Computational Support for Metabolomics: Databases, Analysis, and Visualization

Pedro Mendes

http://www.vbi.vt.edu/~mendes

Scope

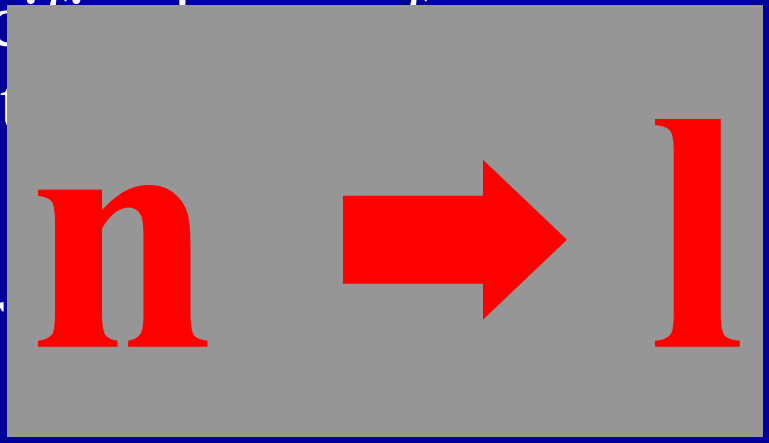
Have data, want knowledge

- Definitions
- Database issues
- Data analysis
- Biochemical networks for visualization

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Definitions

- Metabolic profiling
  - Usually targeted to spec
    metabolites, often quant
- Metabolomics
  - Comprehensive measur
    complement of specific
- Metabolic fingerprinting
  - Collection of metabolic patterns which are used for characterization of particular states



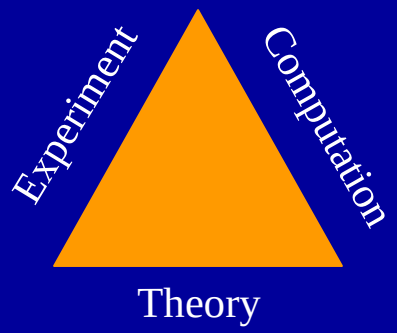
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Definitions

- Metabonic profiling
  - Usually targeted to specific classes of metabolites, often quantitative
- Metabonomics
  - Comprehensive measurement of the metabolite complement of specific species or cell types
- Metabonic fingerprinting
  - Collection of metabolic patterns which are used for characterization of particular states

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

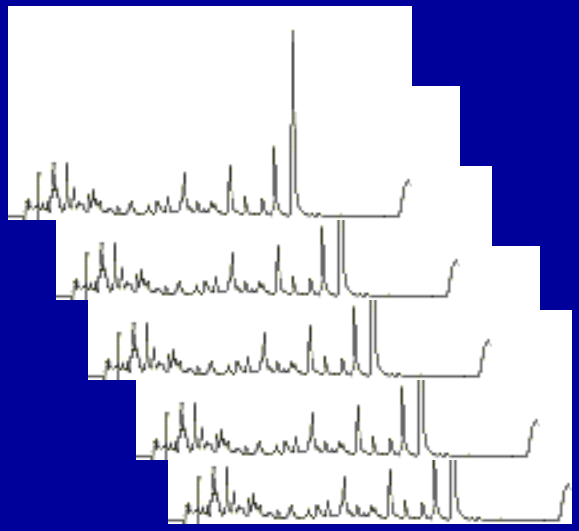
Systems Biology



- Quantitative whole-cell measurements (all mRNAs, all proteins, all sugars, etc.)
- Theoretical basis for explaining observations
- Synthesis of results through mathematical (computer) models

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

What this talk is not about…



ATP	1.277352
NADH	8.063249
NADPH	3.408609
ADP	1.554963
Orthophosphate	3.689995
CoA	8.182627
Pyrophosphate	9.292951
NH3	9.952874
S-Adenosyl-L-methionine	2.86405
AMP	9.696645
S-Adenosyl-L-homocysteine	4.708032
Pyruvate	8.779411
Acetyl-CoA	7.420493
L-Glutamate	7.226892
2-Oxoglutarate	1.709104
UDPglucose	9.610996
D-Glucose	0.222737
Acetate	8.071287
GDP	9.853982
Oxaloacetate	5.882063
Glycine	2.372851
L-Alanine	6.422998
Succinate	4.245424
UDP-N-acetyl-D-glucosamine	5.108658
GTP	6.890642
L-Lysine	6.186211
Glyoxylate	9.115576
L-Aspartate	9.188168
Glutathione	5.232367
UDP-D-galactose	4.903077
Formate	4.424477
L-Arginine	4.278822
L-Glutamine	6.067831
L-Serine	7.690047
Formaldehyde	1.427103
Thiamin diphosphate	6.424938
Alcohol	4.974474
Ascorbate	1.747568
L-Methionine	5.809962
Phosphoenolpyruvate	8.066765
L-Ornithine	9.29803
L-Tryptophan	4.057345
L-Phenylalanine	5.560732
L-Tyrosine	0.989049
Malonyl-CoA	5.070652
Acetaldehyde	3.270844
D-Fructose 6-phosphate	4.458931
Sucrose	8.617488

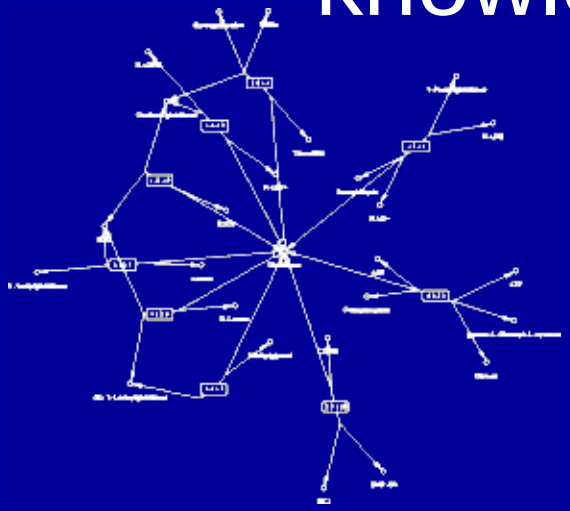
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Rather…

data

ATP	1.277352
NADH	8.063249
NADPH	3.408609
ADP	1.554963
Orthophosphate	3.689995
CoA	8.182627
Pyrophosphate	9.292951
NH3	9.952874
S-Adenosyl-L-methionine	2.86405
AMP	9.696645
S-Adenosyl-L-homocysteine	4.708032
Pyruvate	8.779411
Acetyl-CoA	7.420493
L-Glutamate	7.226892
2-Oxoglutarate	1.709104
UDPglucose	9.610996
D-Glucose	0.222737
Acetate	8.071287
GDP	9.853982
Oxaloacetate	5.882063
Glycine	2.372851
L-Alanine	6.422998
Succinate	4.245424
UDP-N-acetyl-D-glucosamine	5.108658
GTP	6.890642
L-Lysine	6.186211
Glyoxylate	9.115576
L-Aspartate	9.188168
Glutathione	5.232367
UDP-D-galactose	4.903077
Formate	4.424477
L-Arginine	4.278822
L-Glutamine	6.067831
L-Serine	7.690047
Formaldehyde	1.427103
Thiamin diphosphate	6.424938
Alcohol	4.974474
Ascorbate	1.747568
L-Methionine	5.809962
Phosphoenolpyruvate	8.066765
L-Ornithine	9.29803
L-Tryptophan	4.057345
L-Phenylalanine	5.560732
L-Tyrosine	0.989049
Malonyl-CoA	5.070652
Acetaldehyde	3.270844
D-Fructose 6-phosphate	4.458931
Sucrose	8.617488

knowledge



$$\dot{A} = \frac{V_1^f \frac{S}{K_{1S}} - V_1^r \frac{A}{K_{1A}}}{1 + \frac{S}{K_{1S}} + \frac{A}{K_{1A}}} - \frac{\left(V_2^f \frac{A}{K_{2A}}\right)\left(1 - \frac{B}{S \cdot K_{2eq}}\right)\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h-1}}{\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h} + \frac{1 + \left(\frac{C}{K_{2C}}\right)^{h}}{1 + \alpha\left(\frac{C}{K_{2C}}\right)^{h}}}$$

$$\dot{B} = \frac{\left(V_2^f \frac{A}{K_{2A}}\right)\left(1 - \frac{B}{S \cdot K_{2eq}}\right)\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h-1}}{\left(\frac{A}{K_{2A}} + \frac{B}{K_{2B}}\right)^{h} + \frac{1 + \left(\frac{C}{K_{2C}}\right)^{h}}{1 + \alpha\left(\frac{C}{K_{2C}}\right)^{h}}} - \frac{V_3^f \frac{B}{K_{3B}} - V_3^r \frac{C}{K_{3C}}}{1 + \frac{B}{K_{3B}} + \frac{C}{K_{3C}}}$$

$$\dot{C} = \frac{V_3^f \frac{B}{K_{3B}} - V_3^r \frac{C}{K_{3C}}}{1 + \frac{B}{K_{3B}} + \frac{C}{K_{3C}}} - \frac{V_4^f \frac{C}{K_{4C}} - V_4^r \frac{P}{K_{4P}}}{1 + \frac{C}{K_{4C}} + \frac{P}{K_{4P}}}$$

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Databases for Metabolomics

- Laboratory metabolic profile databases
- Species-specific metabolic profile databases
- Generic metabolic profile databases
- Qualitative metabolome databases
- Reference biochemical databases

Mendes (2002) “Emerging Bioinformatics for the Metabolome”. *Briefings in Bioinformatics* **3**, 134-145

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Lab metabolic profile DB

- Act as primary data sources
- Store data about all experimental details (i.e. metadata)
- Narrow in topic, deep in information content
- Should export data in standard formats to allow for interoperability with other DBs

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Species metabolic profile DB

- Collect all experiments published for one species
- Collect data for other types of experiments too (e.g. sequencing, microarrays)
- These are the primary point of entry for species-related information
- Examples:
  - TAIR
  - MaizeDB

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Generic metabolic profile DB

- Collects all published metabolic profiles
- Profiles must allow comparisons
- Due to size constraints, will not store much raw data, but will reference where it is (lab DB, literature)
- Few of these, preferably all mirroring the same data

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Qualitative metabolome DBs

- List all metabolites observed
- The data could be for a single organism or organized taxonomically
- This would be the equivalent to gene databases, as the list is of all potential metabolites that could be seen in an organism

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

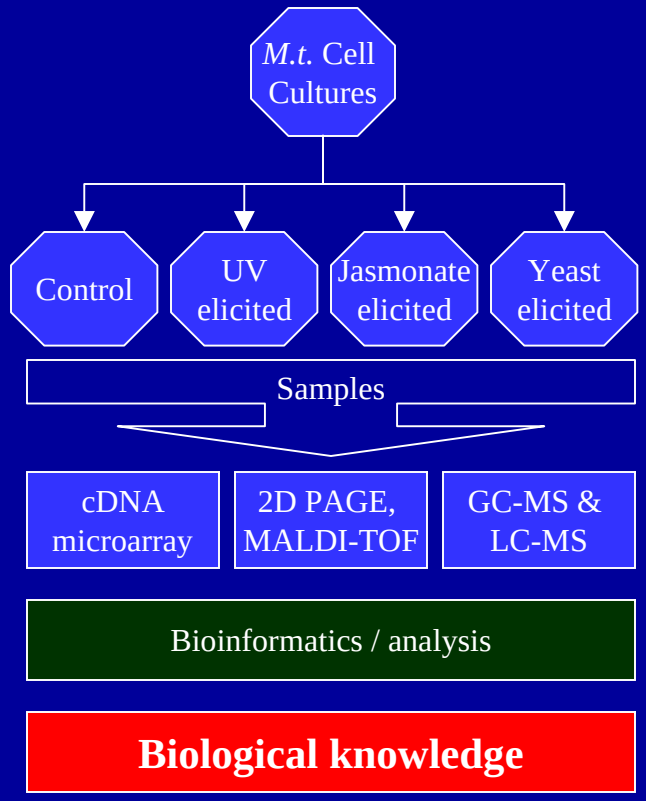
Reference biochemical DBs

- Contain reference information about biochemistry
- Many exist already: KEGG, WIT, EcoCyc, PathDB, PFMM, UM-BBD, BRENDA, SoyBase, etc.
- KEGG is perhaps the most popular, due to its nice pathway diagrams

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

An integrated approach to functional genomics and bioinformatics in a model legume

Mendes, Dixon, Sumner, May, Weller, Smith

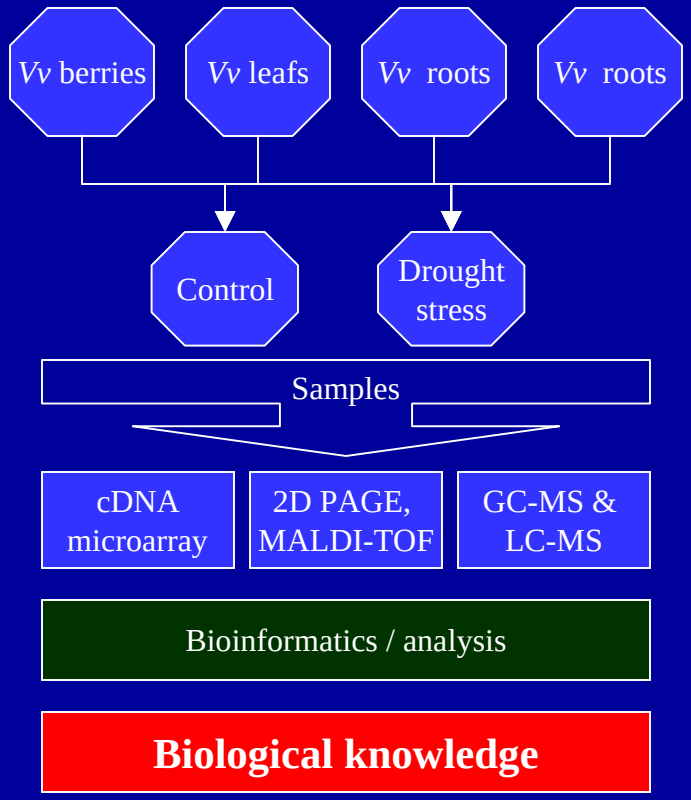


- Focus on the isoflavonoid pathway
- The data set contains information relevant to other processes too
- Quantification is important
- The ultimate aim is to derive causal relationships

http://medicago.vbi.vt.edu

Integrative Functional Genomic Resource Development in *Vitis vinifera*: Abiotic Stress and Wine Quality.

Cramer, Cushman, Mendes, Schooley



- Focus on flavor and stress compounds
- The data set contains information relevant to other processes too
- Quantification is important
- The ultimate aim is to derive causal relationships

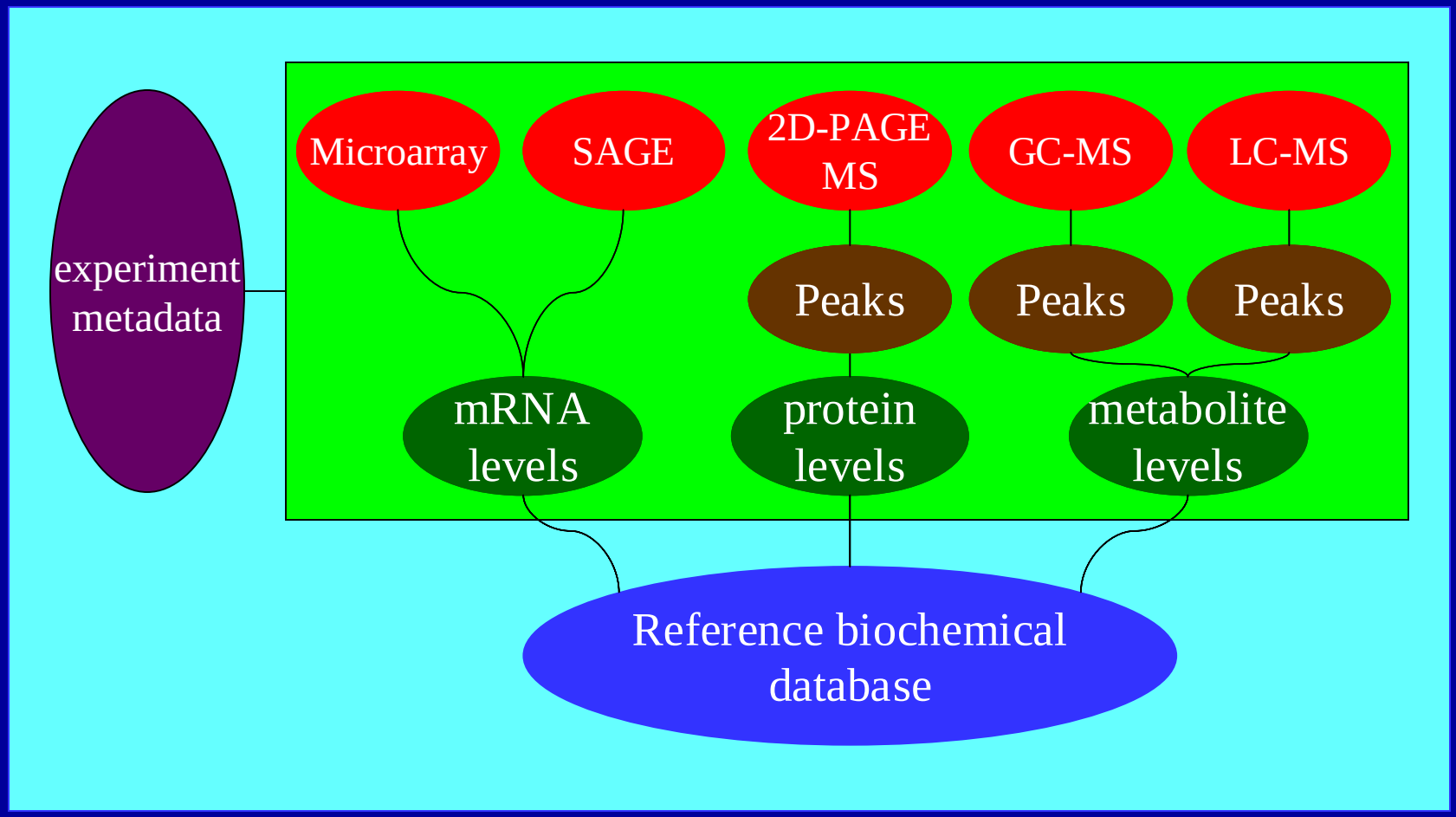
UNR

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

NSF

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

DOME, a Database for Functional Genomics



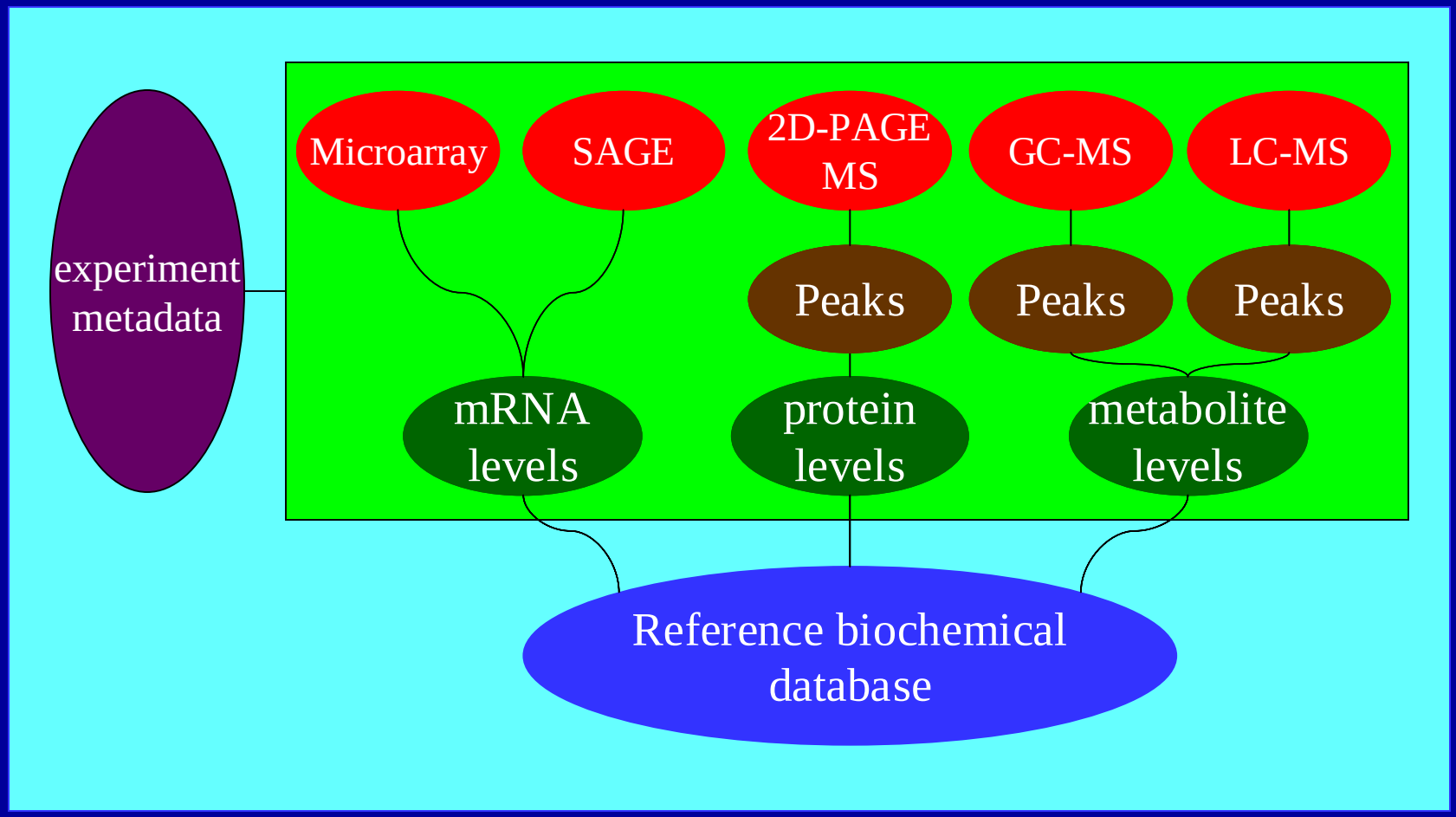
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Metadata

- To allow merging data of several types, it is essential that the database captures the data about the experiments.
- In DOME, experimental designs are represented in a finely structured classification.
- The database is general enough that it can store different kinds of experiments, such as time series or steady state experiments.
- The experimental metadata is what allows one to initially combine microarray, proteomic and metabolite profile samples.

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

DOME, a Database for Functional Genomics



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Desired properties of reference DBs

- All metabolites should be specific
  - no "an alcohol“, “acyl-CoA”, “amino acids”, etc.
- Enzymes should be single-entities
  - include all isoenzymes, etc.
- It is preferable to have fewer data if these are more reliable!
- Facts should be substantiated:
  - References
  - Classify evidence
- None of the existing reference DBs comply with all the requirements listed here

Wittig, U. & De Beuckelaer, A. (2001), *Brief. Bioinformatics* **2**, 126-142.

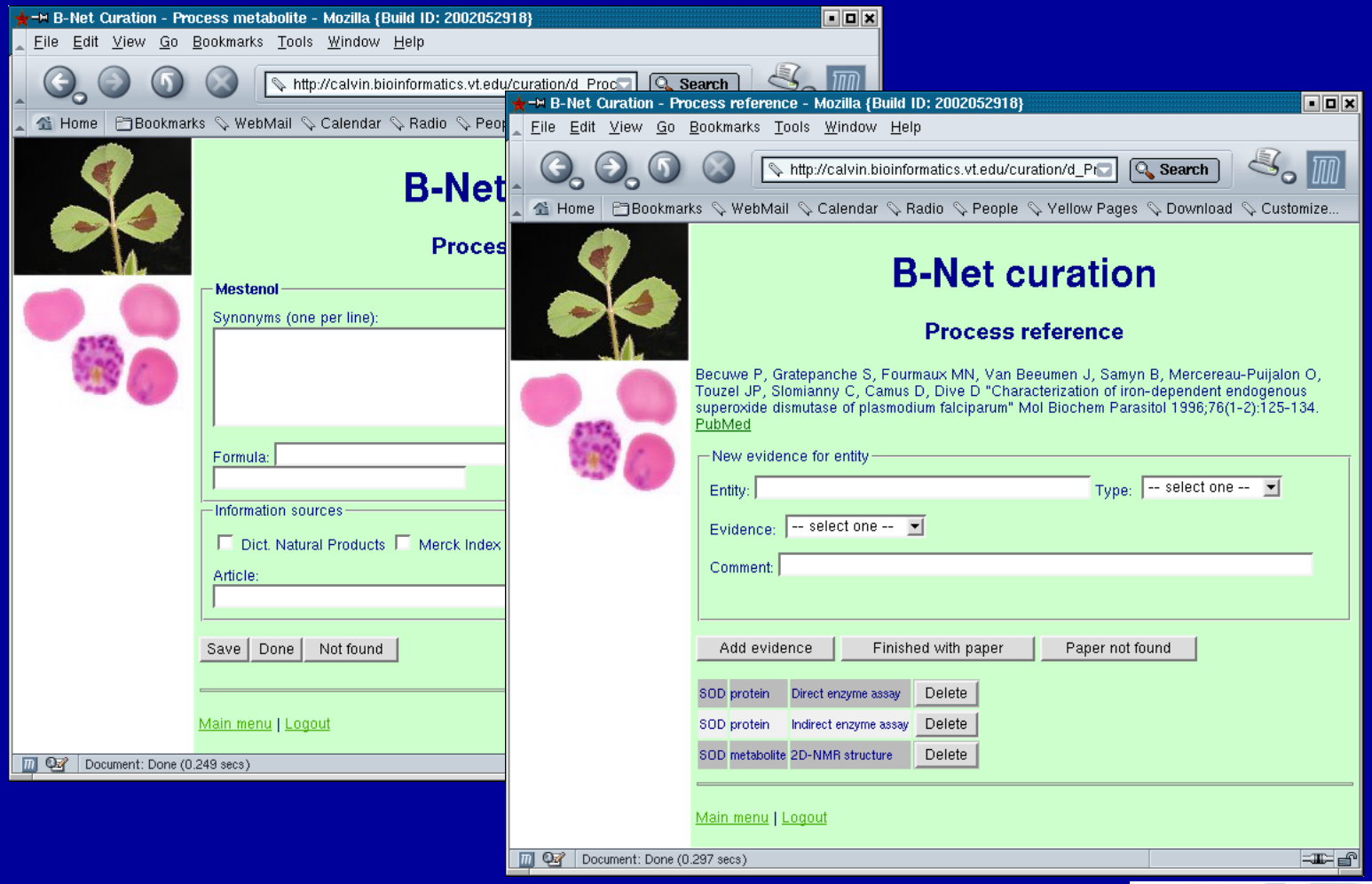
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

B-Net, a reference biochemical database

- To provide a reference for our metabolomic analyses and visualization
- Species specific
- Stores facts recovered from the literature
- Original metabolite data from KEGG, NIST and TAIR, reactions from EC, followed by draconian curation
- Also serves as a qualitative metabolome database

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Computer-assisted curation



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Algorithms for Data Analysis

- Unsupervised methods (looking for patterns)
  - Clustering, PCA
  - Self-organizing maps
- Supervised methods (calibration)
  - Nonlinear regression
  - Feed-forward neural networks
  - Genetic algorithms
- System identification (reverse engineering)
  - Bayesian belief networks
  - Metabolic control analysis
  - Nonlinear dynamics

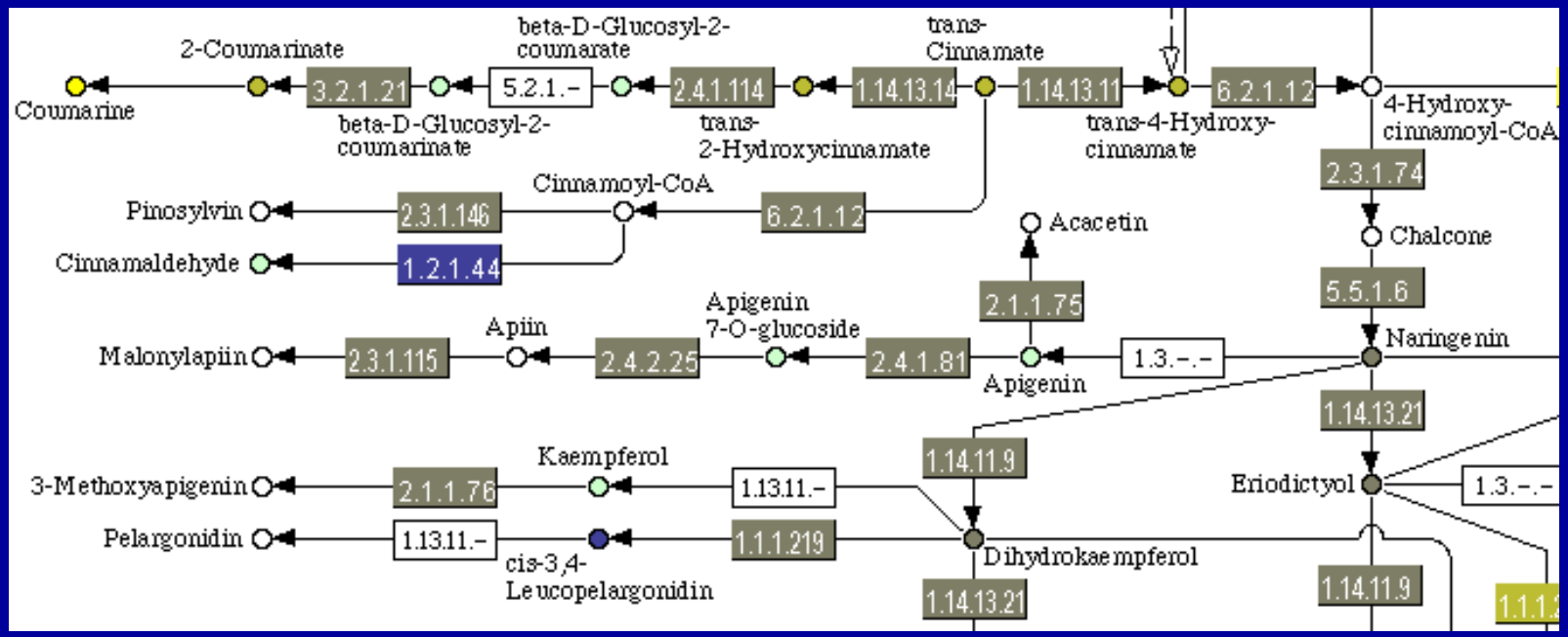
VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Metabolic networks for visualization and data integration

- Metabolic diagrams can be used to visualize metabolomic data together with gene expression or proteomics
- Useful to display data from one sample or to compare two samples
- KEGG has nice diagrams that can be used for this purpose

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

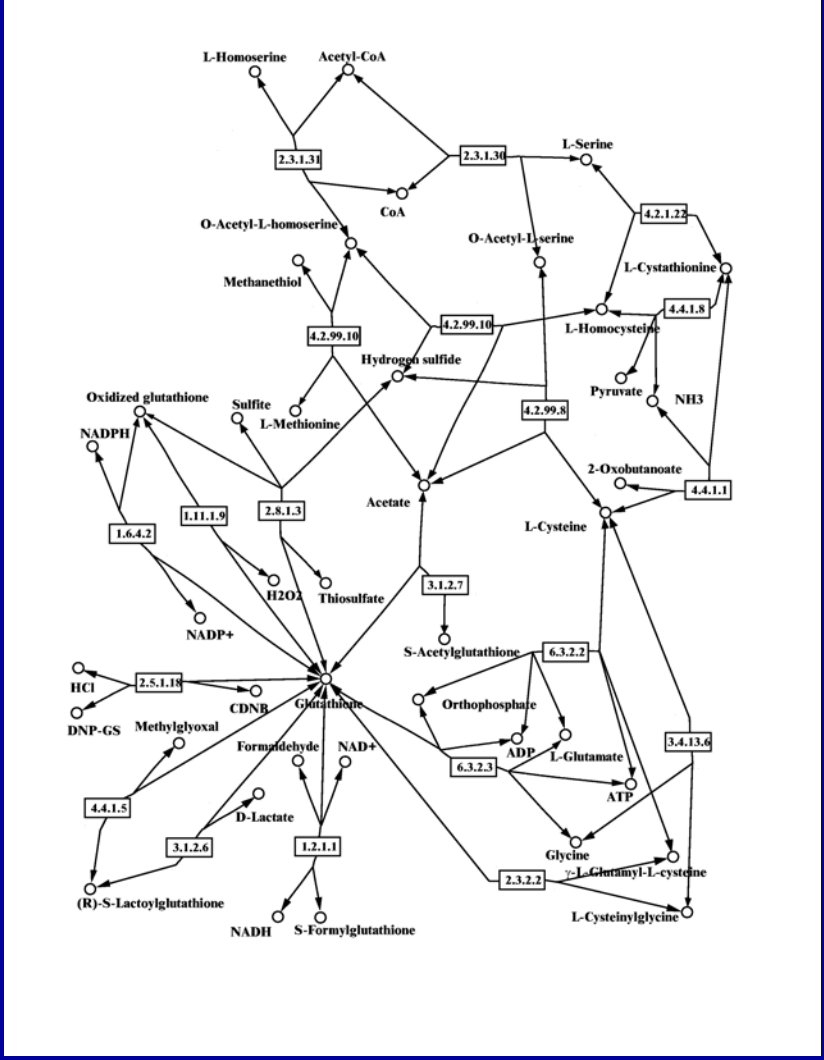
Data integration through metabolic networks



- KEGG maps do not include all side reactions
- Not all biochemistry can be reduced to a template

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Known glutathione neighborhood in *S. cerevisiae*



- All metabolites but H2O are included
- Each metabolite appears only once
- Reactions are included if:
  - Reaction observed directly
  - Gene for enzyme detected in genome

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

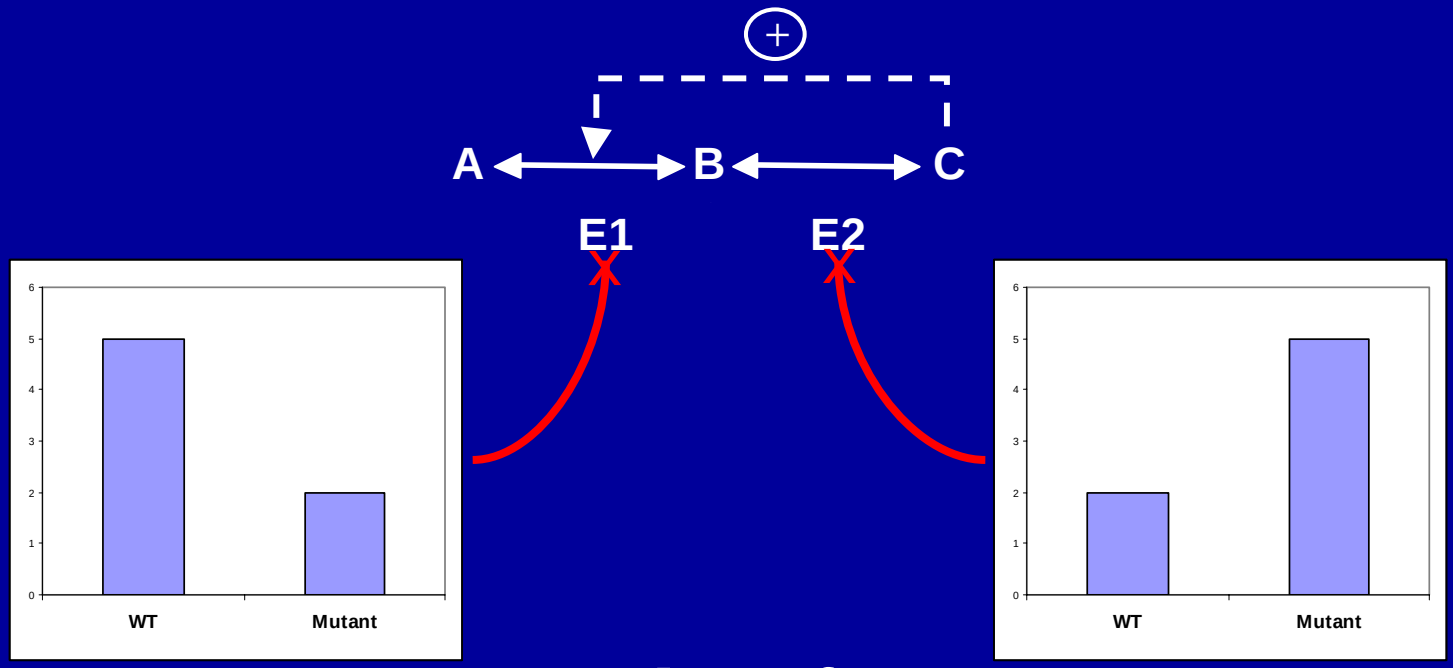
Small World Inside Metabolic Networks

Rank by degree	connectivity	Rank by distance	Importance no.
Glutamate	51	Glutamate	2.46
Pyruvate	29	Pyruvate	2.59
CoA	29	CoA	2.69
2-oxoglutarate	27	Glutamine	2.77
Glutamine	22	Acetyl CoA	2.86
Aspartate	20	Oxoisovalerate	2.88
Acetyl CoA	17	Aspartate	2.91
Phosphoribosyl PP	16	2-Oxoglutarate	2.99
Tetrahydrofolate	15	Phosphoribosyl PP	3.10
Succinate	14	Anthranilate	3.10
3-Phosphoglycerate	13	Chorismate	3.13
Serine	13	Valine	3.14
Oxoisovalerate	12	3-Phosphoglycerate	3.15

- Wagner & Fell (2001), *Proc. R. Soc. Lond. B* **268**, 1803-1810.

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Beware of simplistic ad-hoc interpretations…



But what if…?

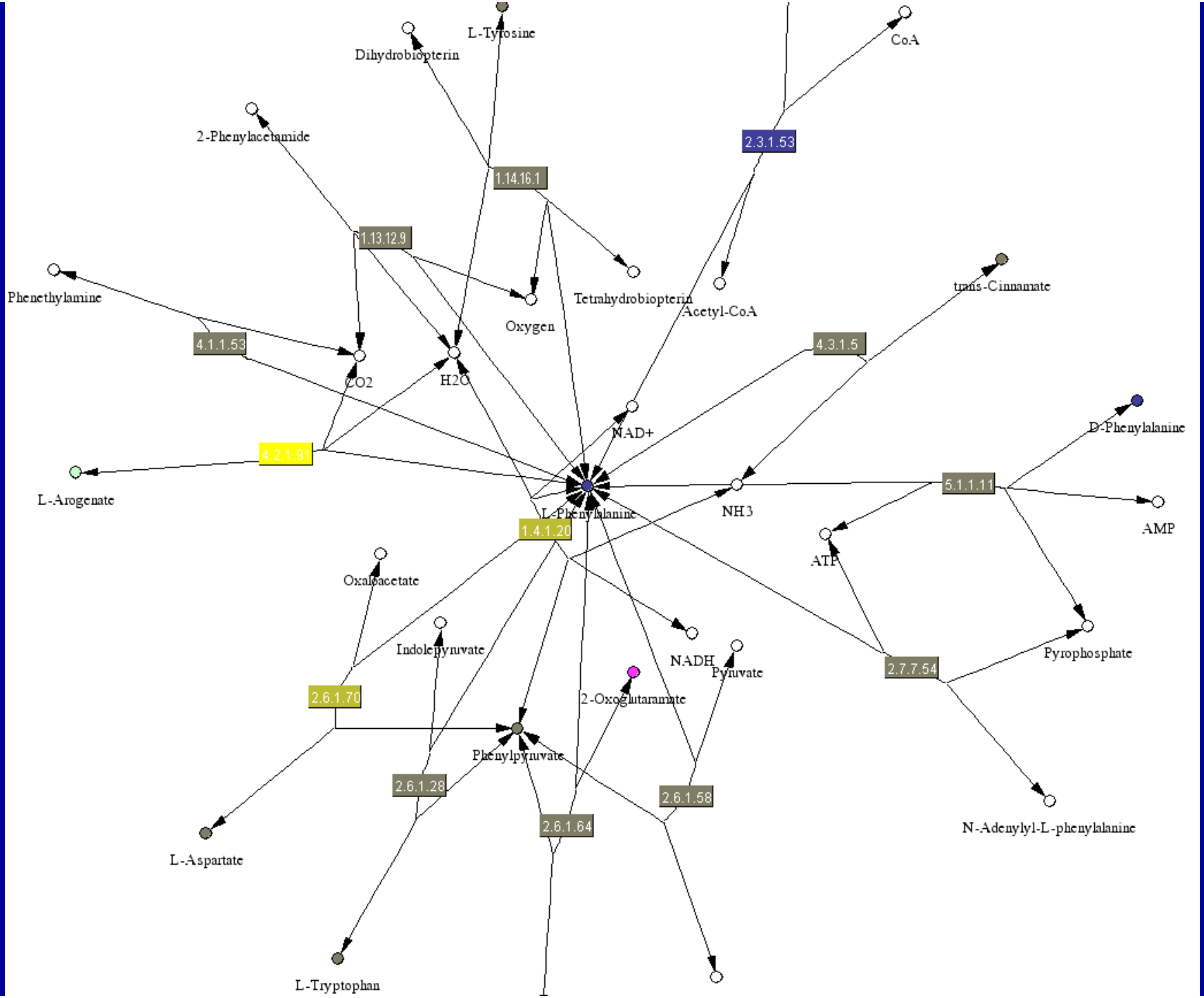
- Chance, B., Holmes, W., Higgins, J. & Connelly, C. M. (1958) Localization of interaction sites in multi-component transfer systems: theorems derived from analogues., *Nature. 162*, 1190-1193.
- Heinrich, R. & Rapoport, T. A. (1974) A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector., *Eur. J. Biochem. 42*, 97-105.

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Metabolite neighborhoods

- Metabolic pathways are an artificial concept derived from historical developments
- Biochemical network is highly interconnected
- Difficult to visualize the whole network
- Metabolite neighborhood is the set of reactions that are connected to a center metabolite, plus all other metabolites that are part of those reactions

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI


L-Tyrosine
Dihydrobiopterin
CoA
2-Phenylacetamide
2.3.1.53
1.14.16.1
1.13.12.9
Phenethylamine
Tetrahydrobiopterin
Acetyl-CoA
trans-Cinnamate
Oxygen
4.1.1.53
4.3.1.5
CO2
H2O
NAD+
D-Phenylalanine
4.2.1.91
L-Arogenate
L-Phenylalanine
NH3
5.1.1.11
AMP
1.4.1.20
ATP
Oxaloacetate
Indolepyruvate
NADH
Pyruvate
Pyrophosphate
2.7.7.54
2.6.1.70
2-Oxoglutaramate
Phenylpyruvate
2.6.1.28
2.6.1.58
2.6.1.64
L-Aspartate
N-Adenylyl-L-phenylalanine
L-Tryptophan

Concentration *versus* Flux

$$J_x(t) = \frac{dx}{dt} \approx \frac{x_t - x_{t-\tau}}{\tau}$$

- Flux is no more than the time derivative of concentration
- But flux is independent of concentration:

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI

Acknowledgements

- Stefan Hoops
- Paul Brazhnik
- Dingjun Chen
- Ana Martins
- Aejaaz Kamal
- X. Jing Li
- Alberto de la Feunte
- Diogo Camacho
- Wei Sha
- Mudita Singhal
- M. Kulkarni
- Liang Xu
- Jessica Caldwell
- Robin Oakes
- Fernina Taliaferro
- Anh Tran
- Olga Brazhnik
- Sinan Güler
- Rohan Luktuke

- VBI: Shulaev, Laubenbacher
- Noble Foundation: Dixon, Sumner, May
- U. Nevada-Reno: Cramer, Cushman, Schooley
- Johns-Hopkins: Sullivan

- National Science Foundation
- Commonwealth of Virginia

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH VBI