http://shah-lab.uchsc.edu

Bioinformatics
School of Medicine
University of Colorado

- High-throuhgput biology is yielding a vast amount of data the genome, transcriptional regulation and proteome of many microbial organisms.
- Yet, the complete metabolic network of even the simplest fully sequence living system has not been elucidated
- We want to develop computational inference methods to enhance the pace of metabolic pathway discovery

- Elucidating metabolic pathways through computational inference over biomolecular data
- Two intertwined predictive goals ...
  - Analysis: Piecing together plausible views of microbial metabolic pathways
  - Engineering: Rationally designing new metabolic capabilities
- Application to specific biological problems and experimental collaboration

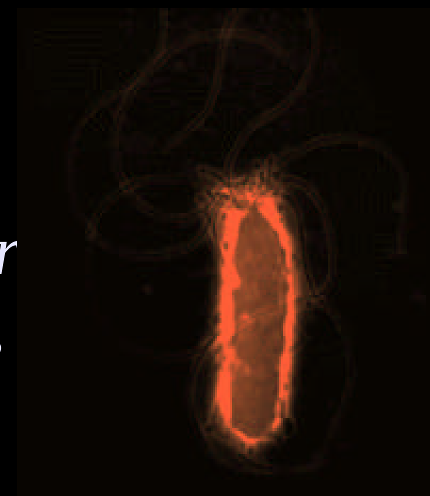
- Traditional methods for experimental determination of pathways are labour-intensive and time-consuming
- There is no high- throughput experimental strategy yet for pathway discovery
- With the availability of whole microbial genomes it is possible to theoretically identify putative proteins and their functions, computationally
- Computational reconstruction of pathways is feasible

- Understanding the complexity of biological systems requires an integrative approach
- Informatics is key for representing biochemical concepts and making them amenable to computation
- Computational approaches are crucial for inference over biological knowledge
- We are developing such a computational framework to support metabolic inference

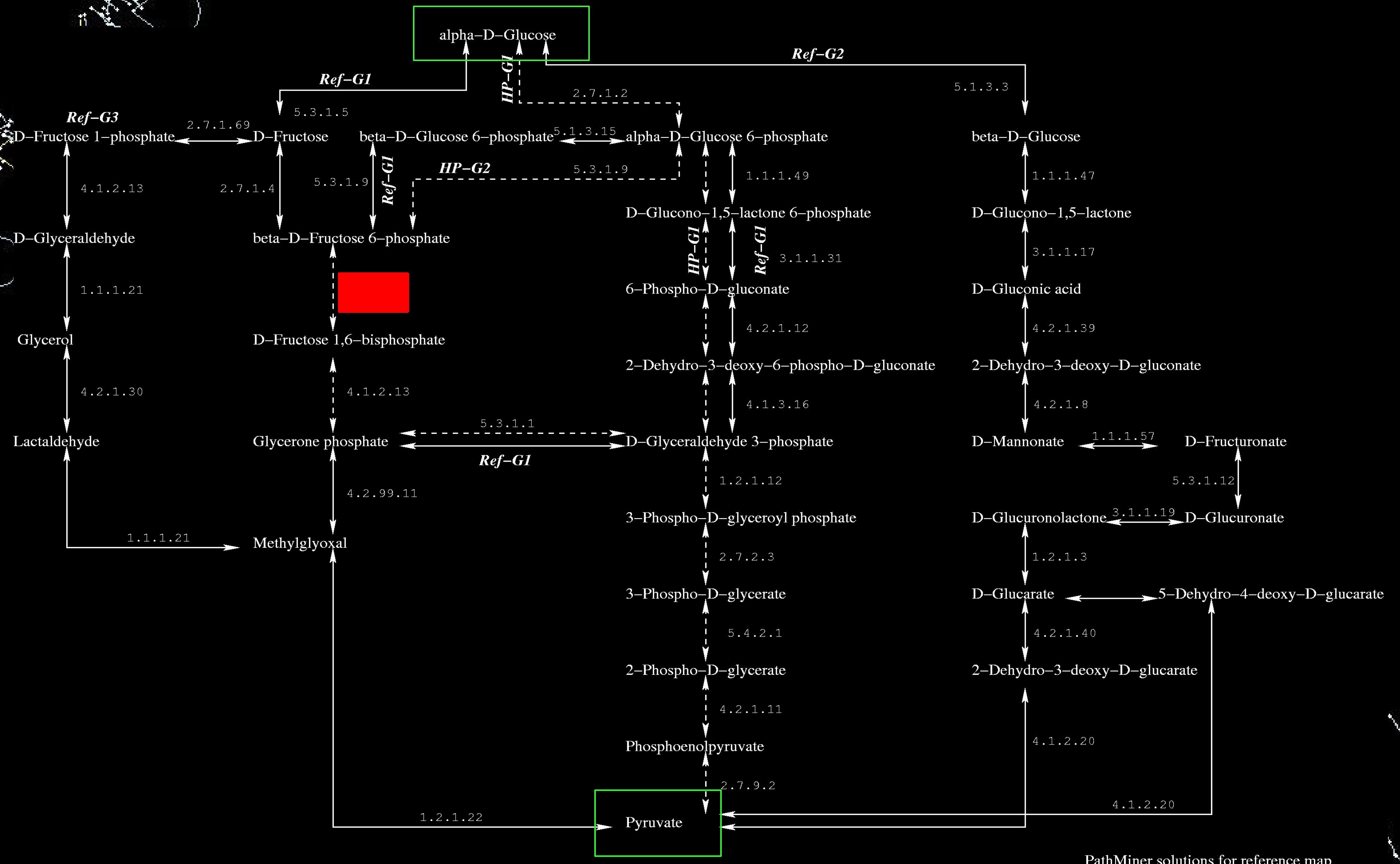
- Many proteins in microbial genomes have not be functionally characterized
- Piecing together functional characterizations of enzymes into plausible metabolic pathways is not straightforward
- Many microbial organisms have new enzymes and novel pathways -- how do we identify these ?

- Given and input compound ***and an output compound, find a series of enzyme-catalyzed transformations that convert the input to the output***
- ***For example: What is the pathway from alpha-d-glucose to pyruvate in E.coli ?***
- ***In E.coli this series of enzyme catalyzed transformations is known as glycolysis***

- Not really !
- For example, *H.pylori is responsible for peptic ulcers; treatments exist but there is no cure*



- *There are many open questions about its its intermediary metabolic pathways*
- ***How is glucose metabolised in H.pylori ?***



In *H.pylori* EDP is preferred over glycolysis

- Even reference or *standard* pathways are not always followed precisely in microbial organisms
- In many organisms alternative biochemical routes or *detours* have been observed (Cordwell, 1999)
- These alternative pathways can use known or unknown enzymes
- How do we infer such pathways in general computationally ?

I. Known pathway but some key enzymes are missing (from annotation)

i. *Homologues have diverged and undetectable by sequence similarity.*

ii. *Enzyme(s) from another superfamily catalyze steps in pathway.*

iii. *There is a non-obvious pathway detour*

II. When no known pathway exists between two compounds then the inference is harder. Consider,

i. A new sequence of known enzyme-catalyzed transformations are involved

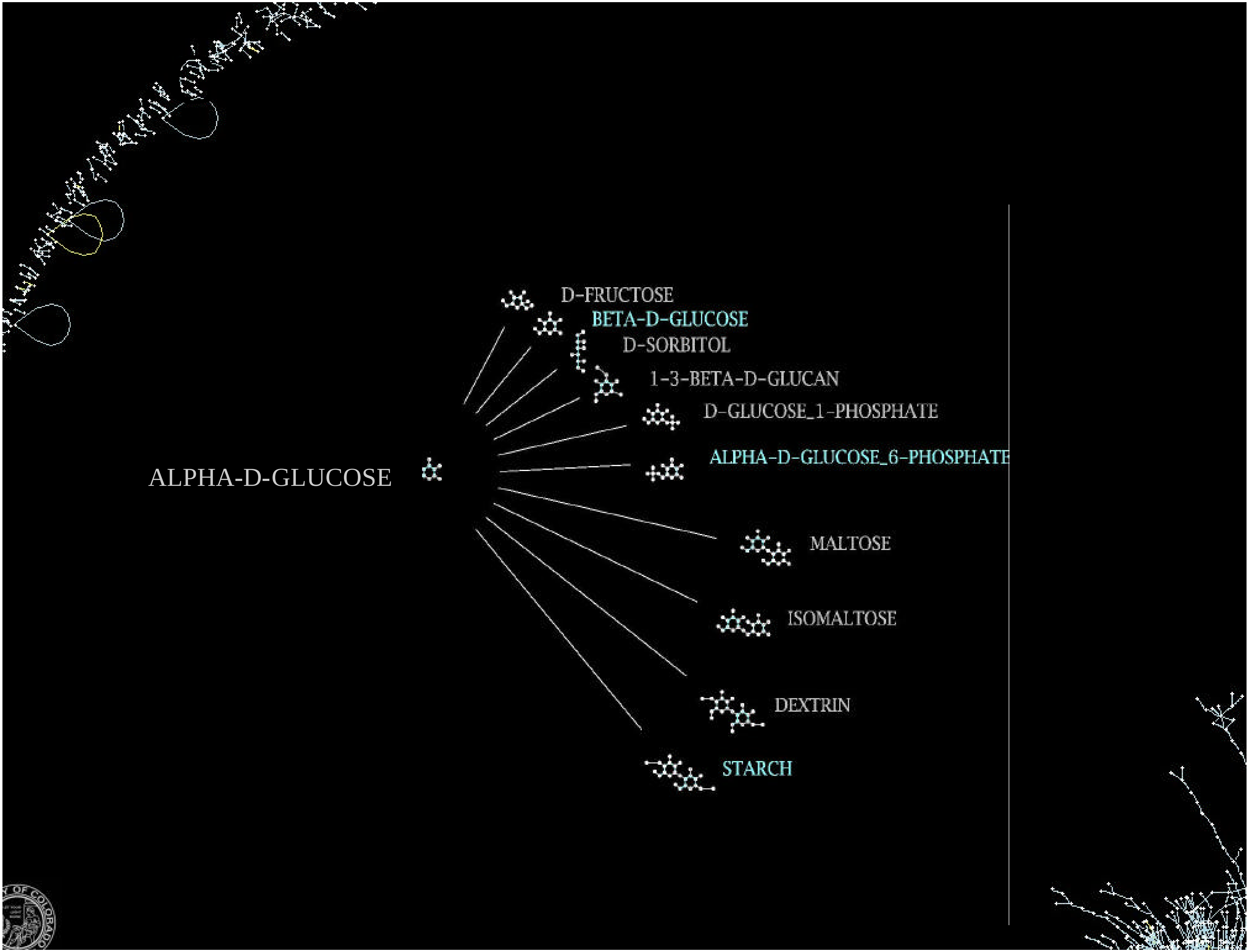
ii. A biochemical pathway must be identified, *de novo*. *That is, a plausible sequence of novel enzymatic functions must be identified*

- We represent biochemistry rationally to enable computations with it and to define novel types of biocatalytic functions
- This representation is the basis for: \
  - Integrating available biomolecular and biochemical data
  - Making inferences about functions and pathways

- We have developed a computational representation of metabolism that resolves biocatalysis into two parts:
    - *The **chemical** component captures the chemical nature of the underlying transformations between compounds.*
    - *The **biological** component captures the enzymatic roles of gene products in terms of specific transformations*

- We abstract metabolism as a hyperdim. state-space in which compounds are points and transformations are state-transitions
- Each compound is represented in *symbolic* terms by its chemical structure components. Eg: carbon dioxide
  - x(CO2) = ((C 1)(O 2)(C=O 2))
- The representation also includes the molecular graph to infer adjacency of any atom or bond
- We have 10,429 compounds from KEGG

- Transformations are state-transitions captured by vector differences between states
- Transformation between alpha-D-glucose-6-phosphate (adg6p) and alpha-D-glucose (adg) is represented as:
  - T(adg6p,adg) = x(adg6p)- x(adg)
    = ((P 1)(O 4)(P-O 3))
- We build transformations from 5,241 reactions


ALPHA-D-GLUCOSE
D-FRUCTOSE
BETA-D-GLUCOSE
D-SORBITOL
1-3-BETA-D-GLUCAN
D-GLUCOSE_1-PHOSPHATE
ALPHA-D-GLUCOSE_6-PHOSPHATE
MALTOSE
ISOMALTOSE
DEXTRIN
STARCH

- Each transformation is associated with enzymes
- Enzymes are described by EC numbers, gene names
- Enzymes can catalyze multiple transformations
- We have around 3,081 defined enzymes

- For each organism, we have the complete set of putative proteins and their assigned functions, including:
  - Enzymes
  - Transporters
- We also have all sequence data from SwissProt and GenBank
- We have the complete genomes for 100 organisms

- By integrating a large amount of metabolic information we can now make inferences with it:
  - *Predict metabolic pathways from genomic data by finding plausible biochemical routes*
  - *Predict biocatalytic functions from protein superfamilies to suggest possible functions of putative protein (from genomic data)*

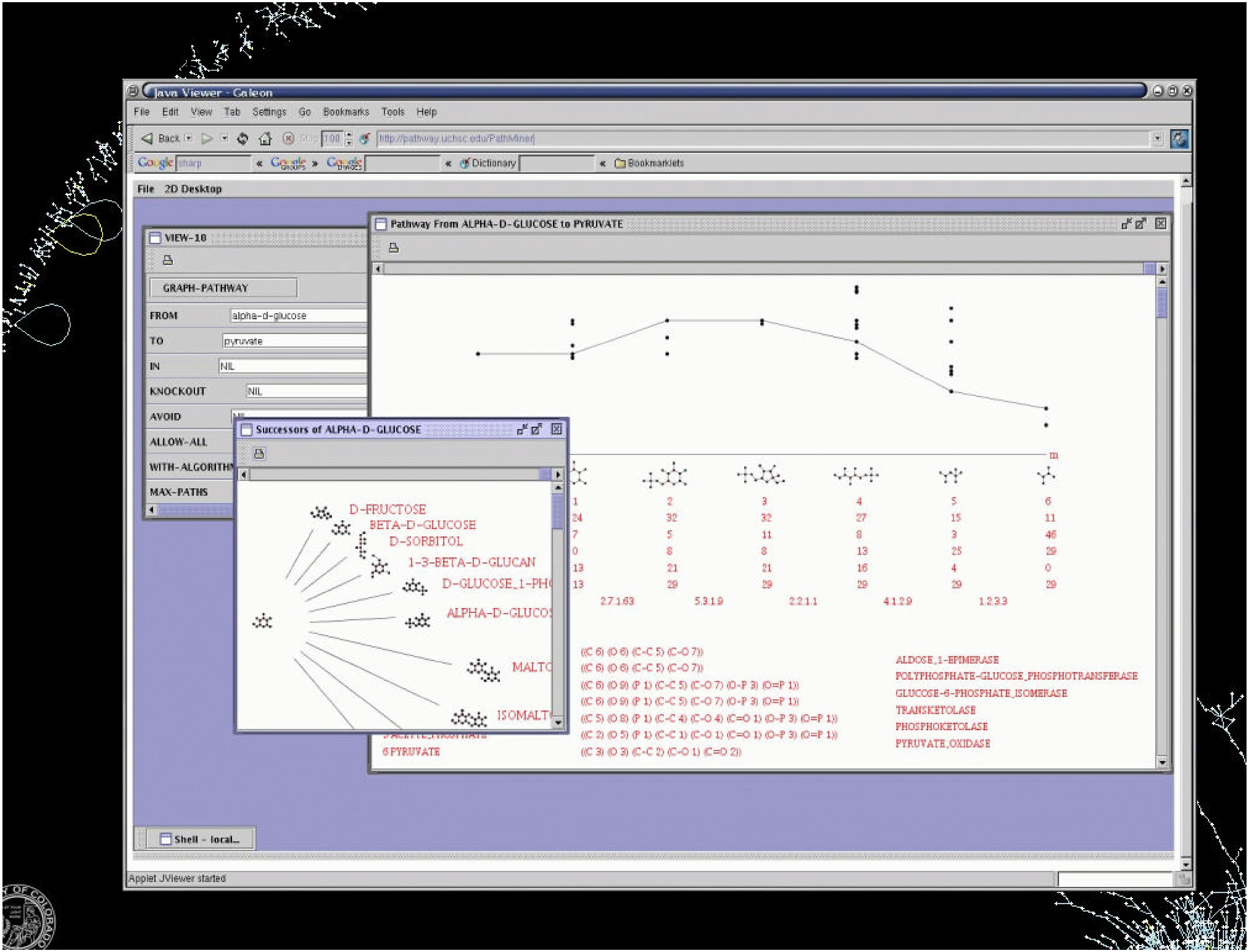
- *Since the sequencing of the first microbial genome, H.influenza, a number of computational methods have been developed to reconstruct reference pathways. Eg. Magpie, PathoLogic, and WIT*
- *Reconstruction is an important starting point for understanding pathways in an organism but there are generally many missing enzymes and gaps in such pathways*
- *We needed strategy to infer new pathways*

- Consider our of metabolism: compounds are states, transformations are state-transitions, and compounds have chemical successors
- We elucidate a metabolic pathway computationally by state-space search
- Each predicted pathway is series of state-transitions
- This produces a combinatorially large number of possible solutions. How can we pick a reasonable subset ?

- *Heurisic search is an informed search technique that uses a best-first algorithm to explore a state-space to find a pathway from initial to final state.*
- *As opposed to blind search (BFS or DFS), informed search methods use an evaluation function (F) to measure the cost of a path*
- *F can be calculated in different ways:*
    - *Greedy - minimize cost to goal (F=H)*
    - *A* - minimize sum of cost so far (G) and cost to goal (F=G+H)*

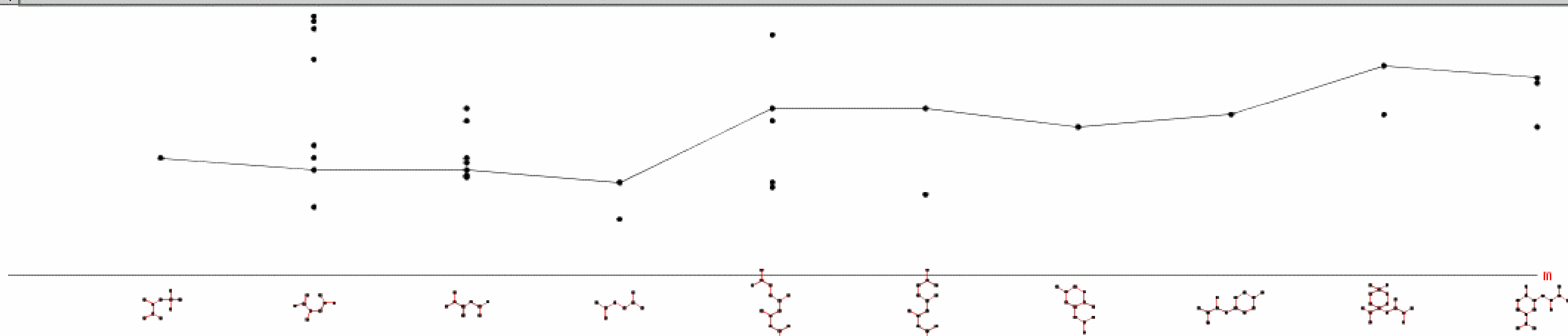
- To predict metabolic pathways by heuristic search, we must calculate the heuristic evaluation function, ***F***
- In general, there are complex factors that determine the cost of a pathway. We wanted a simple concept to compute F
- We decided to test the chemical distance between states to estimate biochemical cost of a pathway from x(0) to x(L), where x(m) is an intermediate state in the pathway:
    - F(0,m,L) = G(0,m) + H(m,L)

- Based on this algorithm, have developed PathMiner, an interactive computational framework for automated metabolic pathway elucidation
- A* search used in PathMiner always finds a pathway that is optimal in ***F*, *not the shortest pathway, and A* search is significantly faster than blind search***
- ***We are using PathMiner for elucidating***
  - Microbial pathways from genomic annotations
  - Synthetic pathways for engineering


Java Viewer - Galeon
File Edit View Tab Settings Go Bookmarks Tools Help
Back
http://pathway.uchsc.edu/PathMiner
Dictionary
Bookmarklets
File 2D Desktop
VIEW-10
GRAPH-PATHWAY
FROM alpha-d-glucose
TO pyruvate
IN NIL
KNOCKOUT NIL
AVOID
ALLOW-ALL
WITH-ALGORITHM
MAX-PATHS
Pathway From ALPHA-D-GLUCOSE to PYRUVATE
Successors of ALPHA-D-GLUCOSE
D-FRUCTOSE
BETA-D-GLUCOSE
D-SORBITOL
1-3-BETA-D-GLUCAN
D-GLUCOSE_1-PH
ALPHA-D-GLUCOS
MALTO
ISOMALT
1 2 3 4 5 6
24 32 32 27 15 11
7 5 11 8 3 46
0 8 8 13 25 29
13 21 21 16 4 0
13 29 29 29 29 29
2.7.1.63 5.3.1.9 2.2.1.1 4.1.2.9 1.2.3.3
((C 6) (O 6) (C-C 5) (C-O 7))
((C 6) (O 6) (C-C 5) (C-O 7))
((C 6) (O 9) (P 1) (C-C 5) (C-O 7) (O-P 3) (O=P 1))
((C 6) (O 9) (P 1) (C-C 5) (C-O 7) (O-P 3) (O=P 1))
((C 5) (O 8) (P 1) (C-C 4) (C-O 4) (C=O 1) (O-P 3) (O=P 1))
((C 2) (O 5) (P 1) (C-C 1) (C-O 1) (C=O 1) (O-P 3) (O=P 1))
((C 3) (O 3) (C-C 2) (C-O 1) (C=O 2))
ALDOSE_1-EPIMERASE
POLYPHOSPHATE-GLUCOSE_PHOSPHOTRANSFERASE
GLUCOSE-6-PHOSPHATE_ISOMERASE
TRANSKETOLASE
PHOSPHOKETOLASE
PYRUVATE_OXIDASE
6 PYRUVATE
Shell - local...
Applet JViewer started

- We are testing PathMiner by investigating pathways in different microbes: *H.pylori*, *D.radiodurans and S.oneidenosis MR-1*
- *In H.pylori we found a number of pathways that are congruent with experimentally determined pathways, including:*
  - *Glucose metabolism*
  - *Pentose phosphate pathway*
  - *TCA*

Pathway 2 From PHOSPHOENOLPYRUVATE to CHORISMATE

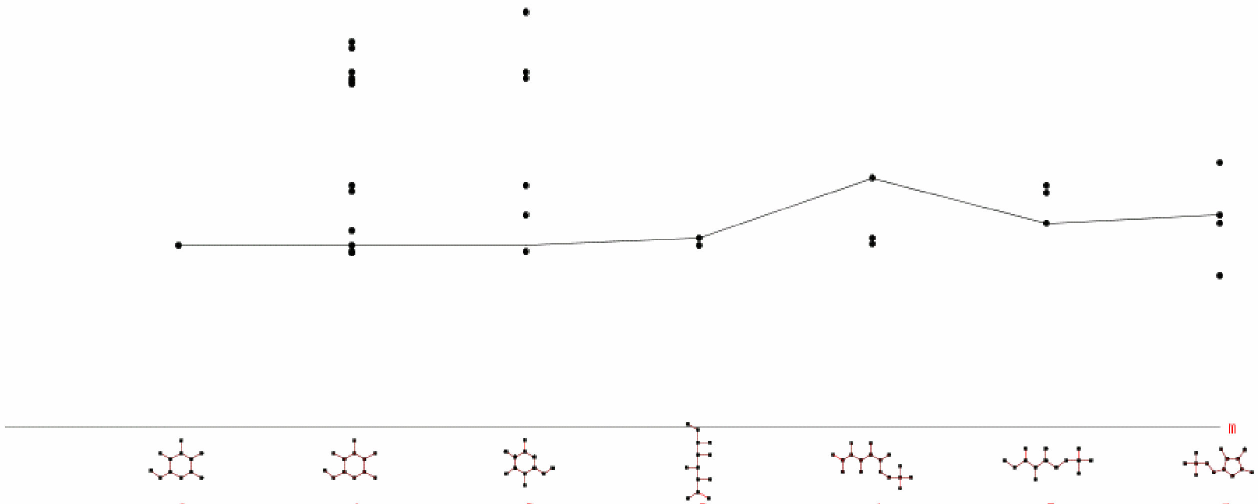


m	0	1	2	3	4	5	6	7	8	9
Features	19	17	17	15	27	27	24	26	34	32
Successors	10	10	4	9	3	2	2	3	4	5
G(0,m)	0	12.0	14.0	18.0	492.56506	492.56506	501.56506	505.56506	515.56506	523.56506
H(m,L)	23	17	15	17	9	9	8	6	8	0
F(0,m,L)	23	29.0	29.0	35.0	501.56506	501.56506	509.56506	511.56506	523.56506	523.56506

0 PHOSPHOENOLPYRUVATE ((C 3) (O 6) (P 1) (C-C 1) (C-O 2) (C=C 1) (C=O 1) (O-P 3) (O=P 1))
(EC_4.1.1.49) (PHOSPHOENOLPYRUVATE_CARBOXYKINASE_-ATP)
1 OXALOACETATE ((C 4) (O 5) (C-C 3) (C-O 2) (C=O 3))
(EC_1.1.1.37) (MALATE_DEHYDROGENASE)
2 S-MALATE ((C 4) (O 5) (C-C 3) (C-O 3) (C=O 2))
(EC_4.2.1.2) (FUMARATE_HYDRATASE)
3 FUMARATE ((C 4) (O 4) (C-C 2) (C-O 2) (C=C 1) (C=O 2))
(EC_3.7.1.2) (FUMARYLACETOACETASE)
4 4-FUMARYLACETOACETATE ((C 8) (O 6) (C-C 6) (C-O 2) (C=C 1) (C=O 4))
(EC_5.2.1.2) (MALEYLACETOACETATE_ISOMERASE)
5 4-MALEYLACETOACETATE ((C 8) (O 6) (C-C 6) (C-O 2) (C=C 1) (C=O 4))
(EC_1.13.11.5) (HOMOGENTISATE_1-2-DIOXYGENASE)
6 HOMOGENTISATE ((C 8) (O 4) (C-C 5) (C-O 3) (C=C 3) (C=O 1))
(EC_1.13.11.27) (4-HYDROXYPHENYLPYRUVATE_DIOXYGENASE)
7 3-4-HYDROXYPHENYLPYRUVATE ((C 9) (O 4) (C-C 6) (C-O 2) (C=C 3) (C=O 2))
(EC_1.3.1.43 EC_1.3.1.52 EC_1.3.1.12) (CYCLOHEXADIENYL_DEHYDROGENASE 2-METHYL-BRANCHED-CHAIN-ENOYL-COA_REDUCTASE PREPHENATE_DEHYDROGENASE)
8 PREPHENATE ((C 10) (H 1) (O 6) (C-C 8) (C-H 1) (C-O 3) (C=C 2) (C=O 3))
(EC_5.4.99.5) (CHORISMATE_MUTASE)
9 CHORISMATE ((C 10) (O 6) (C-C 6) (C-O 5) (C=C 3) (C=O 2))

plet JViewer started

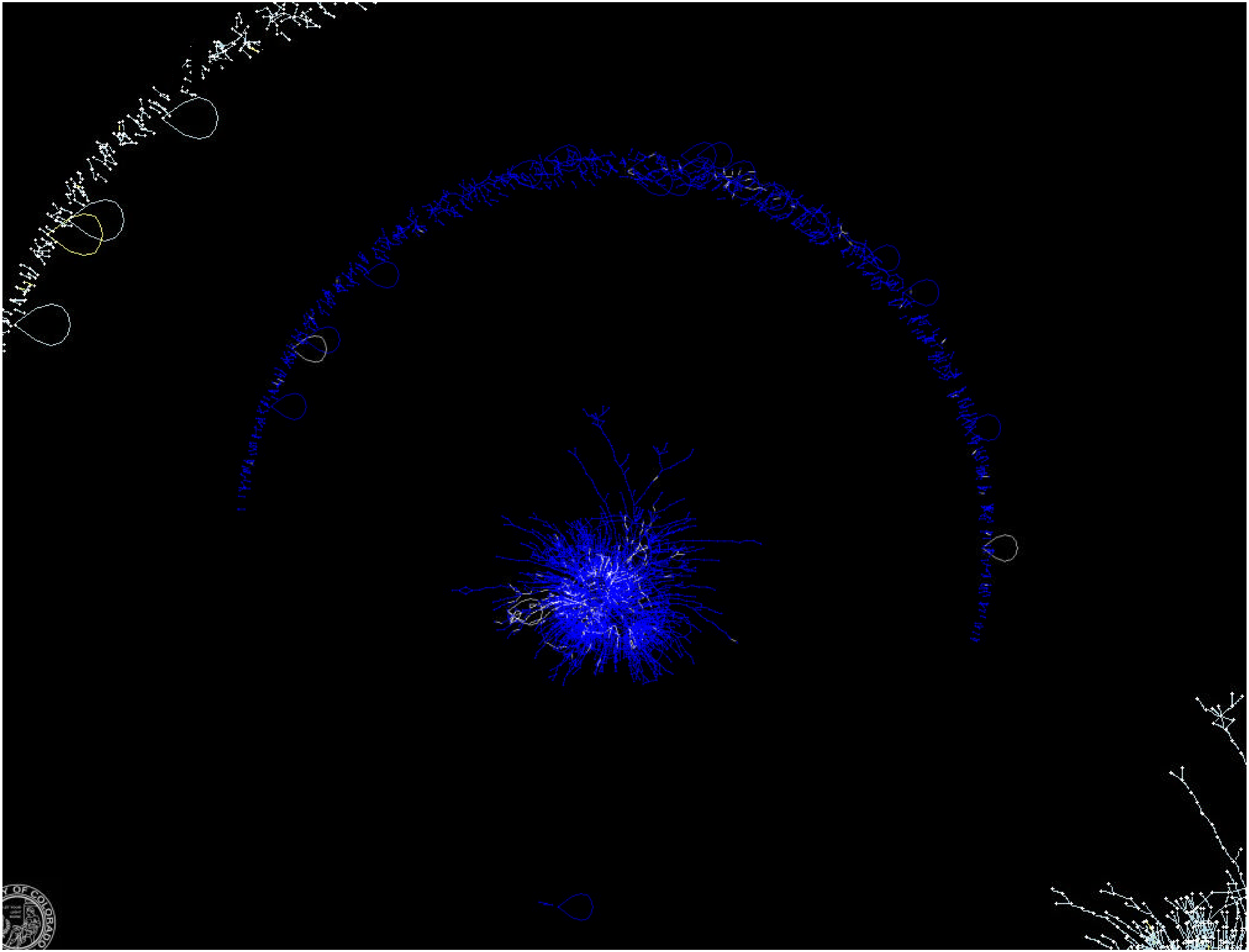
Pathway 4 From ALPHA-D-GLUCOSE to D-RIBOSE_5-PHOSPHATE



m	0	1	2	3	4	5	6
Features	24	24	24	25	33	27	28
Successors	12	7	3	4	4	5	8
G(O,m)	0	0.0	2.0	3.0	11.0	17.0	20.0
H(m,L)	10	10	10	9	5	3	0
F(O,m,L)	10	10.0	12.0	12.0	16.0	20.0	20.0

0 ALPHA-D-GLUCOSE ((C 6) (O 6) (C-C 5) (C-O 7))
(EC_5.1.3.3) (ALDOSE_1-EPIMERASE)
1 BETA-D-GLUCOSE ((C 6) (O 6) (C-C 5) (C-O 7))
(EC_1.1.1.47) (GLUCOSE_1-DEHYDROGENASE)
2 D-GLUCONO-1-5-LACTONE ((C 6) (O 6) (C-C 5) (C-O 6) (C=O 1))
(EC_3.1.1.17) (GLUCONOLACTONASE)
3 D-GLUCONIC_ACID ((C 6) (O 7) (C-C 5) (C-O 6) (C=O 1))
(EC_2.7.1.12) (GLUCONOKINASE)
4 6-PHOSPHO-D-GLUCONATE ((C 6) (O 10) (P 1) (C-C 5) (C-O 6) (C=O 1) (O-P 3) (O=P 1))

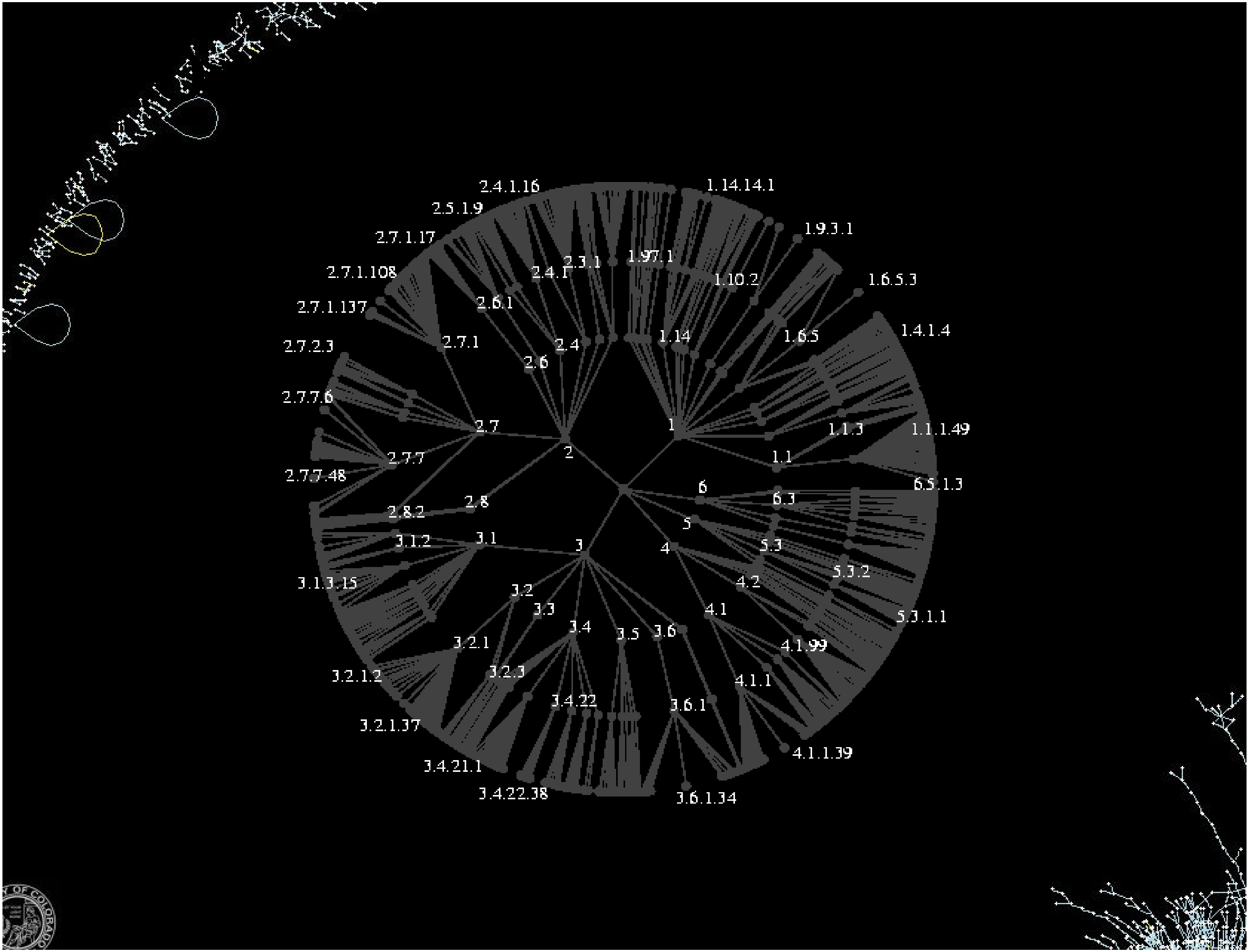
- It is important to consider pathways in the context of broader biochemical processes.
- One way to elucidate the pathways in an organism is to analyze the complete network using functional annotations of genes and known transporters
- We have built a complete network visualizaiton of *D.radiodurans, which we are using to analyze gaps and putative proteins that can fill those gaps.*



- Given the large amount of sequence data how accurately can we infer biocatalytic roles ?
- By systematically computing the correlation of known enzymatic functions with sequence similarity we find:
  - Only 35% of enzymatic functions can be assigned with confidence
  - There are many cases of false positives and false negatives

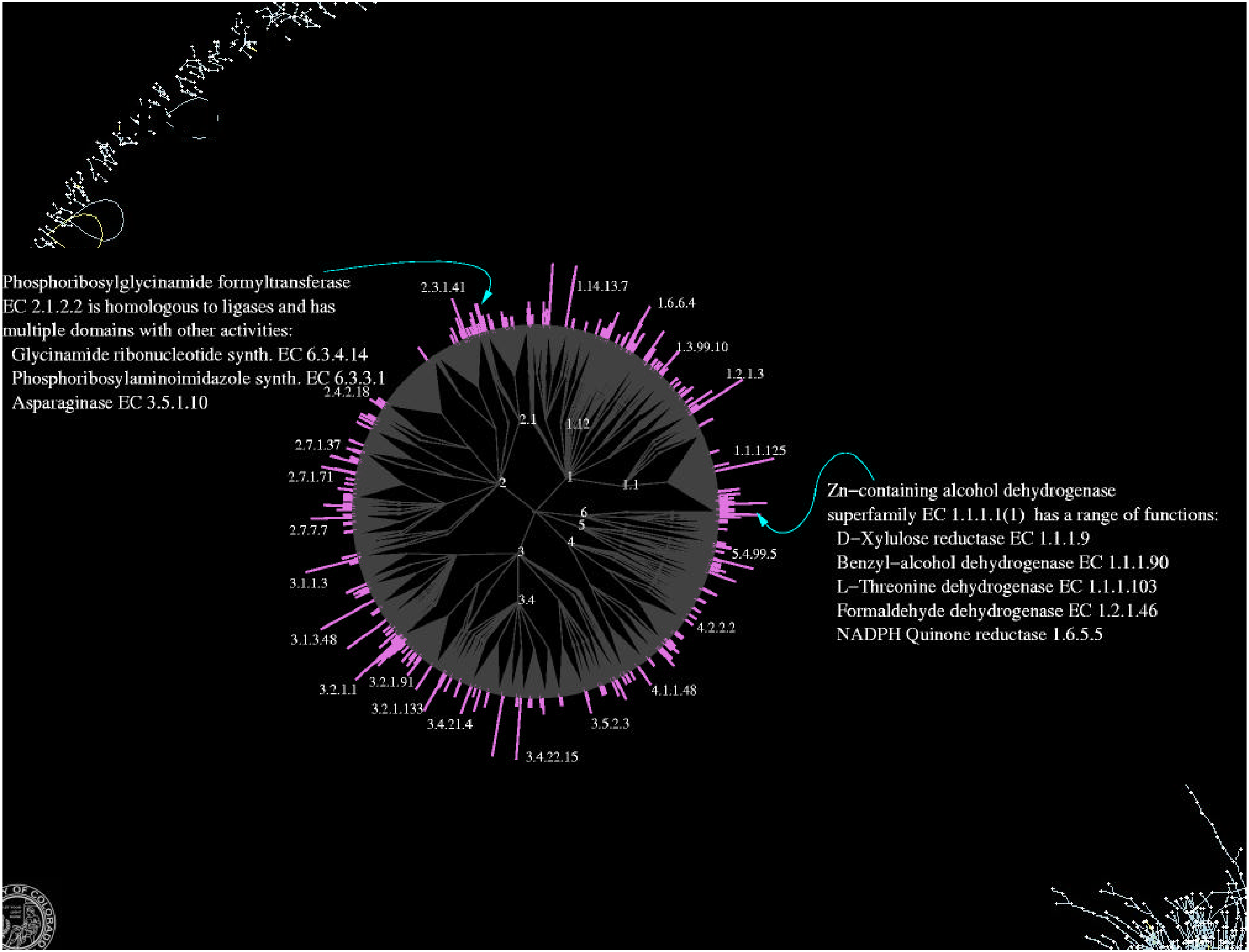
- We begin with a widely used classification scheme, the Enzyme Commission (EC) nomenclature
- The EC defines six broad biocatalytic categories
- Each category has four levels of specification
- There are about 3,500 specific reaction types across all known enzymes
- Though not exhaustive, it covers most enzymes

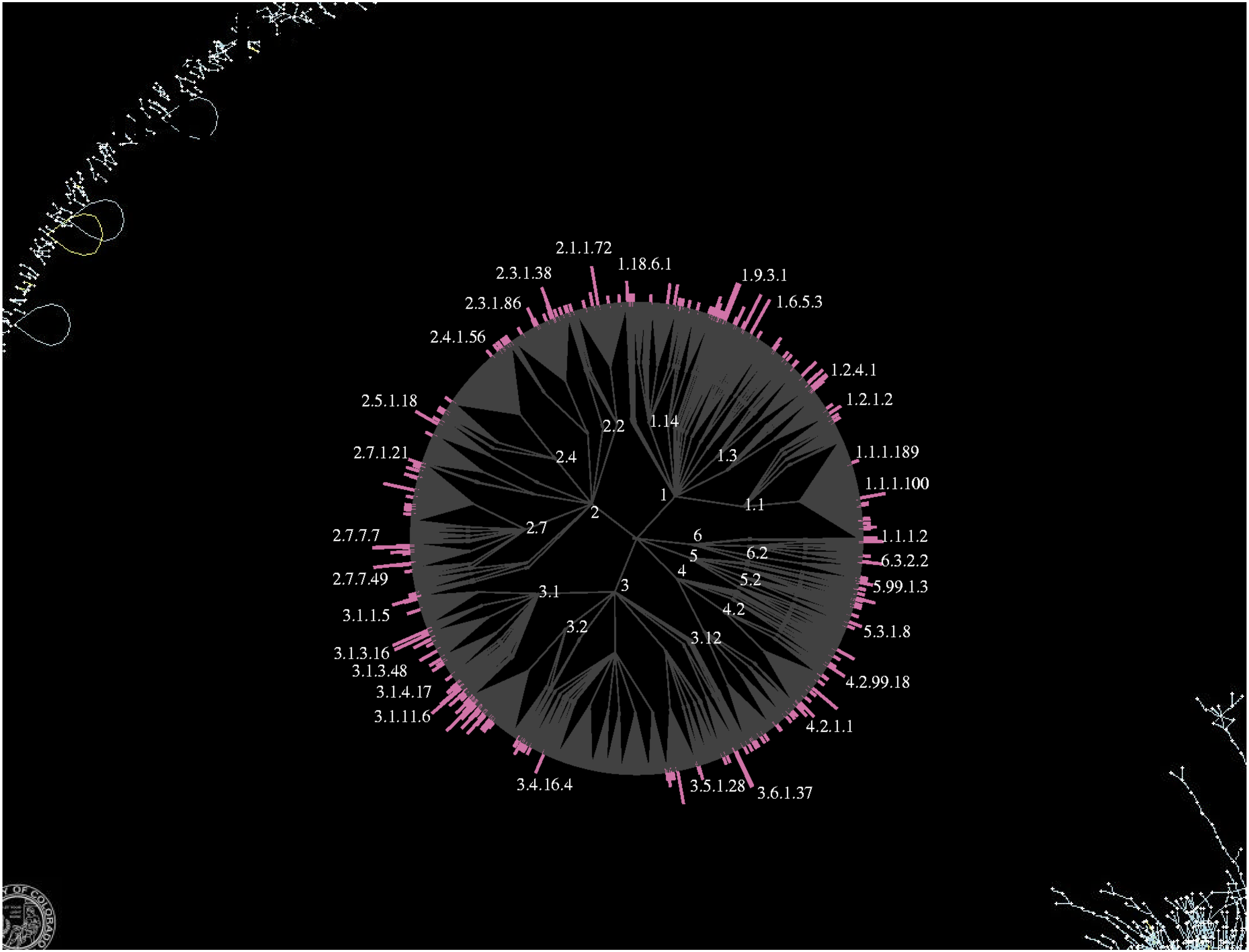
- An EC function is a string of four digits, each number signifying the level in the hierarchy. E.g. EC 1.2.3.4 is oxalate oxidase:
  - Class 1: Oxidoreductase
  - Sub-class 2: Acts on aldehyde or oxo group of donor
  - Sub-sub-class 3: The acceptor is oxygen
  - Serial number 4: The specific reaction: oxalate+O_2<=> hydrogen peroxide+CO_2



2.4.1.16
1.14.14.1
2.5.1.9
1.9.3.1
2.7.1.17
2.3.1
1.97.1
1.10.2
1.6.5.3
2.7.1.108
2.4.1
2.6.1
2.7.1.137
1.4.1.4
1.14
1.6.5
2.7.2.3
2.7.1
2.4
2.6
2.7.7.6
2.7
1
1.1.3
1.1.1.49
2
1.1
2.7.7
2.7.7.48
6
6.5.1.3
6.3
2.8
2.8.2
5
3.1
3.1.2
3
4
5.3
5.3.2
3.1.3.15
4.2
3.2
3.3
4.1
5.3.1.1
3.4
3.5
3.6
3.2.1
4.1.99
3.2.3
3.2.1.2
4.1.1
3.4.22
3.6.1
3.2.1.37
4.1.1.39
3.4.21.1
3.4.22.38
3.6.1.34

- The EC classification is manually derived the differences between levels are not consistent across the functional categories
- The scheme does not capture function uniquely. (Eg. Enzymes that transfer groups share characteristics with ligases)
- The hierarchical organization does not allow complex functions to share multiple characteristics. (Eg. A transferase is like a ligase)
- These factors make EC identifiers difficult to compute with

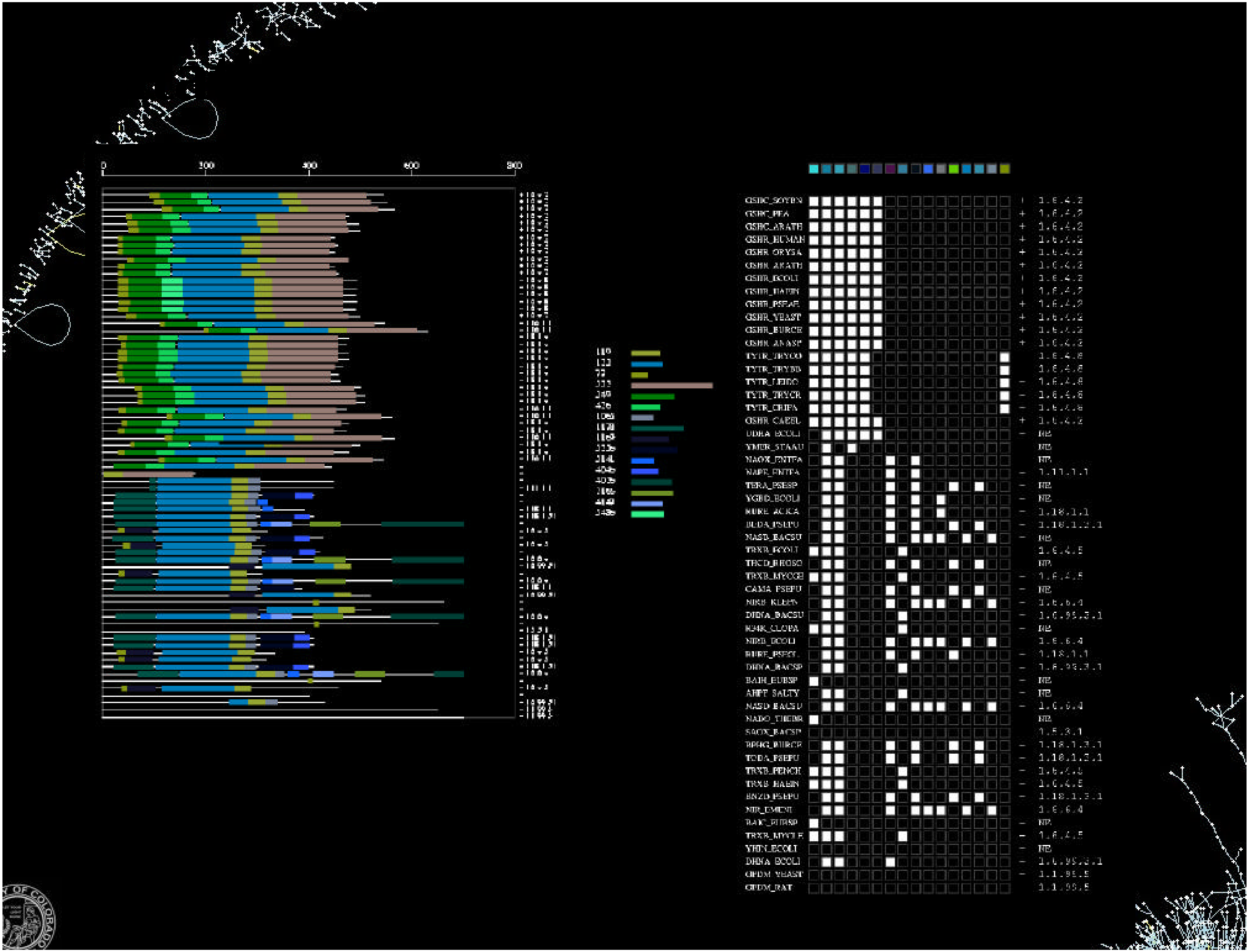

Phosphoribosylglycinamide formyltransferase
EC 2.1.2.2 is homologous to ligases and has
multiple domains with other activities:
Glycinamide ribonucleotide synth. EC 6.3.4.14
Phosphoribosylaminoimidazole synth. EC 6.3.3.1
Asparaginase EC 3.5.1.10
Zn−containing alcohol dehydrogenase
superfamily EC 1.1.1.1(1) has a range of functions:
D−Xylulose reductase EC 1.1.1.9
Benzyl−alcohol dehydrogenase EC 1.1.1.90
L−Threonine dehydrogenase EC 1.1.1.103
Formaldehyde dehydrogenase EC 1.2.1.46
NADPH Quinone reductase 1.6.5.5
2.3.1.41
1.14.13.7
1.6.6.4
1.3.99.10
1.2.1.3
2.4.2.18
2.1
1.12
2.7.1.37
1.1.1.125
2.7.1.71
2
1
1.1
6
5
4
2.7.7.7
3
5.4.99.5
3.1.1.3
3.4
3.1.3.48
4.2.2.2
3.2.1.1
3.2.1.91
4.1.1.48
3.2.1.133
3.4.21.4
3.5.2.3
3.4.22.15


2.1.1.72
1.18.6.1
2.3.1.38
1.9.3.1
1.6.5.3
2.3.1.86
2.4.1.56
1.2.4.1
1.2.1.2
2.5.1.18
1.14
2.2
1.3
2.4
2.7.1.21
1.1.1.189
1
1.1.1.100
1.1
2
2.7
6
1.1.1.2
2.7.7.7
6.2
5
6.3.2.2
4
5.2
2.7.7.49
5.99.1.3
3
3.1
4.2
3.1.1.5
3.2
5.3.1.8
3.12
3.1.3.16
3.1.3.48
4.2.99.18
3.1.4.17
3.1.11.6
4.2.1.1
3.4.16.4
3.5.1.28
3.6.1.37

- Protein function is a property of three dimensional structure and it is hard to make inferences from linear sequence
- Biocatalytic function definitions based on the EC are not always precise and computable
- Specific issues:
  - Distant homologues are hard to identify
  - Proteins in superfamilies have divergent functions
  - We need *sensitive and specific methods*

- Machine Learning (ML) can be used to induce rules that can characterize proteins according to functional classes
- Strategy:
  - Identify superfamilies as relevant data sets for training as they contain examples of divergent functions
  - Functionally relevant representations of proteins based on conserved modules
  - Induction algorithms to infer hypothesis about the correlation between the proteins and their biocatalytic functions

0 75 150 225 300 375 450 525 600 675 750 825
172
137
177
671
705
771
667
875
822
868
1141
685
333
1253
149
664
2552
1903
2913
1114
1765
DRTS_PLACH + 1.5.1.3.1
DRT1_ARATH + 1.5.1.3.1
DRT2_ARATH + 1.5.1.3.1
DRTS_LEIAM + 1.5.1.3.1
DRTS_LEIMA + 1.5.1.3.1
DRTS_PLAFK + 1.5.1.3.1
DRTS_DAUCA + 1.5.1.3.1
DRTS_TOXGO + 1.5.1.3.1
TYSY_MOUSE - 2.1.1.45.1
TYSY_RAT - 2.1.1.45.1
TYSY_HUMAN - 2.1.1.45.1
TYSY_PNECA - 2.1.1.45.1
TYSY_HSVSA - 2.1.1.45.1
TYSY_HSVAT - 2.1.1.45.1
TYSY_CANAL - 2.1.1.45.1
TYSY_YEAST - 2.1.1.45.1
TYSY_CRYNE - 2.1.1.45.1
DYR_HSVS7 + 1.5.1.3.1
DYR_HSVSC + 1.5.1.3.1
TYSY_VZVD - 2.1.1.45.1
DYR_MESAU + 1.5.1.3.1
DYR_MOUSE + 1.5.1.3.1
DYR_PIG + 1.5.1.3.1
DYR_HUMAN + 1.5.1.3.1
DYR_BOVIN + 1.5.1.3.1
DYRA_CITFR + 1.5.1.3.1
DYRA_ECOLI + 1.5.1.3.1
DYRA_ENTAE + 1.5.1.3.1
DYR_HSVSA + 1.5.1.3.1
DYRA_KLEAE + 1.5.1.3.1
DYR_CHICK + 1.5.1.3.1
TYSB_BACSU - 2.1.1.45.1
TYSY_ECOLI - 2.1.1.45.1
DYRA_STAAU + 1.5.1.3.1
DYRB_STAAU + 1.5.1.3.1
TYSY_SHIFL - 2.1.1.45.1
TYSY_LACCA - 2.1.1.45.1
TYSY_MYCGE - 2.1.1.45.1
TYSY_STAAU - 2.1.1.45.1
DYRB_ECOLI + 1.5.1.3.1
DYRC_ECOLI + 1.5.1.3.1
TYSY_BPT4 - 2.1.1.45.1
DYR7_ECOLI + 1.5.1.3.1
DYR_AEDAL + 1.5.1.3.1
DYR_DROME + 1.5.1.3.1
DYRA_HAEIN + 1.5.1.3.1
DYR3_SALTY + 1.5.1.3.1



0
400
600
800
GSHC_SOYBN
GSHC_PEA
GSHC_ARATH
GSHR_HUMAN
GSHR_ORYSA
GSHR_ARATH
GSHR_ECOLI
GSHR_PSEAE
GSHR_YEAST
GSHR_BURCE
TYTR_TRYCR
TYTR_CRIFA
GSHR_CAEEL
YMER_STAAU
NAPE_ENTFA
TERA_PSESP
YGBD_ECOLI
NASD_BACSU
TRXB_ECOLI
TRXB_MYCGE
NIRB_ECOLI
BAIH_EUBSP
NADO_THEBR
TODA_PSEPU
TRXB_PENCH
YHIN_ECOLI
GPDM_YEAST
GPDM_RAT
1.6.4.2
1.6.4.8
1.6.4.5
1.6.6.4
1.18.1.1
1.5.3.1
1.1.99.5
NE

- We have built classifiers for the inference of the biocatalytic potential of a putative protein
  - Efficiently annotate each ORF in a genome with putative enzymatic functions
  - Vary the sensitivity and specificity of function inference
  - Search for plausible protein candidates with a biocatalytic function

- We can infer metabolic pathways from genomic data, or just synthetic pathways, through heuristic search
- We can accurately assign enzymatic functions to putative proteins by machine learning.
- By combining function inference with pathway search, we can improve predictions further

- People
  - Daniel McShan (Ph.d. Student)
  - Shilpa Rao
  - Weiming Zhang
  - Minesh upadhyaya
- Funding
  - NSF
  - ONR
  - DoE