5th International Conference on
Biometrics & Biostatistics
Biostatistics 2016
October 20-21, 2016 Houston, Texas, USA

Pathways from the Genome to Risk Factors and Diseases via a Metabolomics Causal Network

Azam M. Yazdani, PhD

UTHealth®
The University of Texas
Health Science Center at Houston

Causal Inference in Observational Studies

Example

- Question

  What is the effect of Aspirin on Headache?

$$Aspirin \overset{???}{\rightarrow} Headache$$

Response to Treatment	Response to Control
$Y_1(t)$	$Y_2(c)$
$Y_4(t)$	$Y_3(c)$
.	.
.	.
.	.
$Y_i(t)$	$Y_j(c)$

- Observations

  Some under Aspirin treatment (t)

  and some as placebo, control (*C*)

- Association

  Comparison of treated and untreated individuals

  via statistical approaches

- Causation

  Considering the two elements, treatment and response is not sufficient.

  A third element is required in addition to treatment and response.

The third element for causal inference

- The third element represents relationships behind observations and is called
  - Assignment Mechanism (AM), (Rubin, 1991)
  - Data generating process, (Pearl, 2000)
  - Regime indicator (Dawid, 2007)

Since assignment mechanism (AM) is more appropriate in biomedicine, we use this phrase.

This third element, *AM*, must be known or estimated to infer causation.

Illumination of the *AM*

- Causal parameter is formalized as $AM\,(K_R)$:

  the illumination of the *AM* is carried out given Knowledge about Response.

- Randomization is a kind of *AM* that treatment is assigned regardless of any covariate.

- Given the causal parameter, we can distinguish confounders from the set of covariates to achieve causal inference.

Yazdani et al., (2015). Journal of Data Mining in Genomic and Proteomics

Conditional Quantity vs Causal Quantity

- Association Study

  Conditional probability

$$P(Y(t)|X) \text{ and } P(Y(c)|X)$$

or $$y = \alpha t + \beta x + e,$$

Conditioning on covariate X does not have a causal interpretation.

- Causal Quantities

$$P(Y(t)|X, AM(K_R)) \text{ and } P(Y(c)|X, AM(K_R))$$

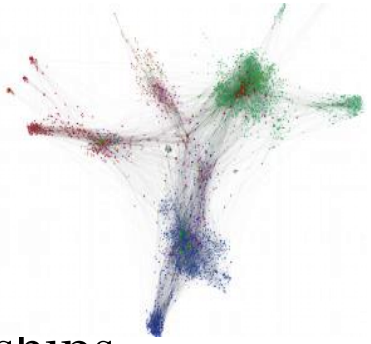
$$y|AM(K_R) = \alpha t + \beta x + e$$

Knowing the causal parameter $AM(K_R)$, covariate X is identified as a confounder.

Given X, as if the AM is randomized.

Yazdani et al., (2014). *International Journal of Research in Medical Sciences*

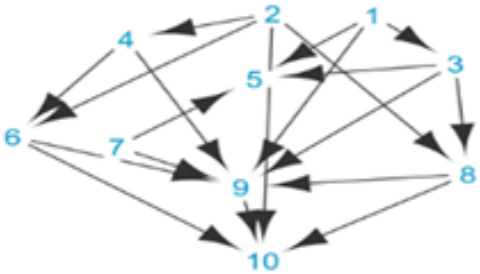
Bayesian Network/DAGs



- To have causal inference in observational studies we need to take into account the underlying relationships.

- In large scales data, Bayesian/causal networks are found the most pragmatic approach.

- A Bayesian network is an illumination of of the $AM(K_R)$.

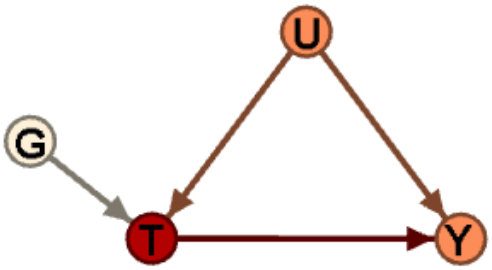


- Difficulty: Direction identification due to Markov equivalence property

$$M_1 \text{——} M_2$$

Data Integration

- Instrumental Variable Methods:
  1. IV is a cause of treatment
  2. IV is not associated with variation in unobserve confounder.
  3. IV is not a direct cause for response.



- **Advantage**: Control for unmeasured confounding
- **Disadvantage:** Increase variation
- Most of the genome IVs are weak and increase the variation

The GDAG algorithm

(Genome granularity Directed Acyclic Graph)

1. Extract near-complete information across the genome by principal component analysis
2. Find a topology over the genome information and variables of interest
3. Include the knowledge to the model
4. Use Bayesian rules to identify directionality among variables of interest.

Granularity means different types of data. (**Dawid**)

Recognized at the 2015 Atlantic Causal Inference Conference and won "The Thomas R. Ten Have Award".

Structural Equation Modeling (SEM)

- The matrix form of SEM under Gaussian assumption is

$$Z = (I - \Lambda)^{-1}(\Gamma X + U), \quad U_{1\times p}^{T} \sim N_p(0, \Delta)$$

where $\Delta_{p\times p} = diag\{\sigma_i^2\}$ and

$$Z^T = (Z_1, \ldots, Z_p) \sim N_p(0, \Sigma_Z)$$

where

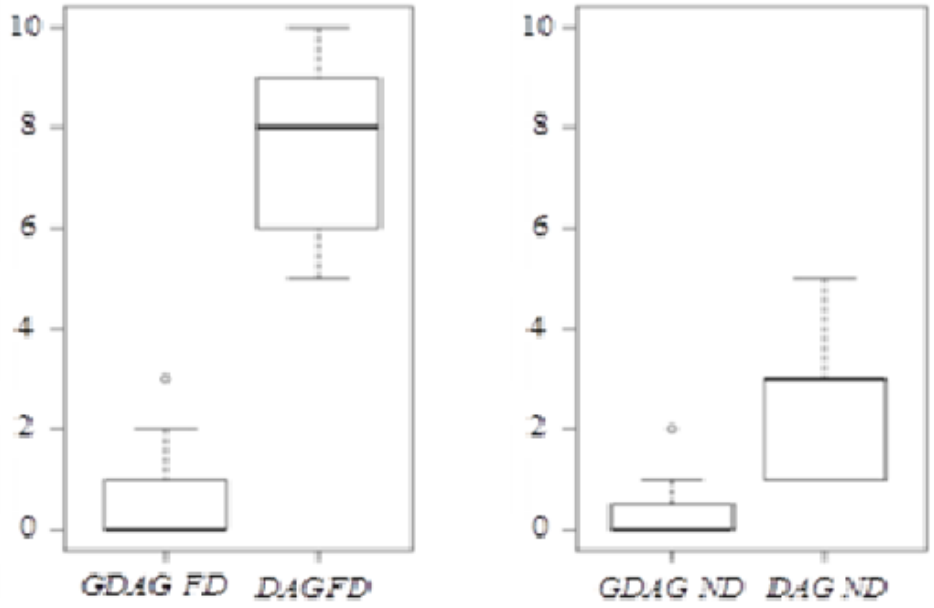
$$\Sigma_Z = (I - \Lambda)^{-1}\Delta((I - \Lambda)^{-1})^T.$$

This assumption is the p-dimensional random vector $(Z_1, \ldots, Z_p)$ is related to the set of nodes in DAG *D (e,v)*, where *e* stands for links and *v* for the nodes corresponding to the *p* variables.

Λ is a lower diagonal matrix with zeros on the diagonal.

The advantage of Using genome information

Comparison of the performance of the algorithm with and without data integration by comparing False Discoveries (FD) and Non Discoveries (ND).



Statistical significant level 0.001 selected by structural Hamming distance.

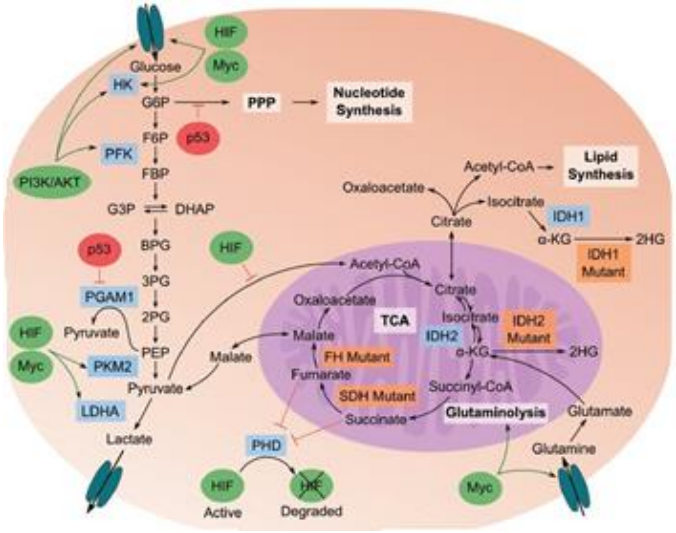
Yazdani et al., (2016). *Journal of Biomedical Informatics*.

Applications

Data

The metabolome is a collection of small molecules representing a variety of physiologic, metabolic, and cellular processes.

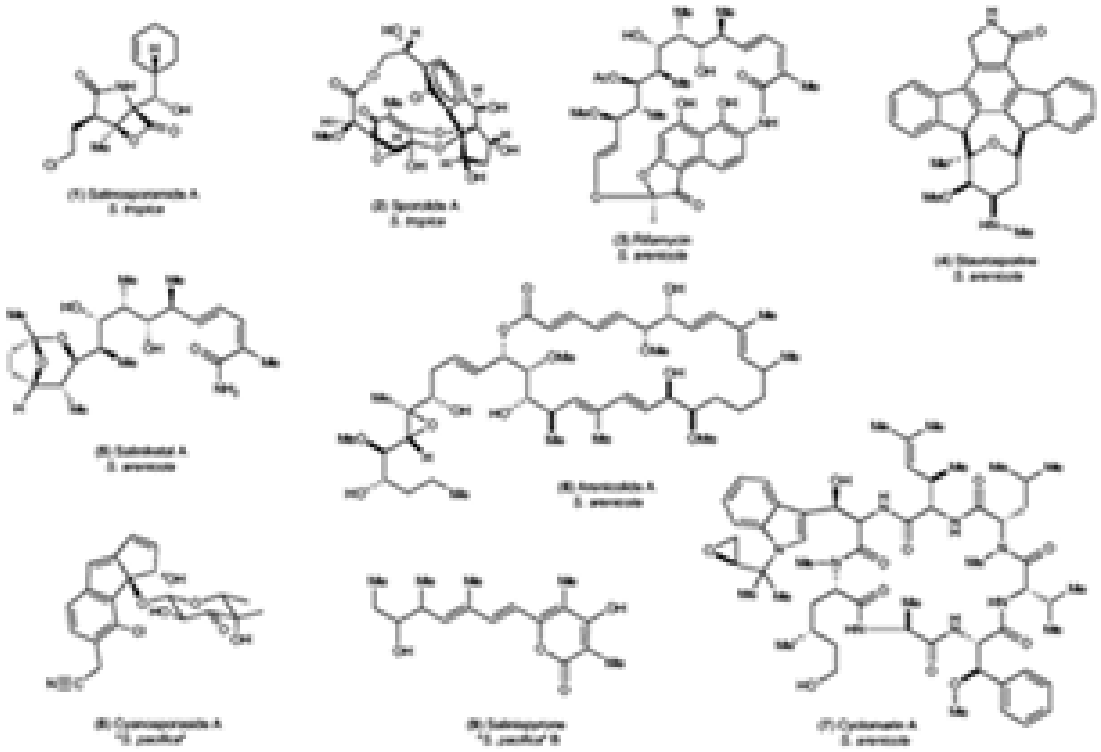




- *2,479* random samples, African-American from Jackson.
- 1,034,945 SNPs, across the genome.
- 122 metabolites transformed to normal, missing values imputed.

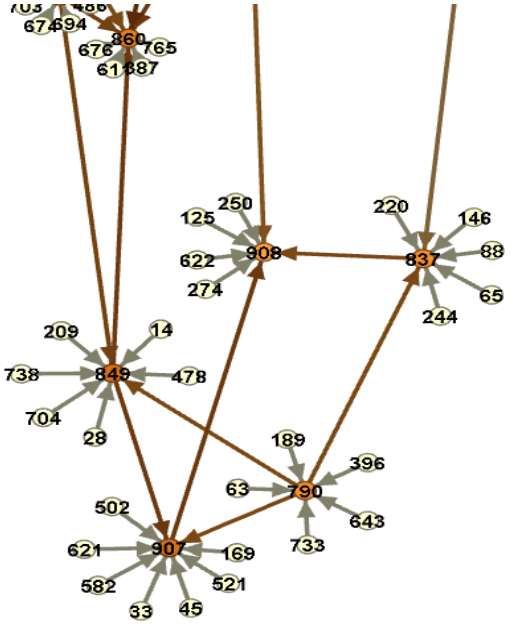
Metabolomics Causal Network

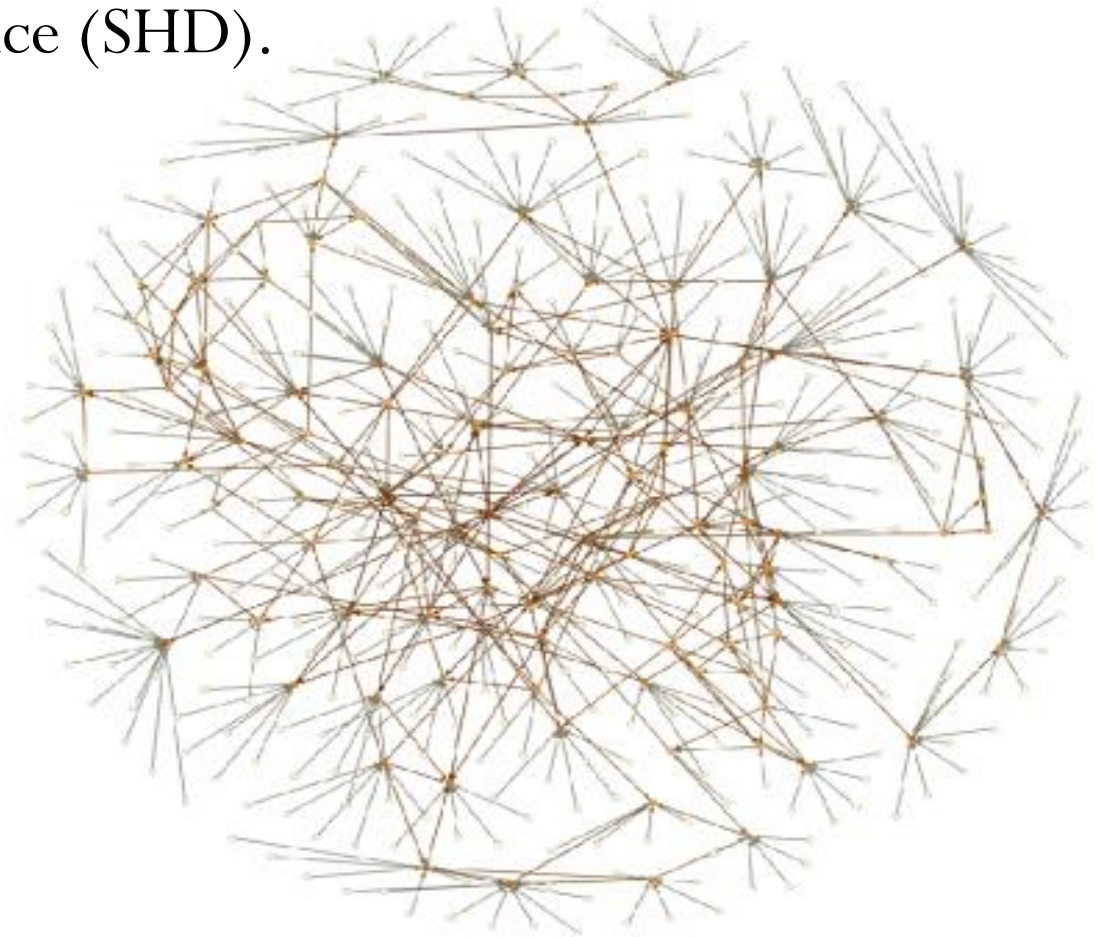
AM ?



Genome and Metabolomics Network

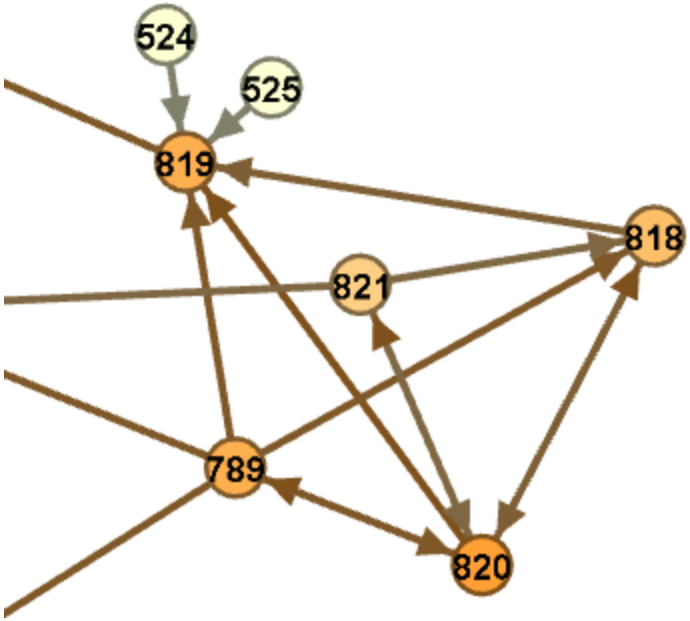
The result is based on statistical significant level 0.001 selected by structural Hamming distance (SHD).





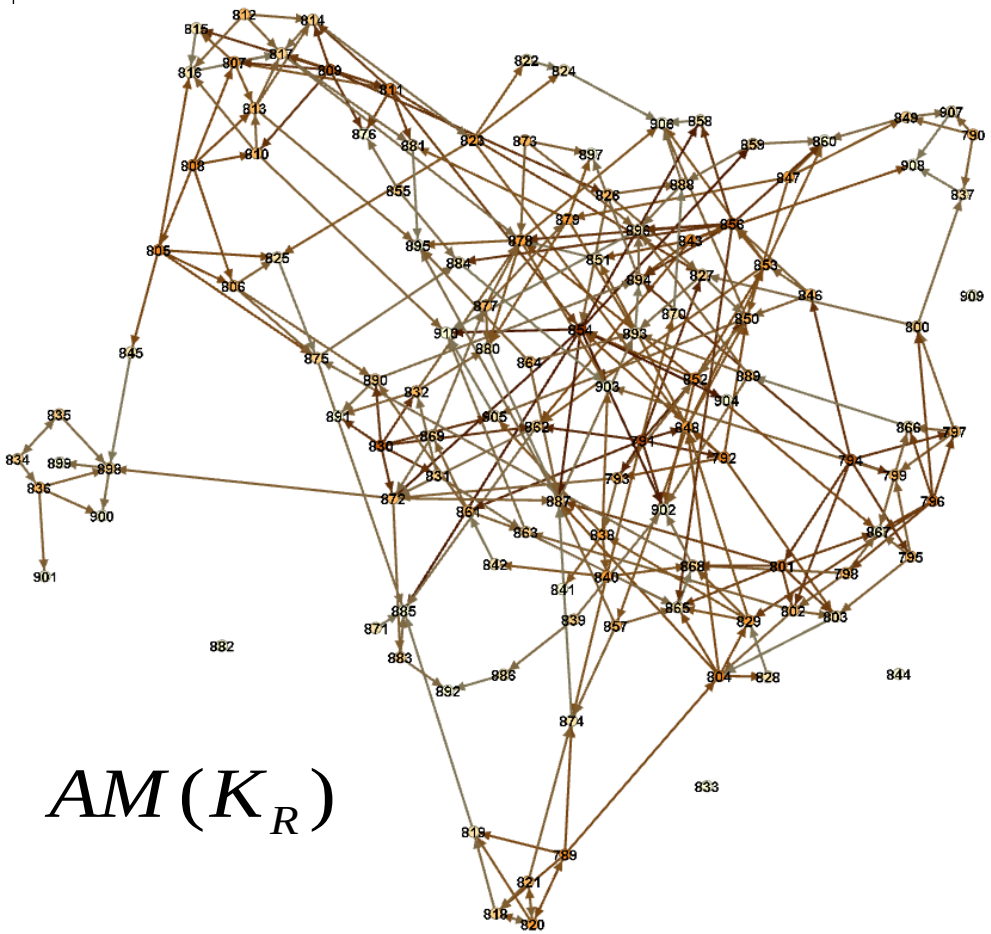
Dietary compounds

- Interesting no genome IV significantly influences dietary compounds.



Visualization of the Metabolomics Causal Network

- The result is based on statistical significant level 0.001 selected by structural Hamming distance (SHD).



$AM(K_R)$

$M_i \perp M_j \mid (S_{M_i}, AM(K_R))$, for $M_j \notin S_{M_i}$,

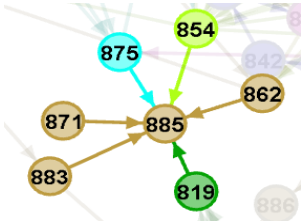
where M_i stands for a metabolite and

S_{M_i} is a set of M_ks with direct effect on M_i.

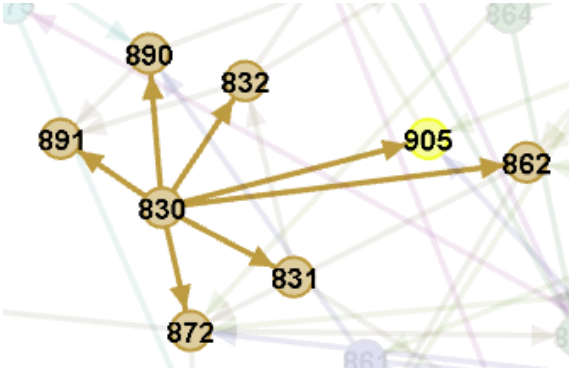
Yazdani et al., (2016). *Journal of Biomedical Informatics*.

Some simple definitions

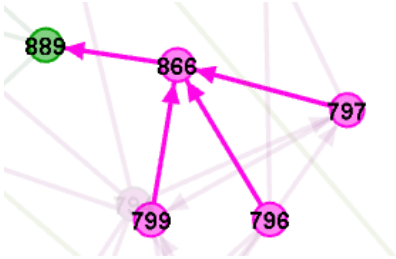
- No influence on other nodes



- No influence from other nodes



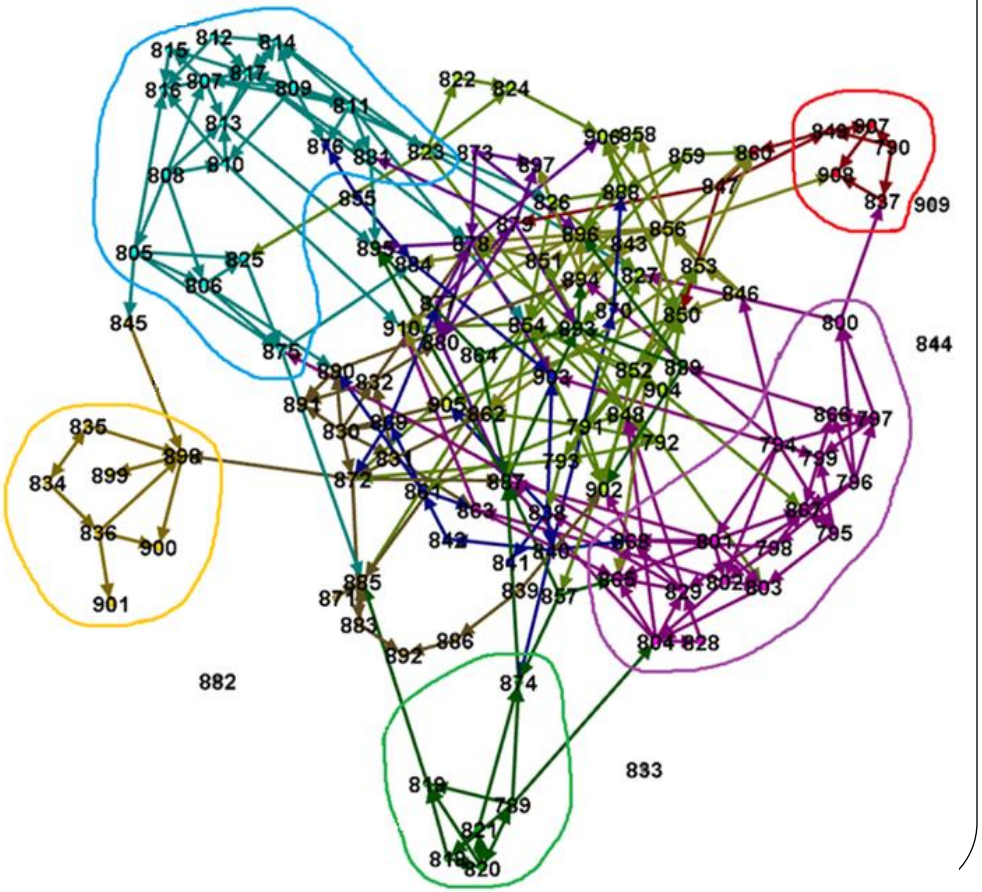
- The effect is not blocked in node 866.



Modules

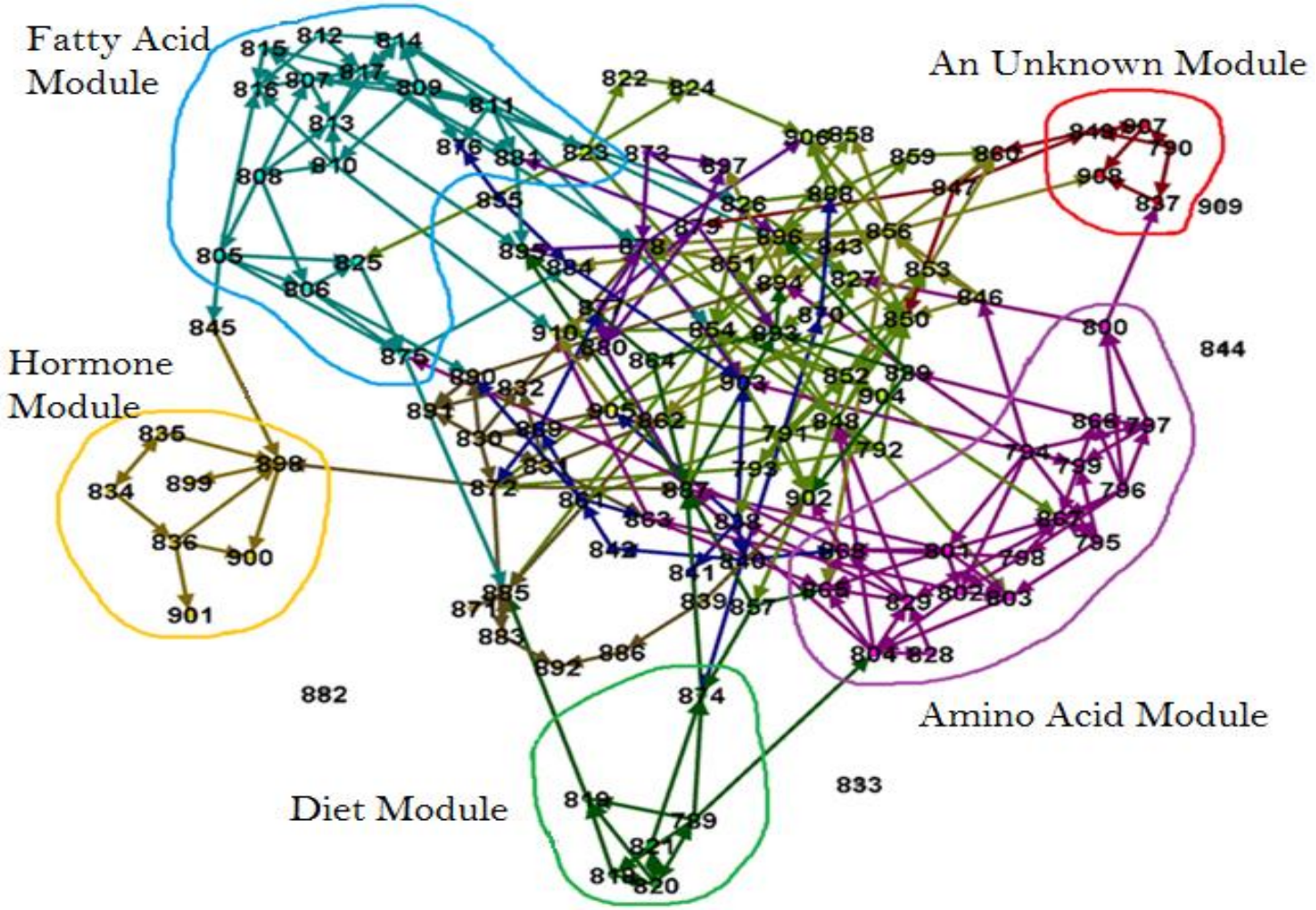
A module is a subset of the network where in the subset are densely connected nodes and sparsely connected with outside nodes.

- Causal networks have two features to determine borders
  1. Directionality
  2. Causal effect sizes



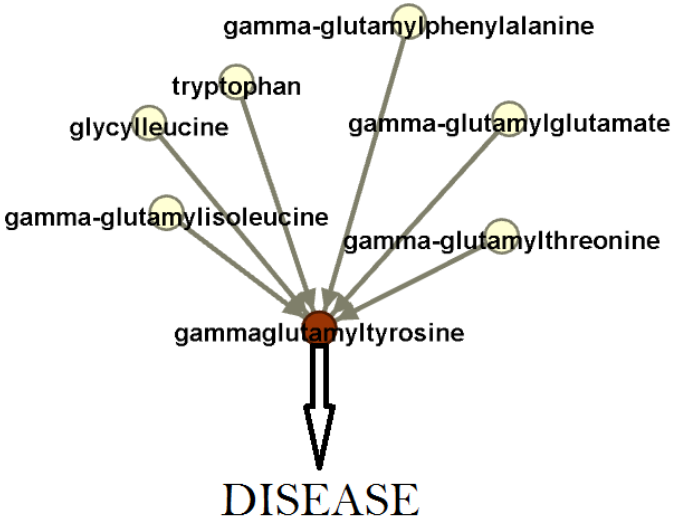
We noticed that the modules largely correspond to structural sub-classes of metabolites.

Modules

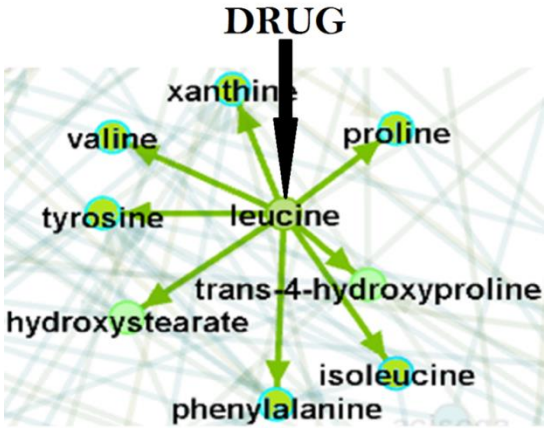

Fatty Acid Module
An Unknown Module
Hormone Module
Amino Acid Module
Diet Module

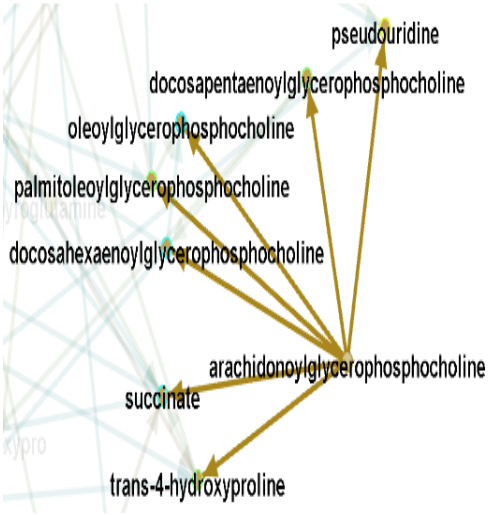
Hypothesized Targets

- For Prediction

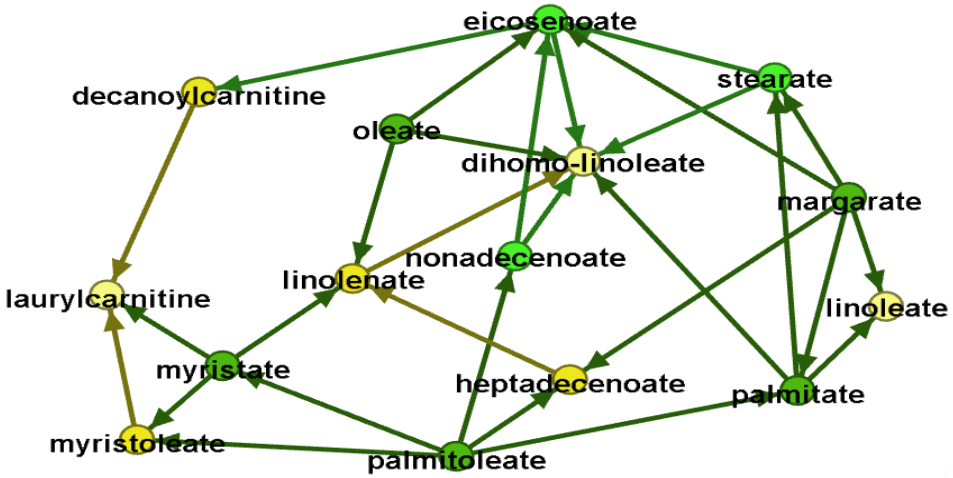


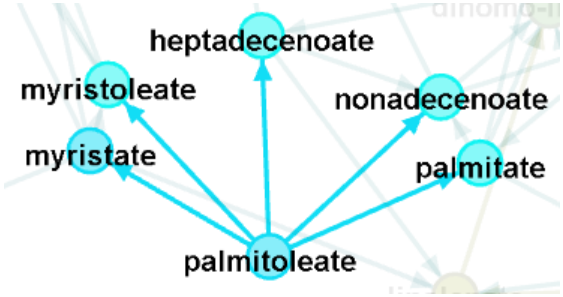
- For Intervention



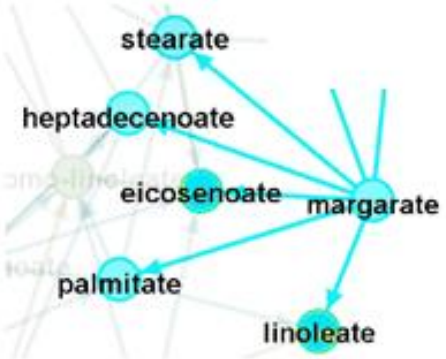


Fatty Acid Module





These two metabolites are not synthesized inside the body.



Dietary intervention on these two metabolites would be predicted to have an influence across the network.

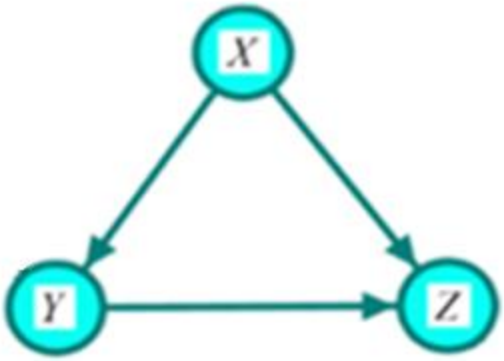
Yazdani et al. (2016). OMICS: A Journal of Integrative Biology. (in press)

Effect Measurements

The effect of *Y* on *Z* ?

The coefficient in this equation represents only the **association between *Y* and *Z***:

$$Z = \alpha' + \lambda Y + e'$$



To find the effect *Y* on *Z*:

1) The effect of *X* is removed from *Y* by $Y \mid AM(K_R) = \alpha_{yx} + \beta_{yx} X + e_{yx}$

2) And then fitting *Z* on residual *Y* on *X*: $Z \mid AM(K_R) = \alpha_z + \gamma e_{yx} + e_z$

The coefficient in this equation is interpreted as **the effect of *Y* on *Z***.

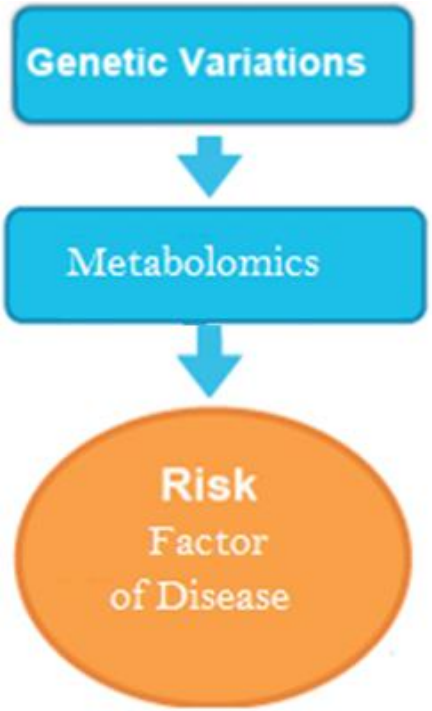
Yazdani et al., (2016). *Journal of Data Mining in Genomics & Proteomics*

Integrating three levels of granularity

Aim: Causal relationship of metabolites with triglyceride levels

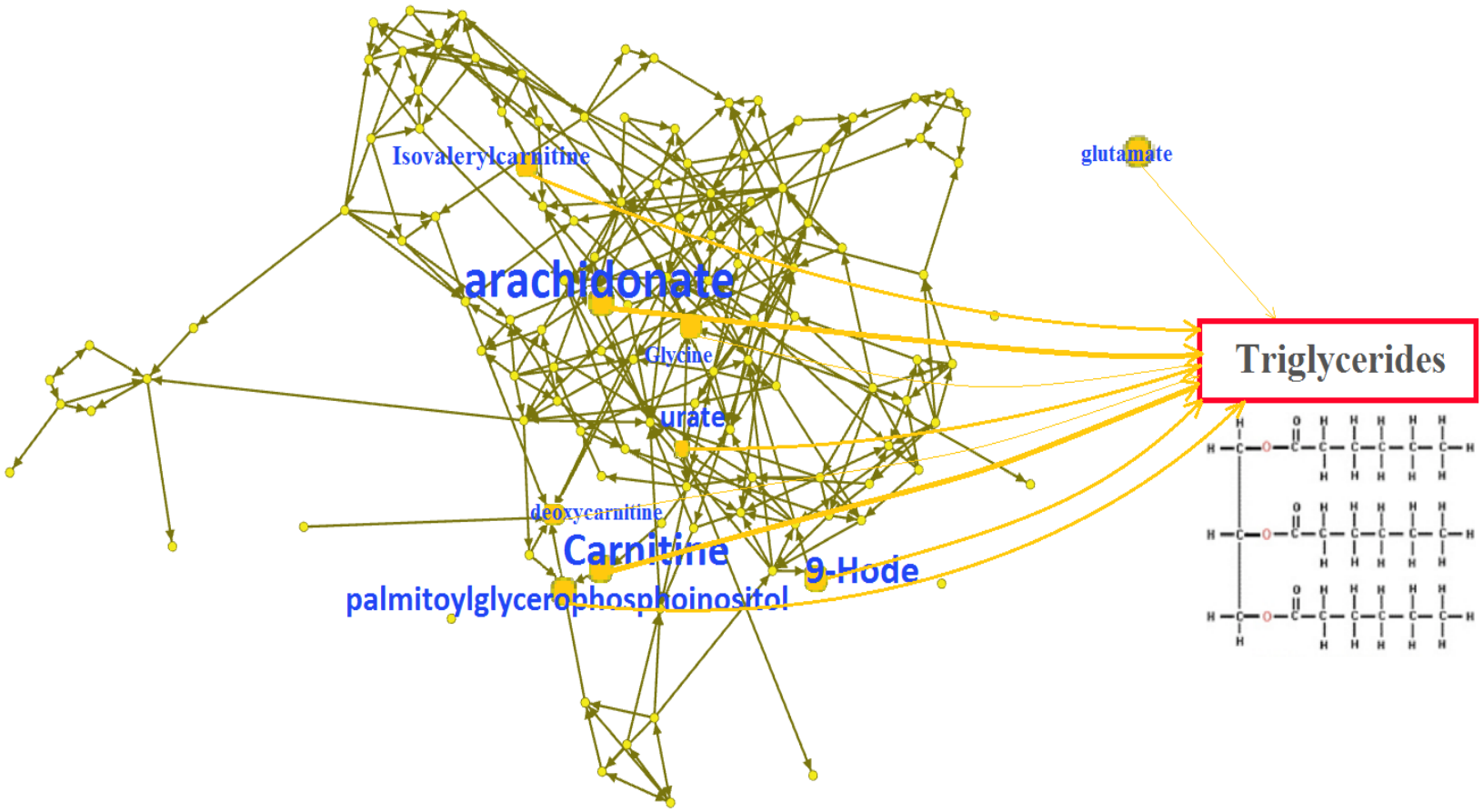
A single analysis:
28 metabolites out of 122
are associated with triglyceride levels.

How many of them have direct effect on the risk factors?



Yazdani et. al, (2016). Metabolomics.

Metabolomic and Triglycerides Network



We found only 9 metabolites have direct effect on triglycerides levels.

$$Trig \perp M_r \,|\, (S_{Trig}, AM(K_R)) \text{ , for } M_r \notin S_{Trig},$$

where *Trig* stands for triglyceride variable and represents a set of metabolites with direct effects on *Trig*.

Pathways from the genome to risk factors via metabolomic

Genome variants selection including common and rare

Methods: SKAT (Wu, et al., 2011) and **CCRS** (Yazdani, et al., 2015)

Yazdani *et al. BMC Bioinformatics* (2015) 16:405
DOI 10.1186/s12859-015-0825-4

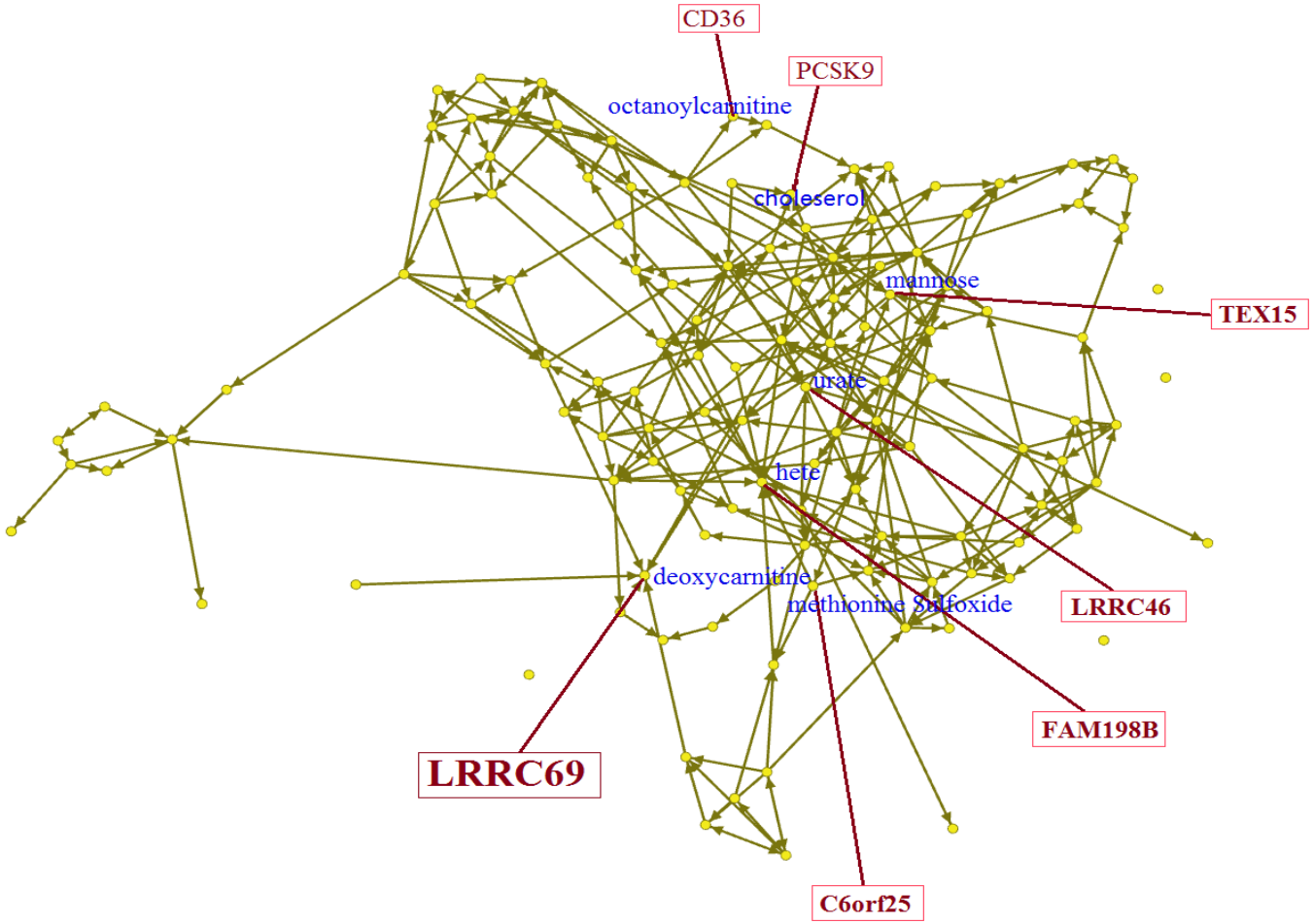
BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Rare variants analysis using penalization methods for whole genome sequence data

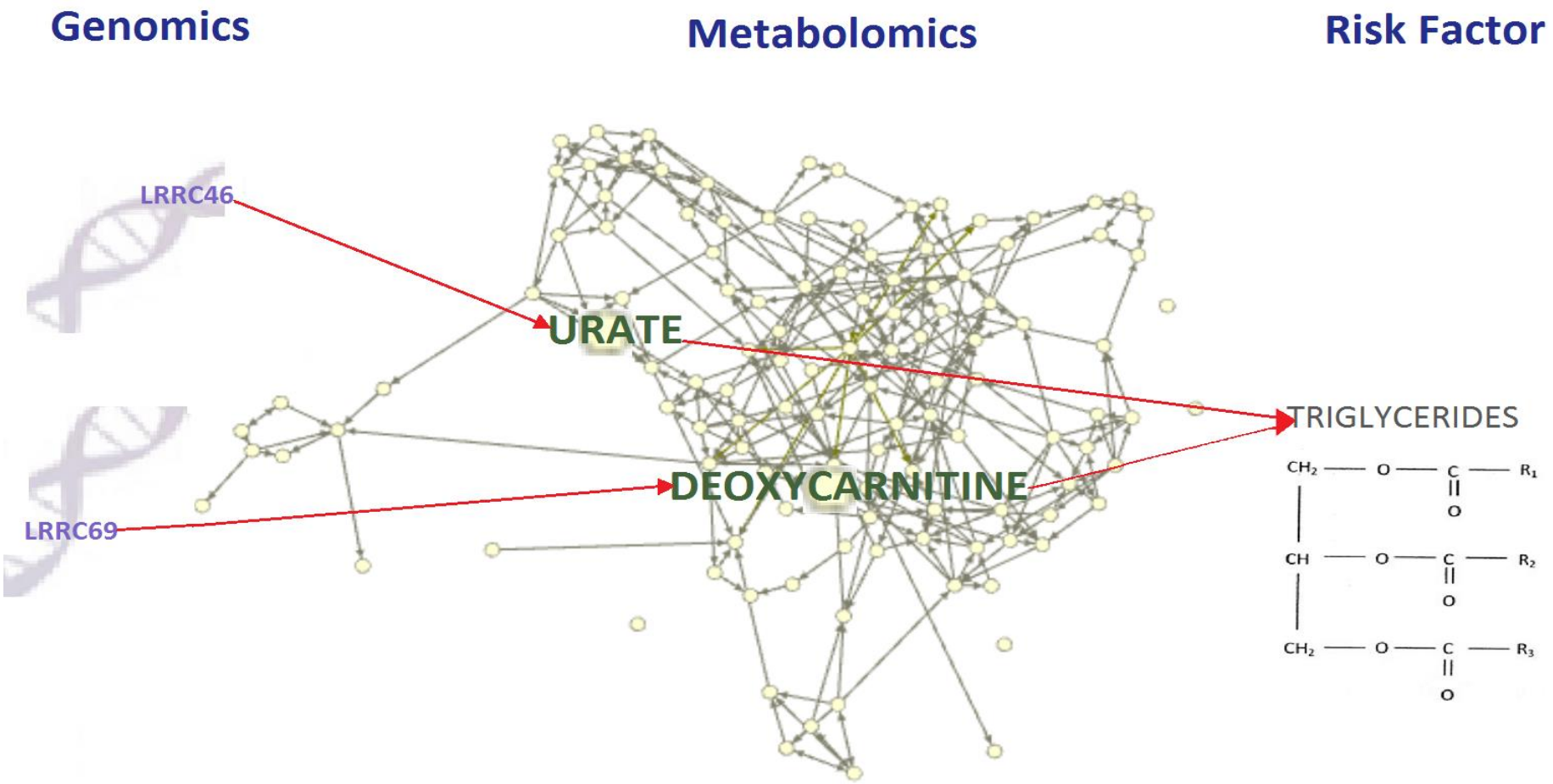
Genes with loss-of-function (LOF) variants

- Seven LOF variants in a strong relationship with metabolites in our study are identified



Pathways from the genome to risk factors via metabolites

These pathways are more stable than connecting a gene directly to a risk factor since we have considered amid metabolomics granularity.



Conclusion

To reduce false discovery, we analyze data in a causal setting and identify network underlying observations.

The GDAG algorithm identifies underlying networks robustly using the extracted information across the genome and creating strong and valid instrumental variables.
Then, we integrated different biological granularities.

Thanks

- Prof. Peter Kischka
  Friedrich-Schiller-University, Jena, Germany

- Prof. Philip Dawid
  Cambridge University

- Prof. Eric Boerwinkle

Others



Akram Yazdani, PhD,



Ahmad Samiei, PhD,

CANCER PREVENTION & RESEARCH INSTITUTE OF TEXAS

Two Fellowship Grants

CANCER PREVENTION & RESEARCH INSTITUTE OF TEXAS

Thank you