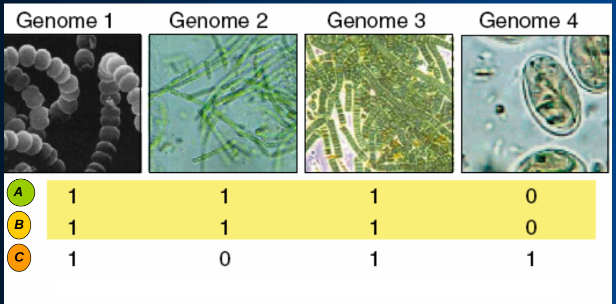
Definition: Pathway Holes are reactions in metabolic pathways for which no enzyme is identified in the PGDB.



Why use genome context data to fill pathway holes?

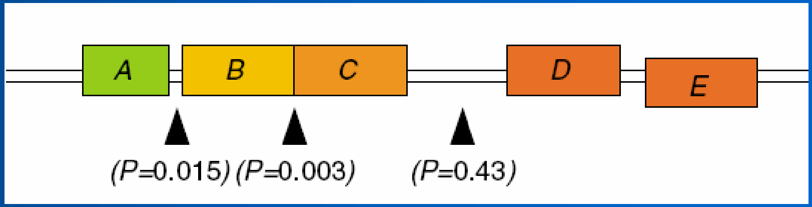
- **Pathway hole filler (PHFiller) generates no hits for many pathway holes**
  - No enzyme sequences for the reaction (orphan enzyme)
  - No homologous sequences in genome (convergent evolution)
  - Organism doesn’t do the reaction
- **About 44% of MetaCyc small-molecule metabolism reaction have no sequences**

Genome Context Methods

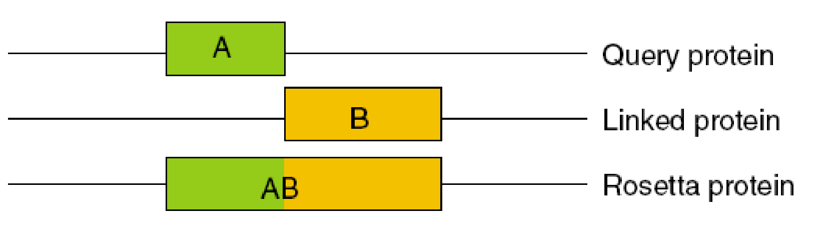


	Genome 1	Genome 2	Genome 3	Genome 4
A	1	1	1	0
B	1	1	1	0
C	1	0	1	1

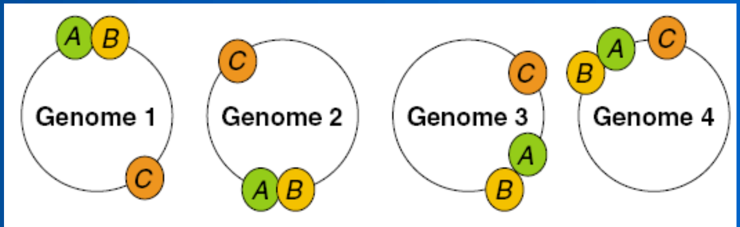
PP (phylogenetic profiles)



GC (gene clusters)



RS (gene fusions)

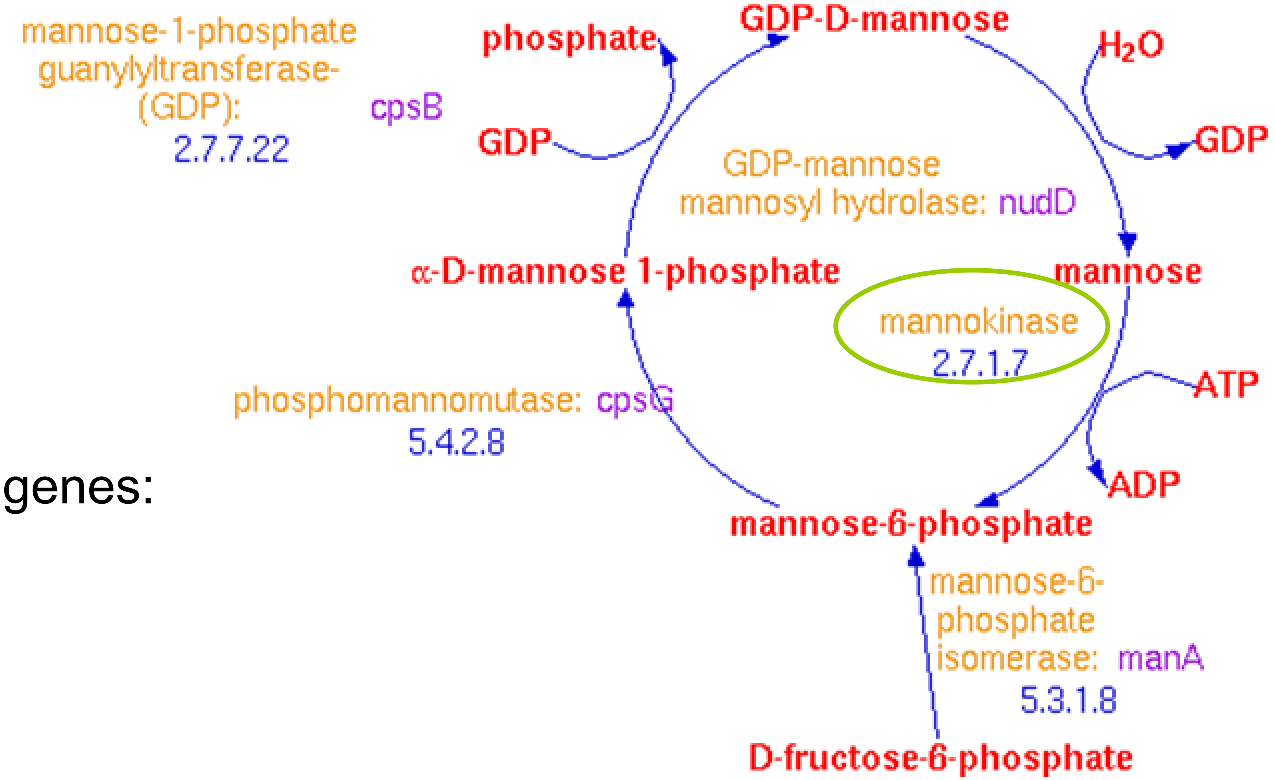


GN (gene neighbors)

From Bowers et al., 2004

Principle behind PHFiller-GC

Use genes related to pathway genes by genome context methods to identify and evaluate pathway hole fillers.



Known genes:

cpsB

cpsG

nudD

manA

Principle behind PHFiller-GC

- Pathway – GDP-mannose metabolism
- Hole – mannokinase
- Known enzymes – cpsG, cpsB, nudD, manA

Gene	PP pairs	GN pairs	RS pairs	GC pairs
cpsG	glmM	cpsB wcaI nudD blmS hflB manA	None	wcaJ cpsB
cpsB	None	wcaI fcl nudD gmd cpsG	yihS	cpsG wcaI
nudD	None	gmd fcl cpsB cpsG wcaF wcaI	None	wcaI fcl
manA	None	cpsG fumA fumC ydgH ydgA tus rstB rstA	None	ydgA

Identification of candidates

- **PHFiller uses BLAST hits to isozyme sequences**
- **PHFiller-GC uses**

  + Genes in a pathway directon

  + Genes functionally associated to a pathway gene by one or more GC method

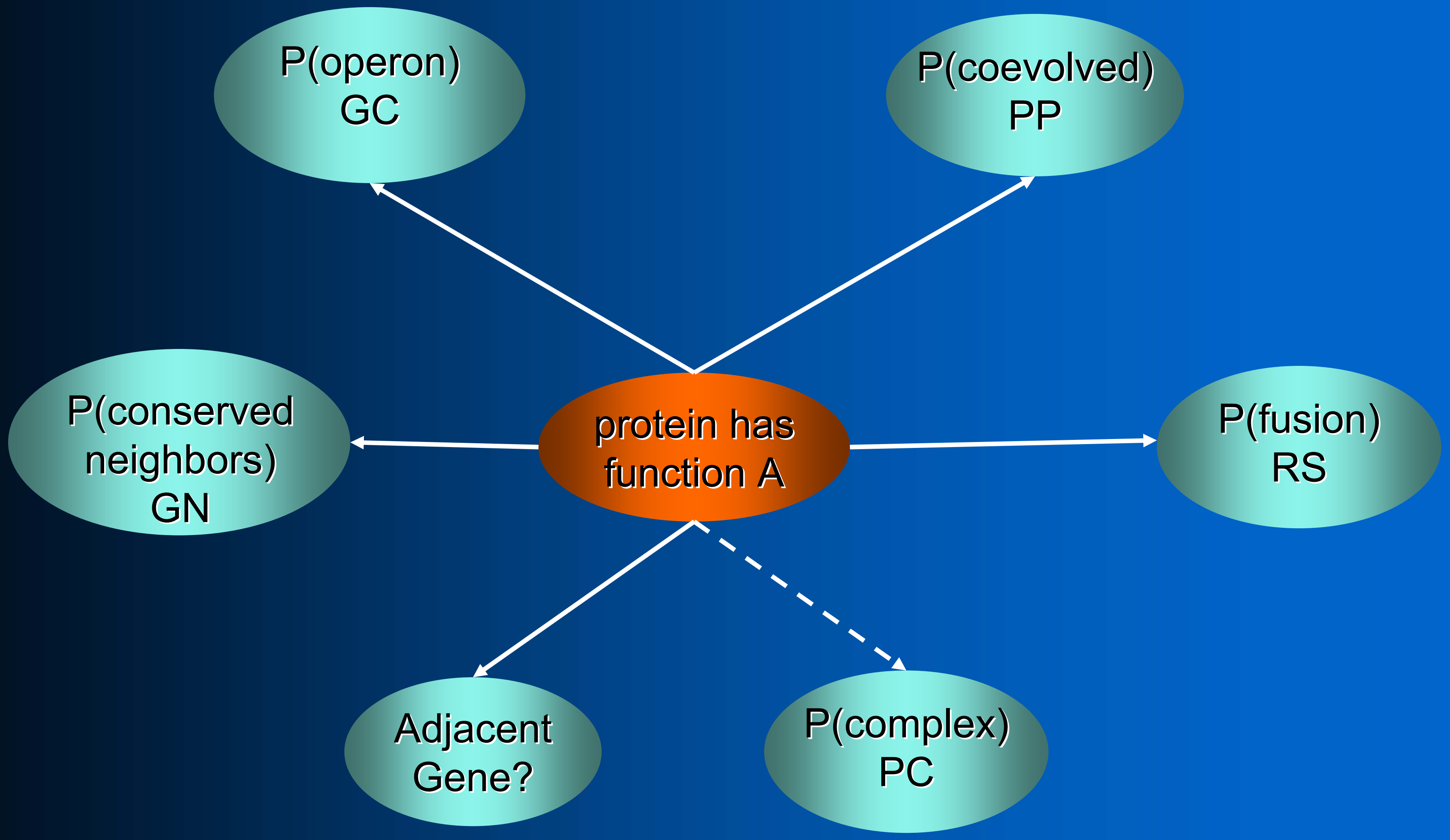
  – Genes catalyzing pathway reactions (generates too many false positives and would probably be considered by biologist as candidate anyway)

Principle behind PHFiller-GC

- Directon genes: 38 genes in a directon with a pathway gene
- Excluded pathway genes: cpsG, cpsB, nudD, manA

Gene	PP pairs	GN pairs	RS pairs	GC pairs
cpsG	glmM	~~cpsB~~ wcaI ~~nudD~~ blmS hflB ~~manA~~	None	wcaJ ~~cpsB~~
cpsB	None	wcaI fcl ~~nudD~~ gmd ~~cpsG~~	yihS	~~cpsG~~ wcaI
nudD	None	gmd fcl ~~cpsB~~ ~~cpsG~~ wcaF wcaI	None	wcaI fcl
manA	None	~~cpsG~~ fumA fumC ydgH ydgA tus rstB rstA	None	ydgA

Use Bayesian classifier to evaluate candidates



Validation

- Pathway criteria
  - Contiguous
  - 2 or more reactions
  - 2 or more known enzymes
- EcoCyc
  - 206 pathways
  - 132 pathways meet criteria => 557 reactions
  - 124 reactions removed (enzyme for multiple rxns same pwy)

⇒ **433 reactions for validation**

⇒ **507 enzymes (547 enzymatic reactions)**

Validation

- **5- or 10-fold cross validation**
- **Steps:**
  1. Identify candidates for each reaction (training and test sets)
  2. Generate training distributions (from training set)
  3. Compute probabilities for each reaction (test set)
  4. Evaluate performance
- **Models:**
  - Full model with all features (AD, GN, GC, RS, PP)
  - Individual features
- **Evaluation – fraction of true hits in the top N candidates for each reaction (“How many genes will I have to test?”)**

Evaluation: Fraction of true hits in top N hits

- N hits identified for (each) reaction R
- Sorted by P(has function R)

e.g., galactonate dehydratase

Hits in order of P(has function)
1. G7790-MONOMER (dgoR)
2. GALACTONATE-DEHYDRATASE-MONOMER (dgoD)
3. YIDT-MONOMER (dgoT)
4. G7160-MONOMER (yfaW)
5. G6839-MONOMER (rspA)

Results

- All true hits vs. best hit
- EcoCyc validation
    - Homology vs. genome context
    - Reactions with no homology data
- Validation in other organisms

Best hit vs. all hits in top N candidates

Best hit

- fraction of reactions with at least one true hit in the top N candidates
- How often do I find at least one enzyme for the reaction?

All hits*

- fraction of all true hits in the top N candidates
- How often do I find all enzymes (or complex components) catalyzing the reaction?

e.g., phenylacetate-CoA oxygenase

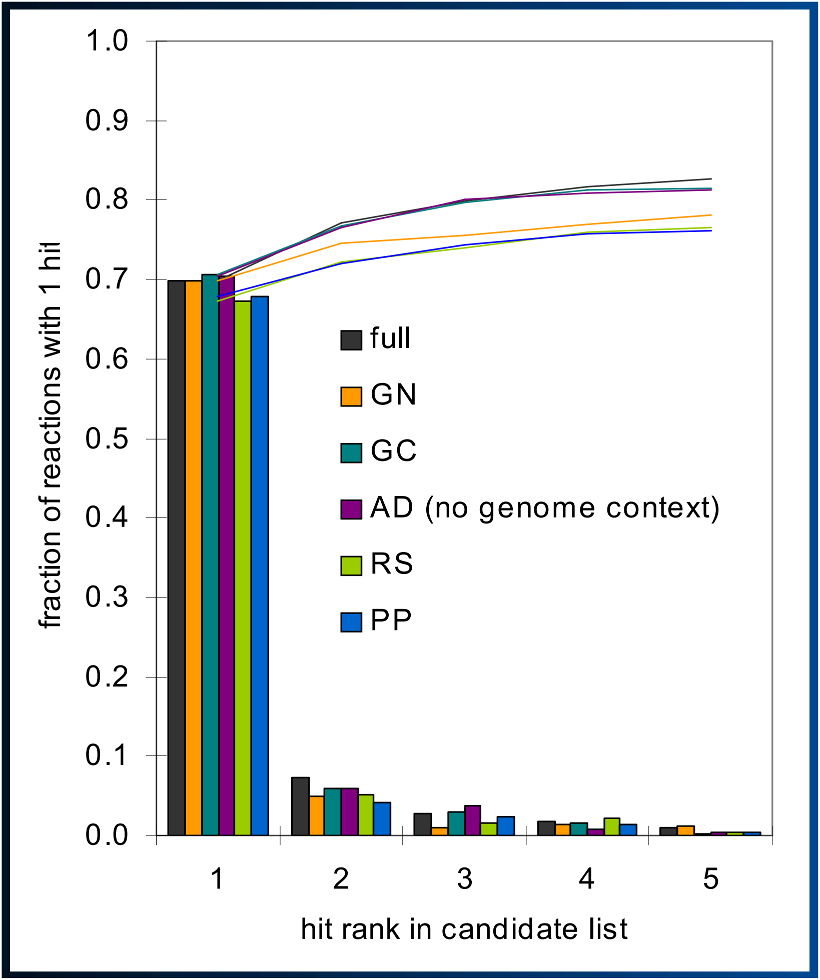
5 monomers in the enzyme complex

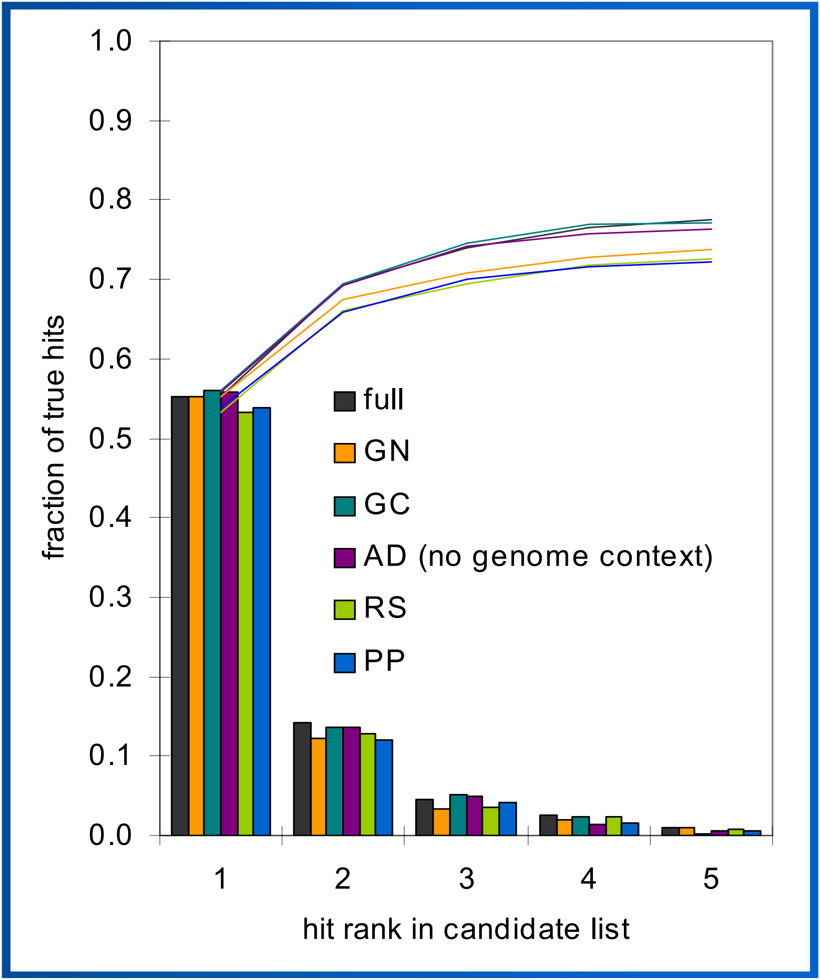
Ranks of the 5 proteins in candidate list: 3, 4, 5, 6, 15

Best hits: only “3” gets counted

All hits: all values contribute to fraction in top ten

Best hit vs. all hits in top N candidates

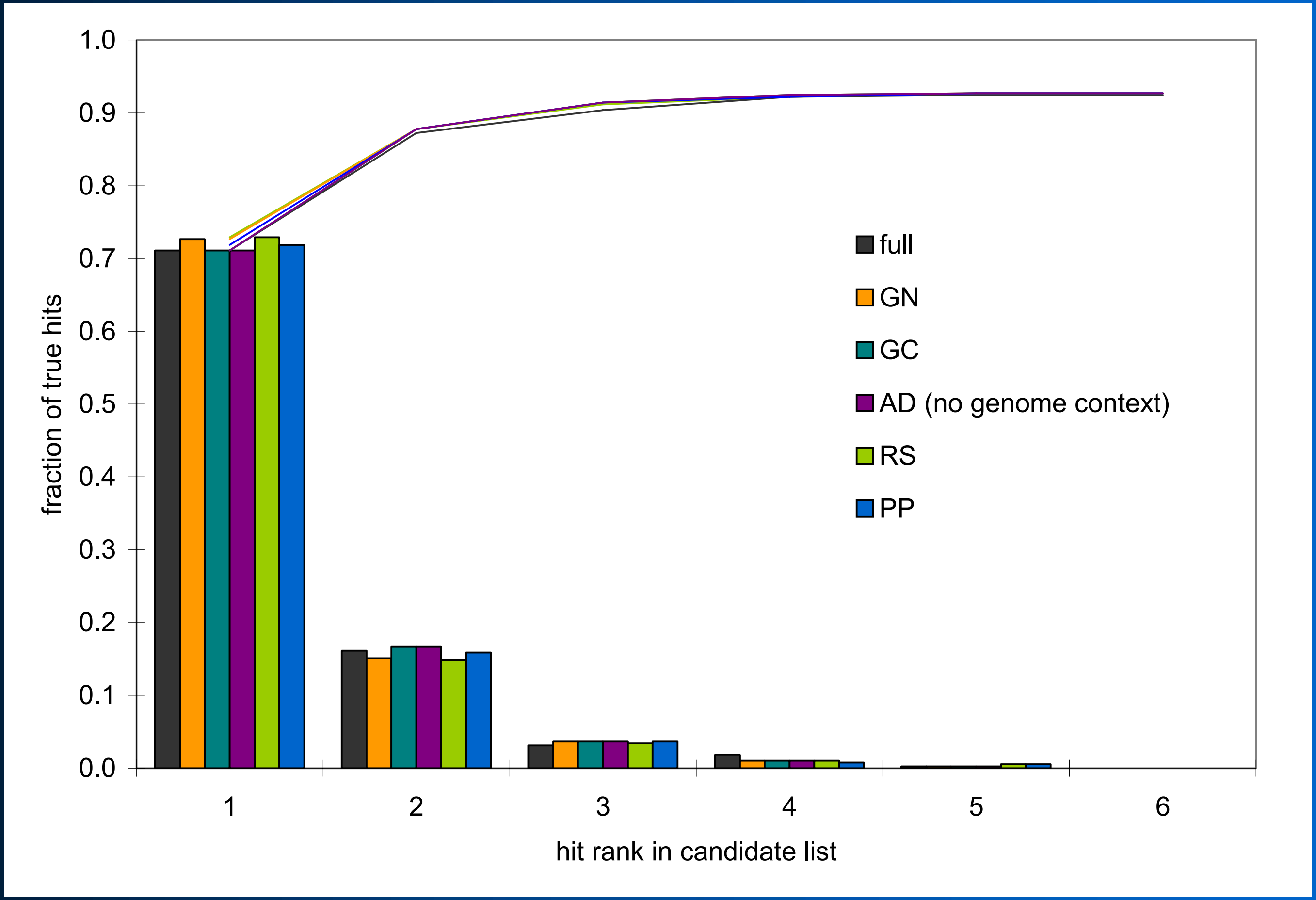




Good results get “diluted” when counting all true hits.

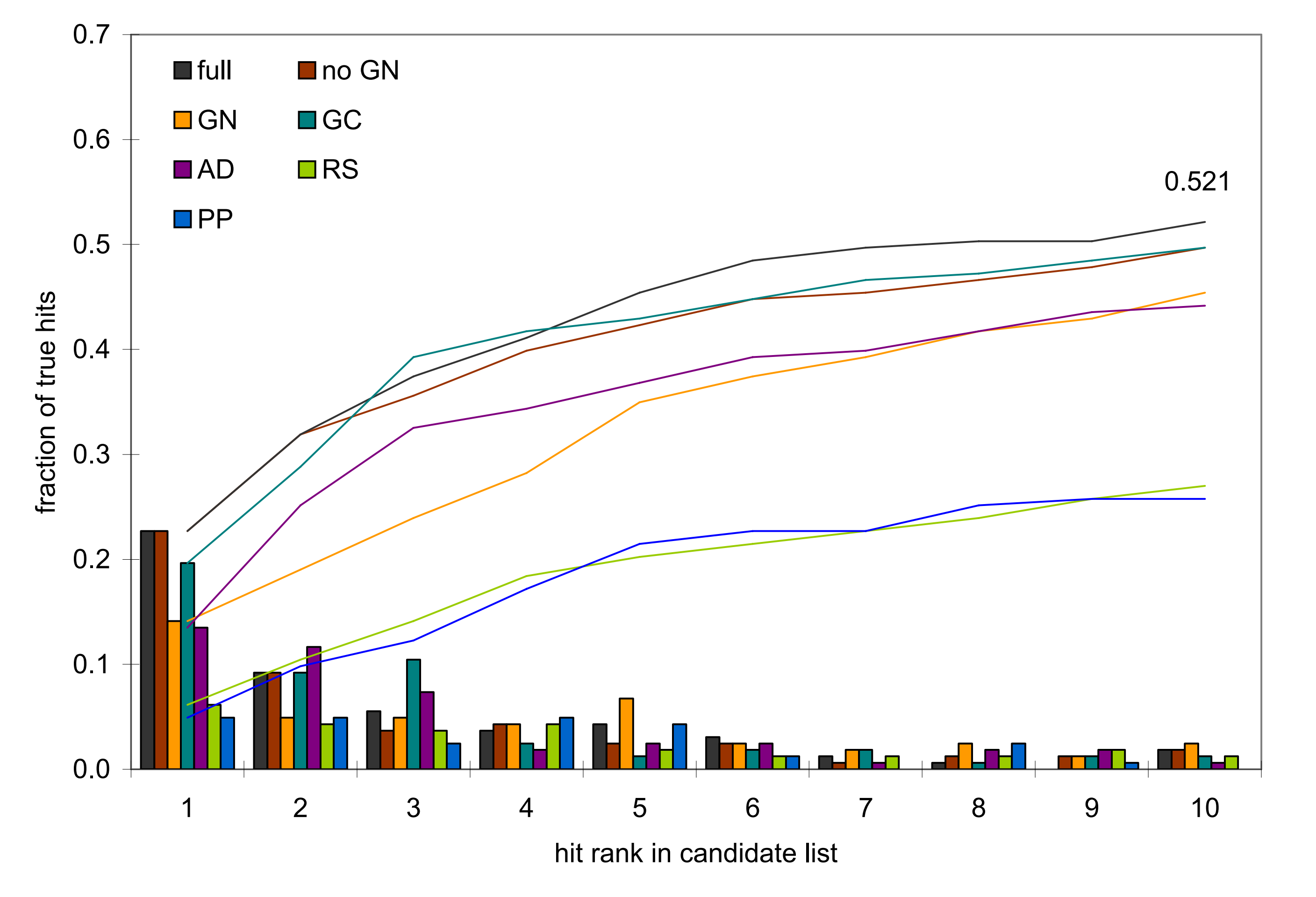
Genome context data can't improve on homology (for EcoCyc anyway)

The 297 reactions with homology data.

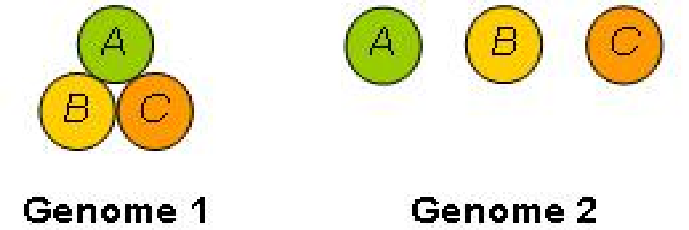


But, for reactions with no homology data, we find 52% of true hits in top ten candidates

The 124 reactions (29% of EcoCyc reactions) without homology data.



Protein complex ortholog method

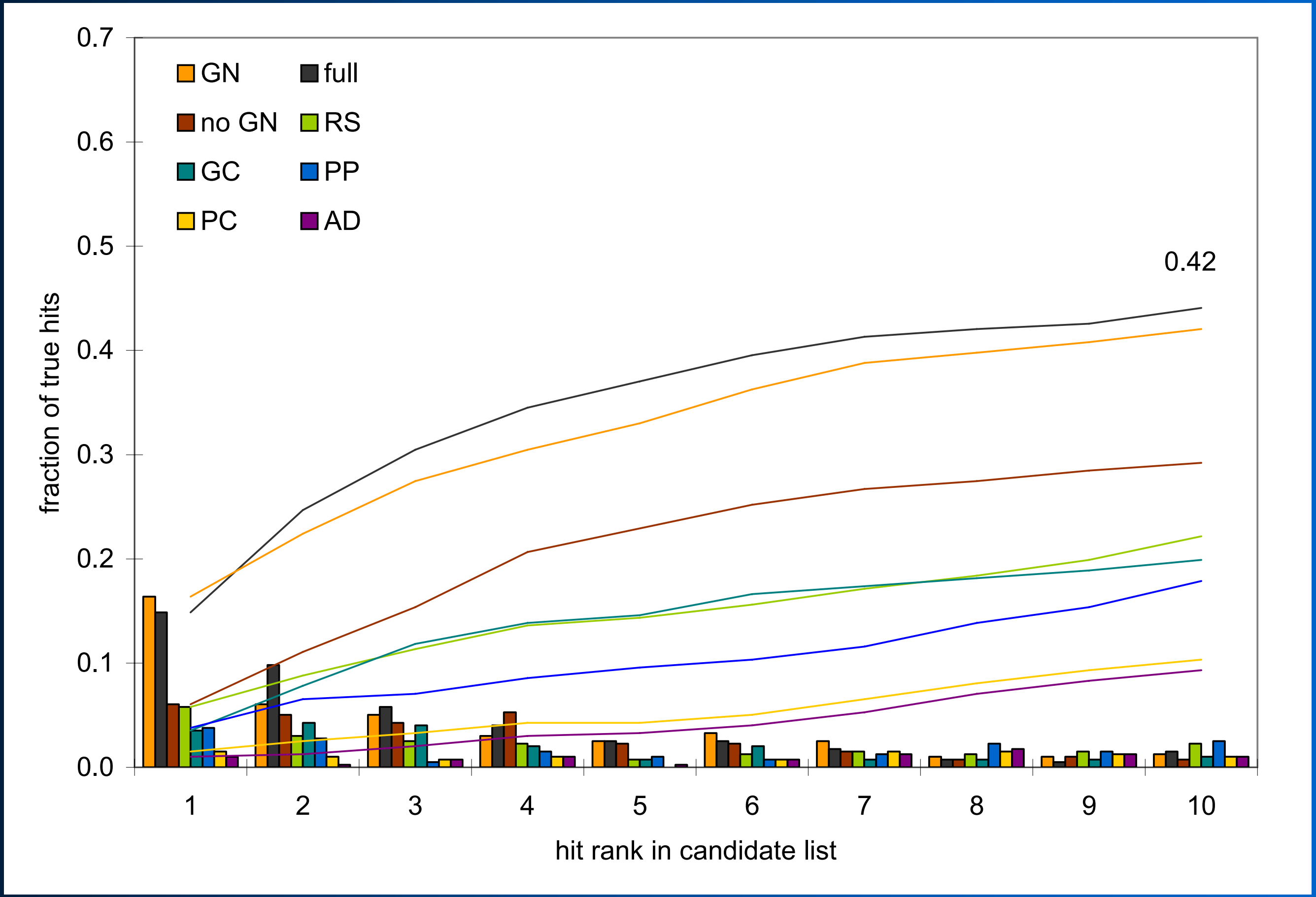


- Analogous to gene fusion method
- If A, B, and C form a known complex in organism A, their orthologs, A’, B’, and C’ are functionally associated in organism B.
- Use complexes from EcoCyc (genome 1)

Results from application to other PGDBs

CauloCyc – *Caulobacter crescentus*

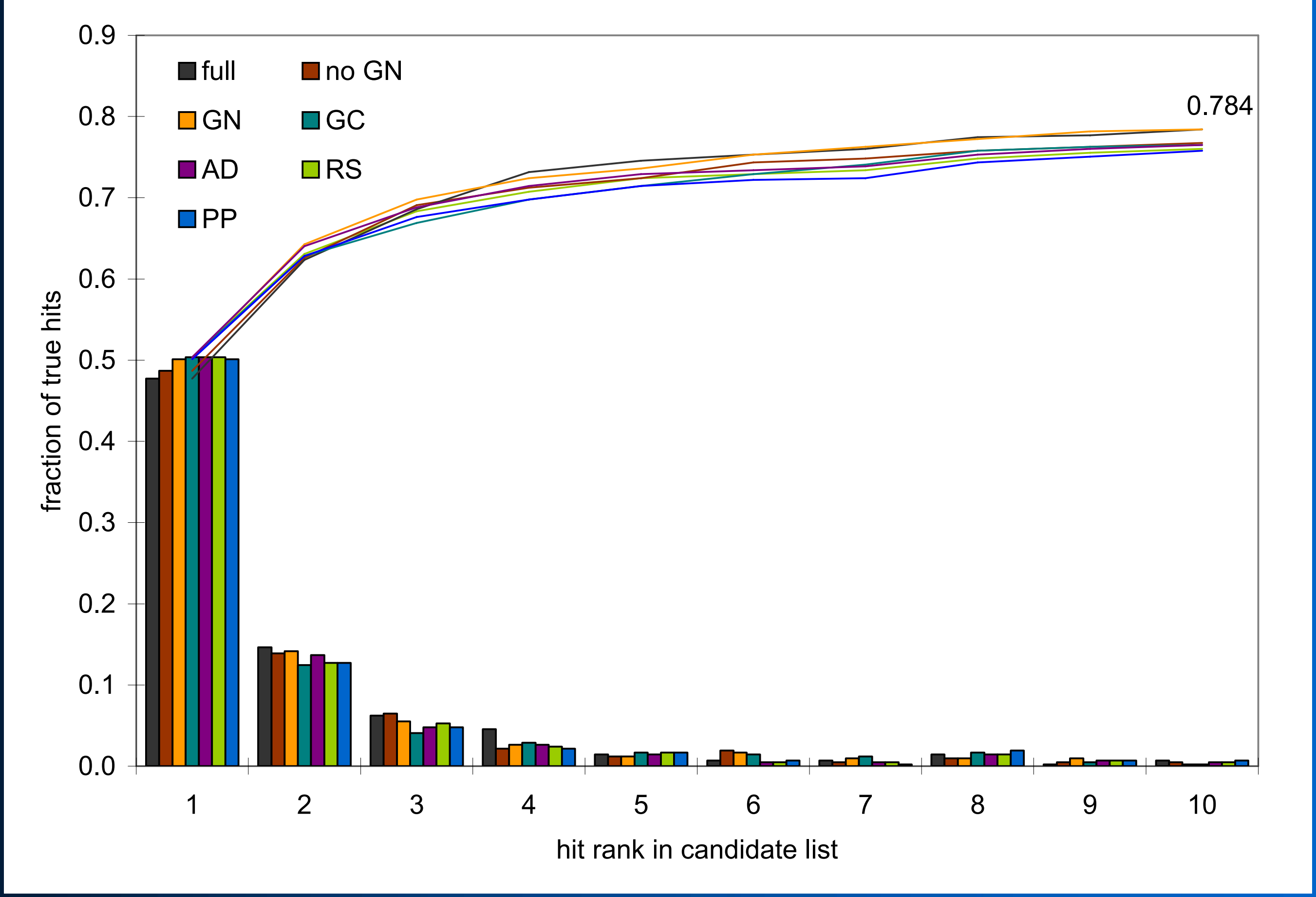
Genome Context data



Results from application to other PGDBs

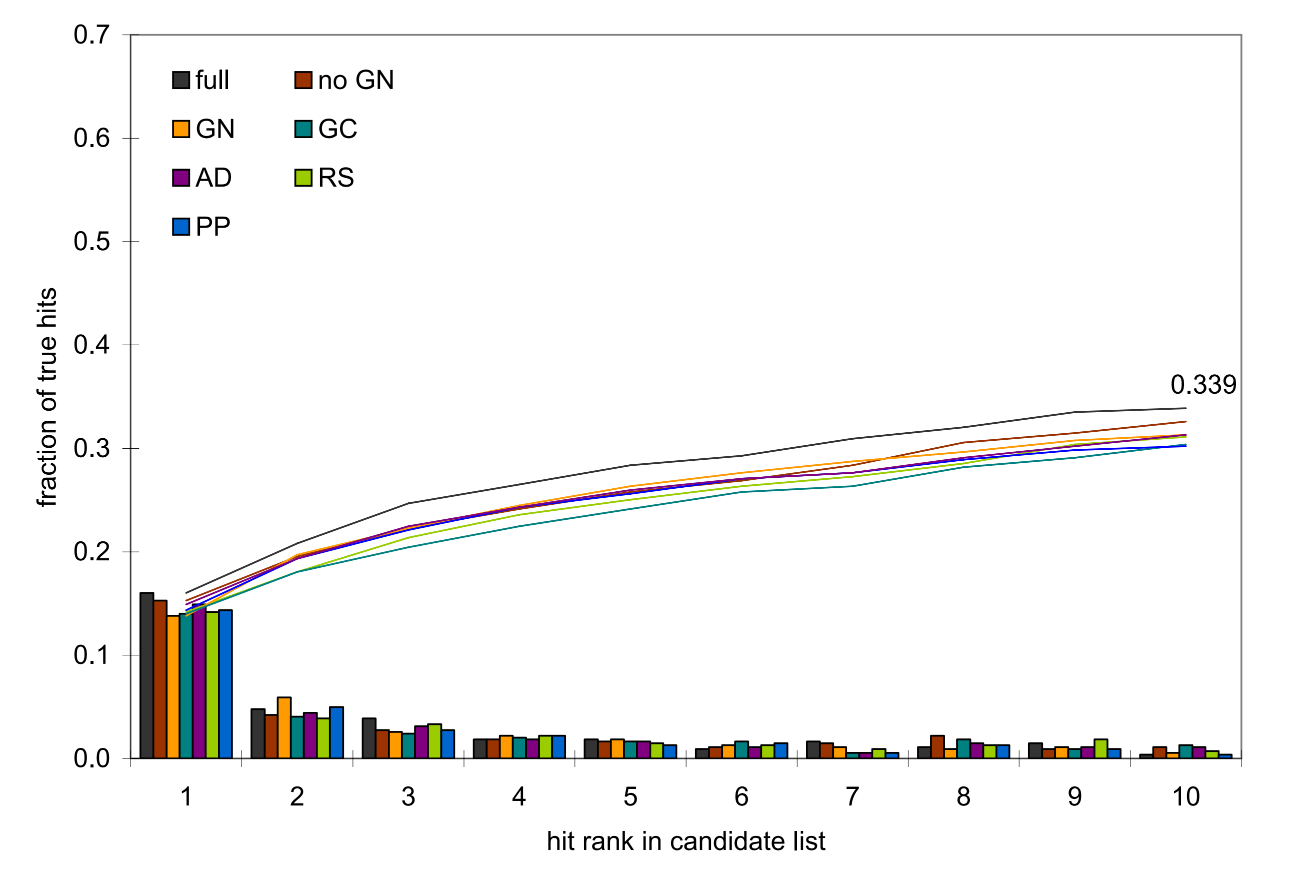
CauloCyc – *Caulobacter crescentus*

Homology + Genome Context data



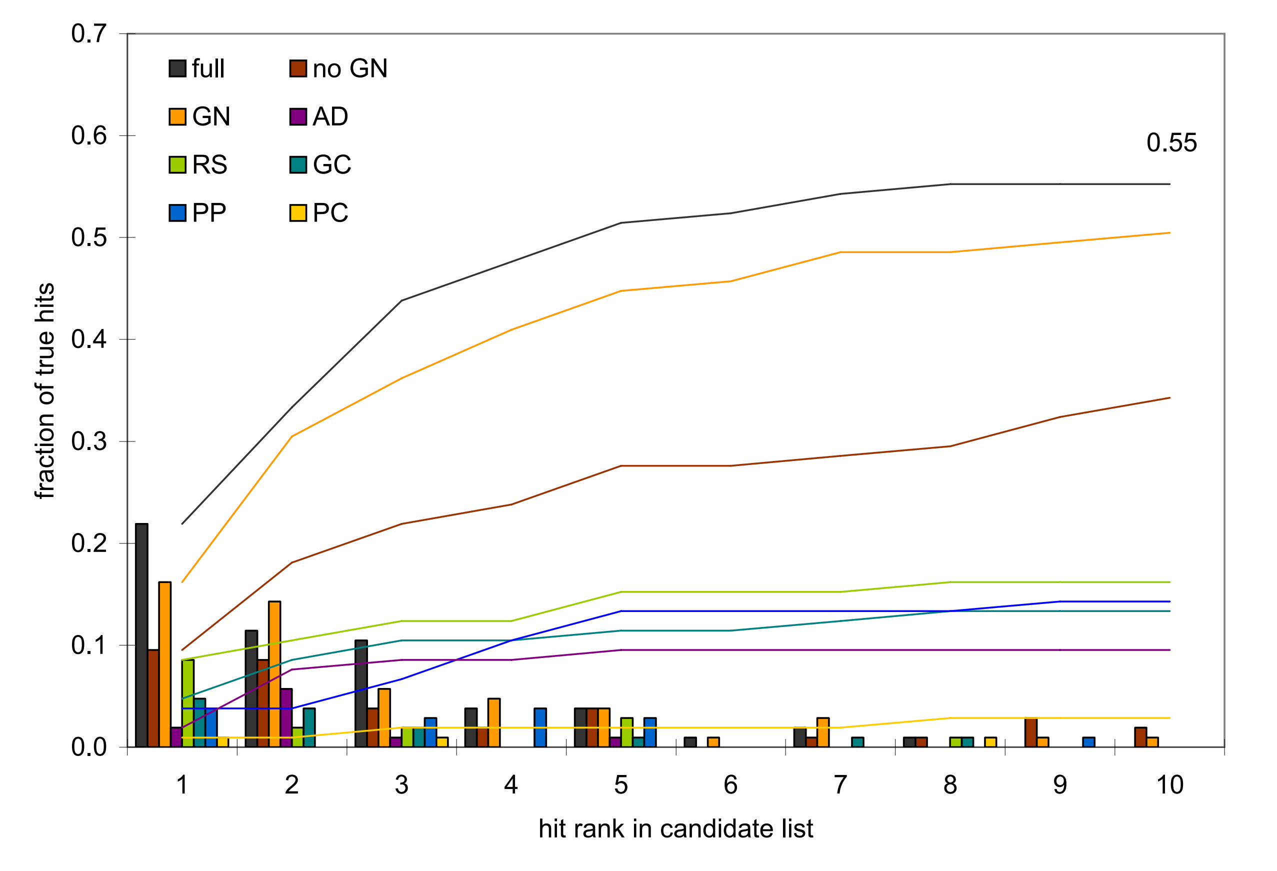
Results from application to other PGDBs

AgroCyc – *A. tumefaciens*
Genome Context data



Results from application to other PGDBs

AfulCyc – *A. fulgidus*
Genome Context data



Results

- For reactions with no homology data, we find all true hits in the top 10 candidates 52% of the time.
- We find the best hit in the top 10 candidates 58% of the time.
- When homology data is available, genome context data does not help.
- Results are comparable for tier 2 and tier 3 organisms.

Implementation in PathoLogic

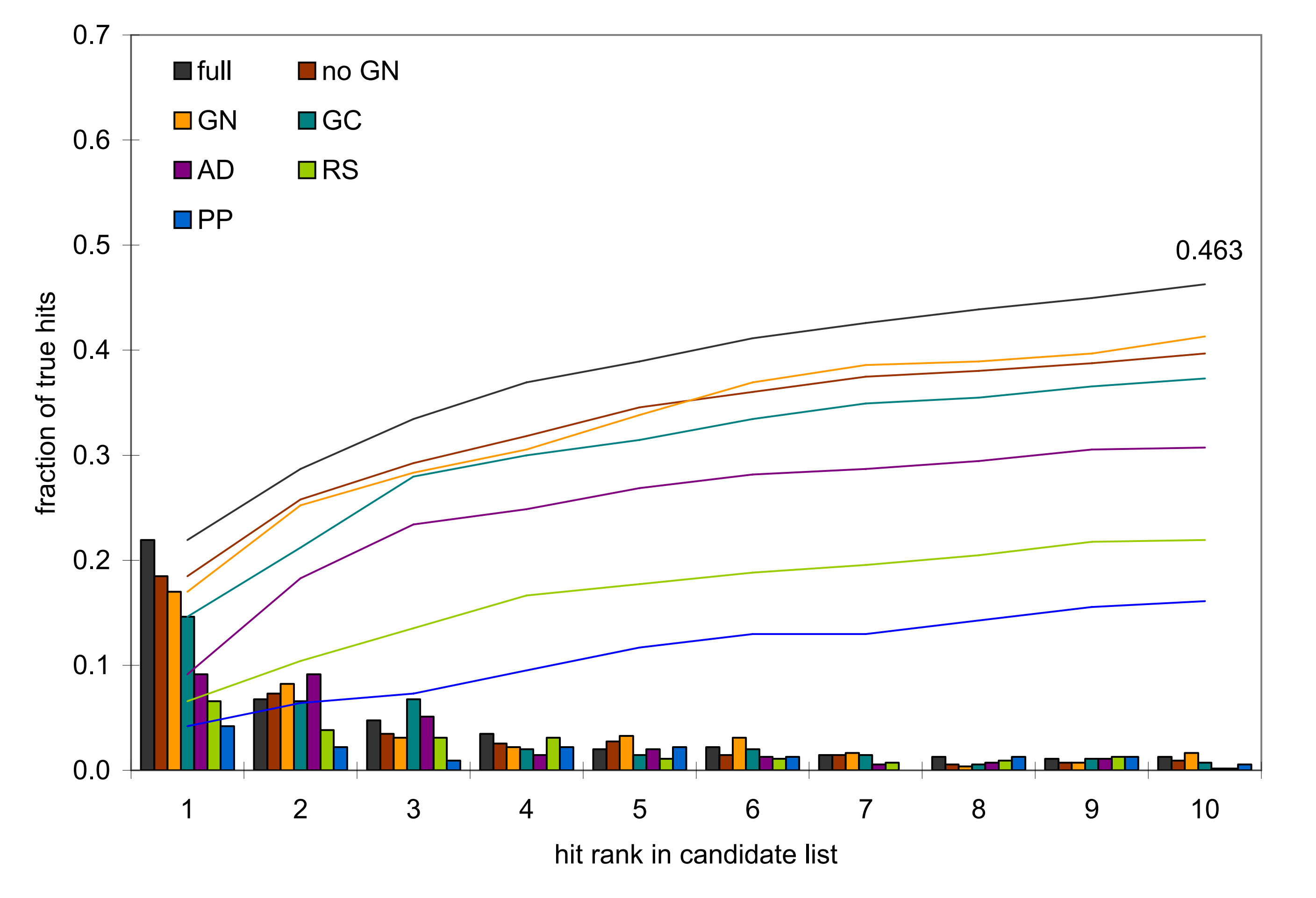
- **Current implementation**
  - Orthologs from CMR “all vs all”
  - MySQL database stores related pairs and orthologs
  - Can compute gene neighbors and phylogenetic profiles
  - Other data (gene fusions, gene clusters) from Prolinks
- **Possibilities for computing genome context data**
  - CoGenT – Christos Ouzounis (gene fusions, phylo. profiles)
  - Gene clusters – PathoLogic operon predictor (no P-value)
  - STRING

Acknowledgements

- Bioinformatics Research Group
- RR07861 and GM70065 from the National Institutes of Health
- DE-FG03-01ER63219 from the U.S. Department of Energy

Results for all reactions excluding homology data…

All 433 qualifying EcoCyc reactions.



Performs better on Tier 3 orgs than Tier 1/2

PGDB	data	% in top 10	# knowns	# true hits
EcoCyc	H+GC	81.0	433 (all)	547
	H	79.0	433 (all)	547
	GC	44.0	297 (w/ homology)	384
	GC	52.1	136 (no homology)	163
Caulo	GC	45.1	294	390
MtbRv	GC	35.5	257	411
Aful2234	GC	55.2	148	129
Telo197221	GC	58.7	186	224
Ssol2287	GC	54.1	91	148

Are the reactions with and without known sequences somehow different?

- **T-test on 10-fold cross-validation results**
- **Compared "fraction in top 10 candidates"**
  - Rxns without homology data = 52.1%
  - Rxns with homology data = 44.0%

These are not **statistically different!**