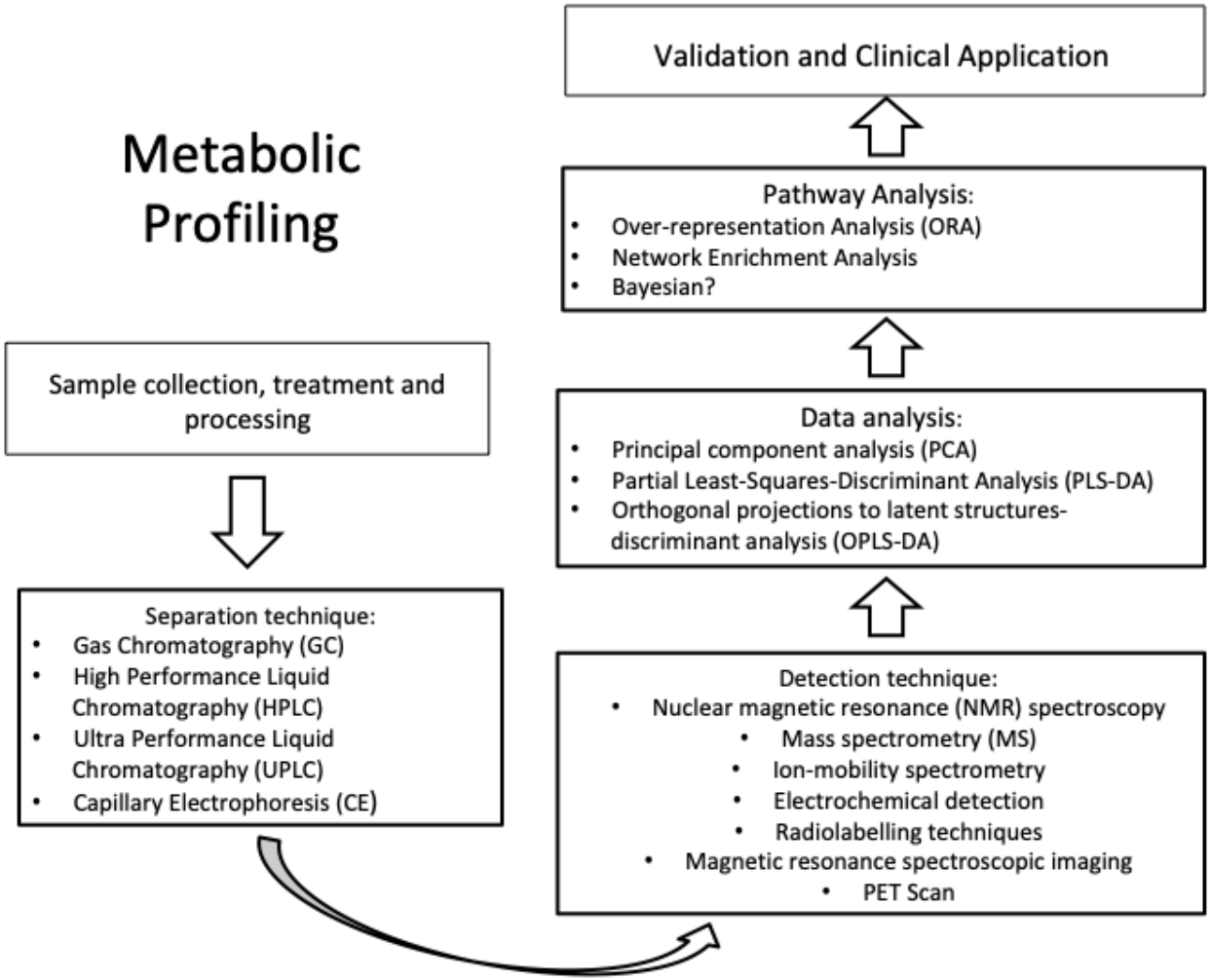
# Pathway Analysis in Metabolomics using Bayesian Statistics

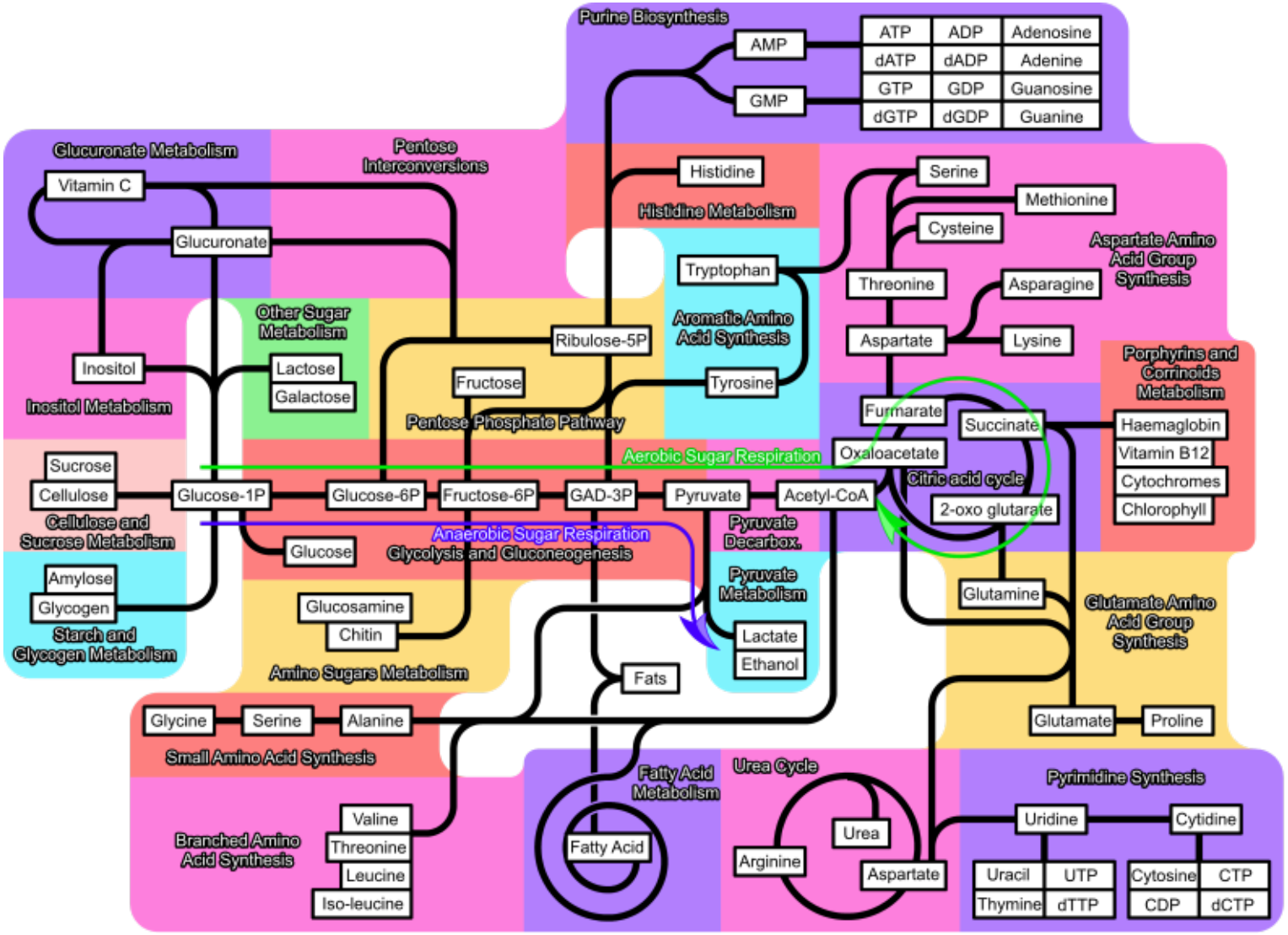
Novia Minaee

15/04/2021

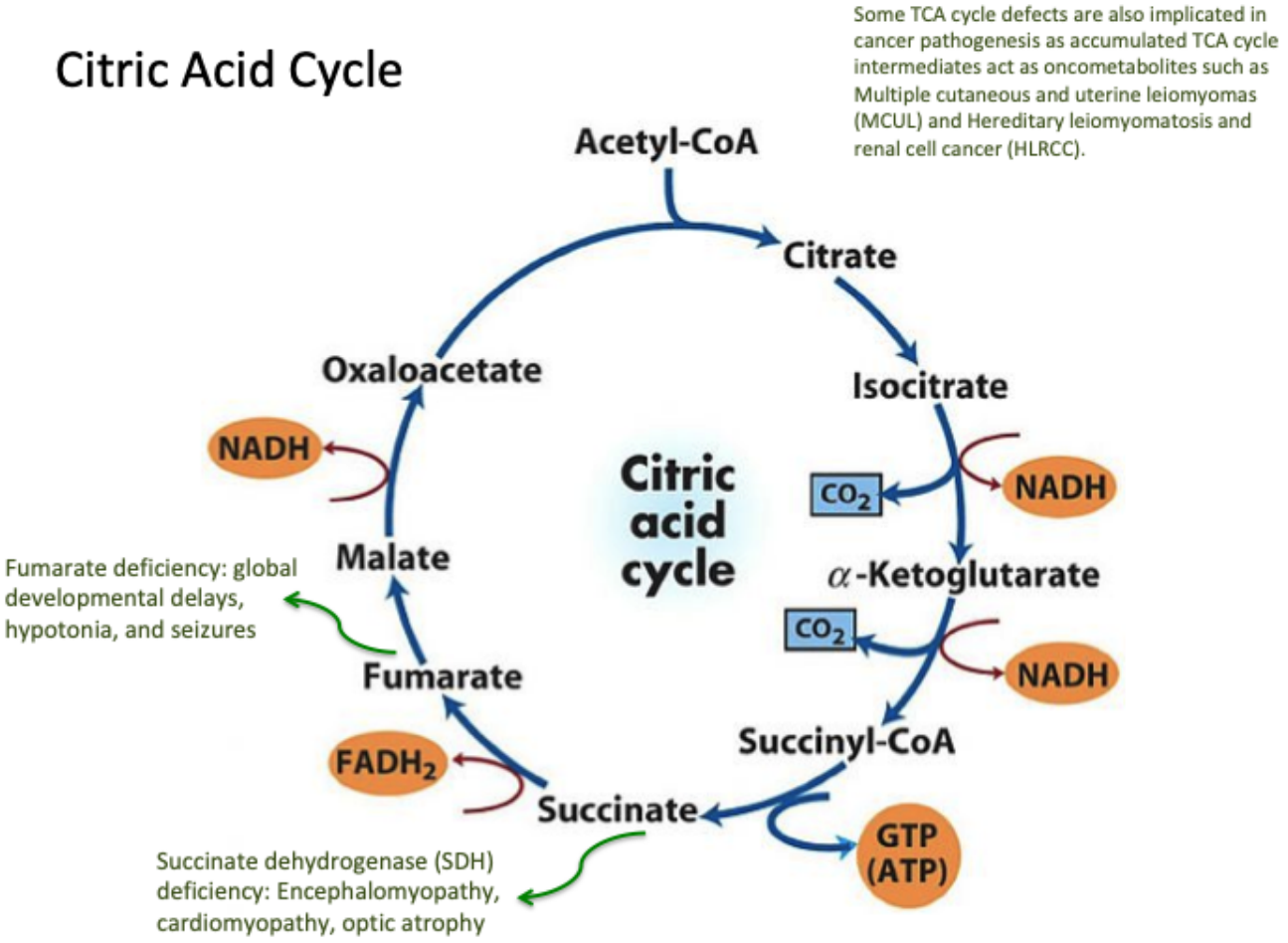
# Metabolic Profiling



# Metabolic journey



# Citric Acid Cycle or Tricarboxylic acid cycle (TCA)



# Challenge

- So many unknown links between pathways that haven't been identified.
- To take into account the interaction between pathways and metabolites within the pathway as one dynamic system.
- To include the pathway interaction into metabolic profiling in the pathway analysis.

# Aim

- To integrate the dynamic nature of a biological system in the analysis through Bayesian Statistics.

# Pathway Analysis Methods

Two most common methods of enrichment analysis are:

1. Over-representation analysis (ORA)
2. Quantitative Enrichment Analysis (QEA)

# Over-representation analysis (ORA)

- Tests if the proportion of affected metabolites within a pathway is statistically meaningful.
- Based on statistical tests on probability distribution like the hypergeometric, binomial or chi-squared.

(Kamburov et al. 2011; Ren et al. 2015; Khatri, Sirota, and Butte 2012; Picart-Armada et al. 2018)

# Quantitative Enrichment Analysis (QEA)

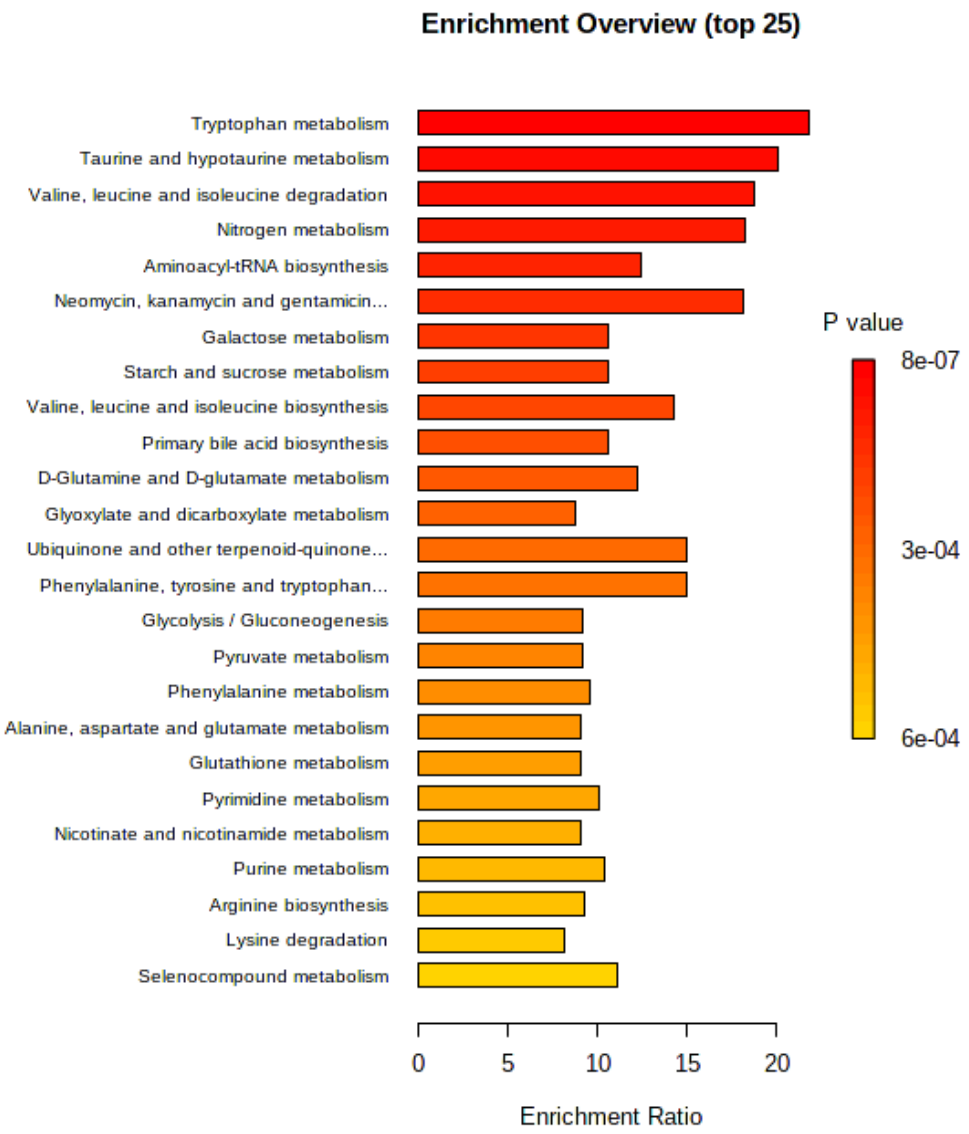
- It is directly based on the compound concentration values as compared to the compound lists used by over representation analysis.
- It is usually more sensitive than over-representation analysis and has the potential to discover "subtle but consistent" changes among compounds within the same biological pathway.

(Xia and Wishart 2010)

# Example of QEA Summary for top 25 Pathway

| | Total Cmpd | Hits | Statistic Q | Expected Q | Raw p | Holm p | FDR |
|---|---|---|---|---|---|---|---|
| Tryptophan metabolism | 41 | 1 | 22.64 | 1.04 | 8.41E-07 | 3.36E-05 | 3.36E-05 |
| Taurine and hypotaurine metabolism | 8 | 1 | 20.89 | 1.04 | 2.53E-06 | 9.87E-05 | 5.06E-05 |
| Valine, leucine and isoleucine degradation | 40 | 1 | 19.48 | 1.04 | 6.03E-06 | 2.29E-04 | 5.76E-05 |
| Nitrogen metabolism | 6 | 1 | 18.94 | 1.04 | 8.41E-06 | 3.11E-04 | 5.76E-05 |
| Aminoacyl-tRNA biosynthesis | 48 | 11 | 12.94 | 1.04 | 8.43E-06 | 3.11E-04 | 5.76E-05 |
| Neomycin, kanamycin and gentamicin biosynthesis | 2 | 1 | 18.89 | 1.04 | 8.64E-06 | 3.11E-04 | 5.76E-05 |
| Galactose metabolism | 27 | 2 | 11.05 | 1.04 | 1.22E-05 | 4.15E-04 | 6.11E-05 |
| Starch and sucrose metabolism | 18 | 2 | 11.05 | 1.04 | 1.22E-05 | 4.15E-04 | 6.11E-05 |
| Valine, leucine and isoleucine biosynthesis | 8 | 2 | 14.89 | 1.04 | 1.63E-05 | 5.21E-04 | 7.24E-05 |
| Primary bile acid biosynthesis | 46 | 2 | 11.08 | 1.04 | 1.99E-05 | 6.18E-04 | 7.97E-05 |
| D-Glutamine and D-glutamate metabolism | 6 | 3 | 12.73 | 1.04 | 4.12E-05 | 1.24E-03 | 1.50E-04 |
| Glyoxylate and dicarboxylate metabolism | 32 | 8 | 9.19 | 1.04 | 5.02E-05 | 1.45E-03 | 1.67E-04 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 9 | 1 | 15.60 | 1.04 | 6.23E-05 | 1.74E-03 | 1.78E-04 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 4 | 1 | 15.60 | 1.04 | 6.23E-05 | 1.74E-03 | 1.78E-04 |
| Glycolysis / Gluconeogenesis | 26 | 2 | 9.56 | 1.04 | 7.29E-05 | 1.90E-03 | 1.82E-04 |
| Pyruvate metabolism | 22 | 2 | 9.56 | 1.04 | 7.29E-05 | 1.90E-03 | 1.82E-04 |
| Phenylalanine metabolism | 10 | 2 | 10.03 | 1.04 | 8.77E-05 | 2.11E-03 | 2.06E-04 |
| Alanine, aspartate and glutamate metabolism | 28 | 7 | 9.41 | 1.04 | 1.34E-04 | 3.08E-03 | 2.92E-04 |
| Glutathione metabolism | 28 | 2 | 9.42 | 1.04 | 1.39E-04 | 3.08E-03 | 2.92E-04 |
| Pyrimidine metabolism | 39 | 2 | 10.51 | 1.04 | 1.56E-04 | 3.28E-03 | 3.12E-04 |
| Nicotinate and nicotinamide metabolism | 15 | 2 | 9.50 | 1.04 | 2.16E-04 | 4.33E-03 | 4.12E-04 |
| Purine metabolism | 65 | 2 | 10.81 | 1.04 | 2.53E-04 | 4.81E-03 | 4.60E-04 |
| Arginine biosynthesis | 14 | 2 | 9.63 | 1.04 | 3.00E-04 | 5.40E-03 | 5.22E-04 |
| Lysine degradation | 25 | 2 | 8.55 | 1.04 | 3.35E-04 | 5.70E-03 | 5.59E-04 |
| Selenocompound metabolism | 20 | 1 | 11.62 | 1.04 | 6.36E-04 | 1.02E-02 | 1.02E-03 |

# Example of QEA Summary Plot for top 25 Pathway



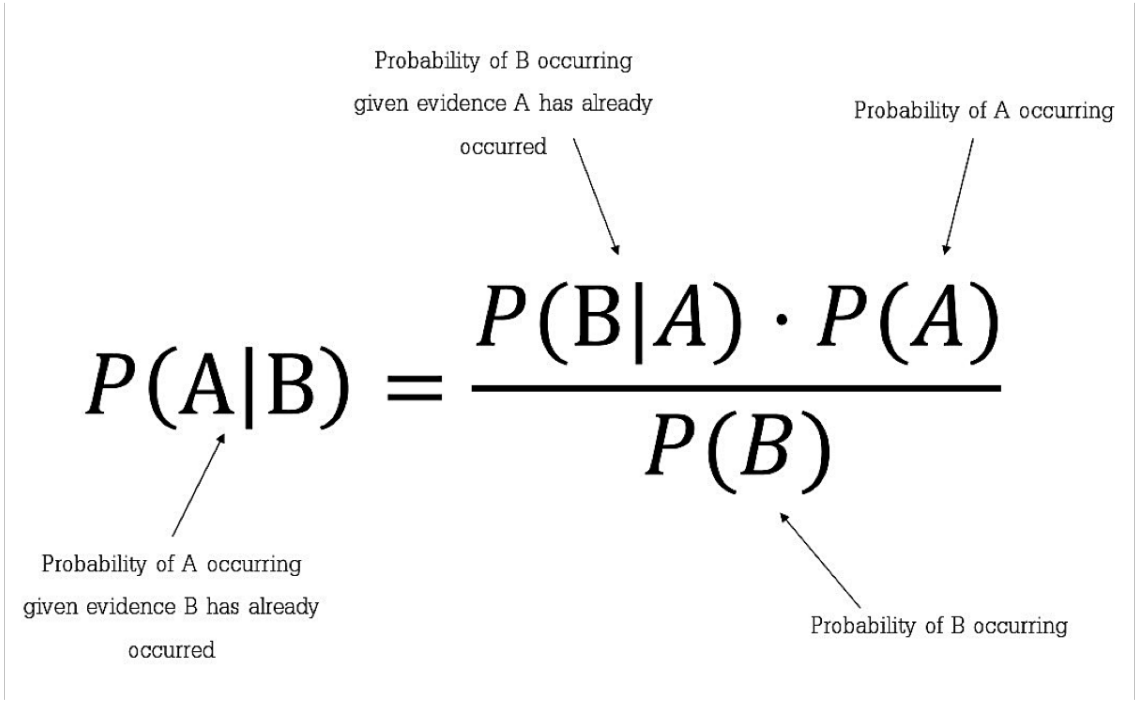
# Current Limitations of ORA and QEA

- Massive information loss (due to only the most significant metabolites are used and the rest are ignored).
- Only a number of identified metabolites is considered.
- It assumes each metabolite and pathway is independent of others (improper independence).
- Does not take the reactions among metabolites into account.

(Kamburov et al. 2011; Ren et al. 2015; Khatri, Sirota, and Butte 2012; Picart-Armada et al. 2018; Xia and Wishart 2010)

# Bayesian Statistics

- It is a probabilistic model that uses Bayes' Theorem to update the probability for a hypothesis as more evidence becomes available.
  - i.e. it uses probabilities as a tool to quantify uncertainty.



# Example

We want to know what is the probability that it rained given that it is wet outside?

A = it rained today, B = wet sidewalk.

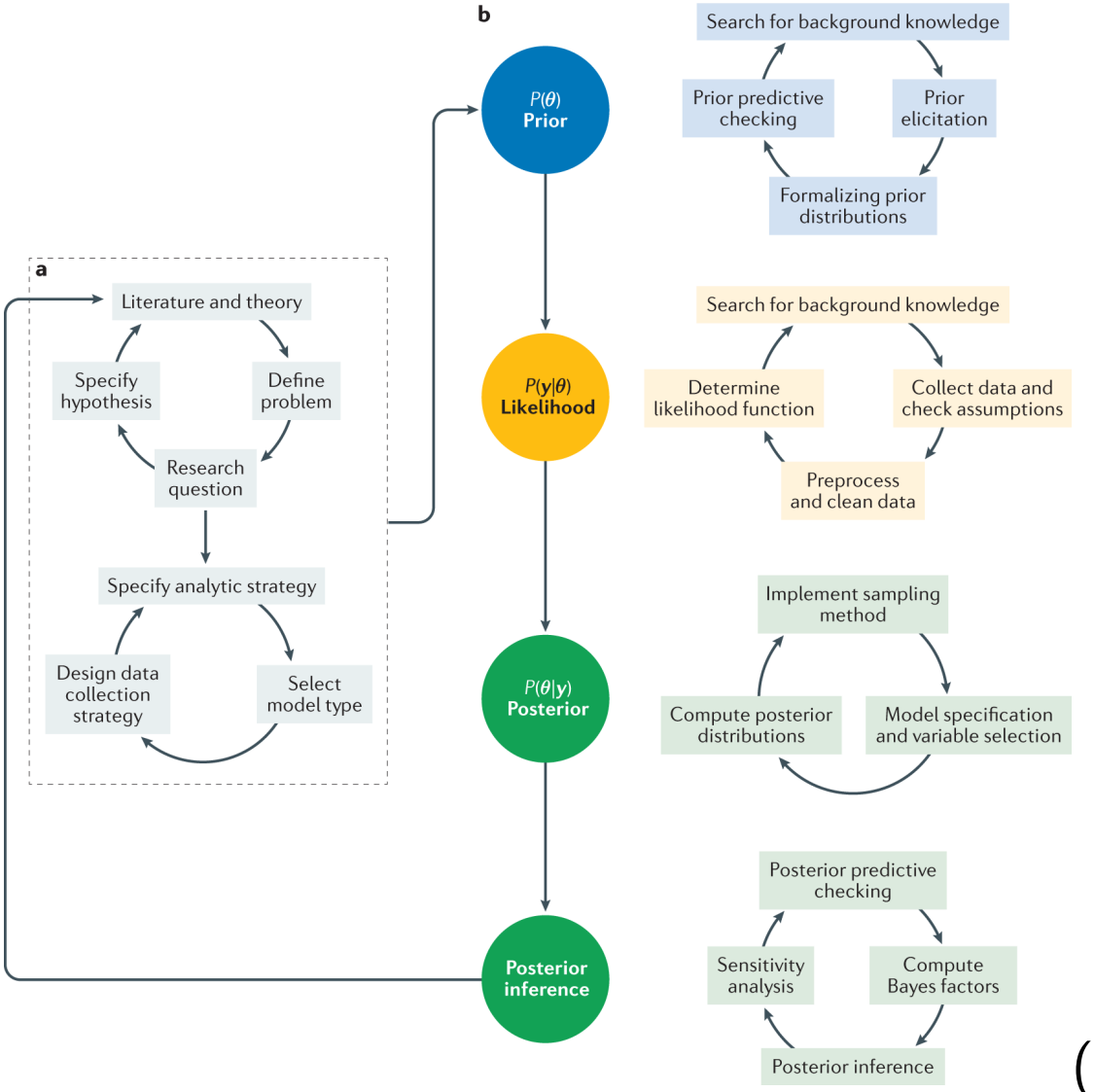
$$p(\text{rain}|\text{wet}) = \frac{p(\text{rain})p(\text{wet}|\text{rain})}{p(\text{wet})} = \frac{p(\text{rain})p(\text{wet}|\text{rain})}{p(\text{rain})p(\text{wet}|\text{rain})+p(\text{no rain})p(\text{wet}|\text{no rain})}$$

p(rain) – our initial beliefs – called prior.

p(rain|wet) – our final beliefs – called posterior.

p(wet) – how plausible is the evidence? what is the likelihood that it is wet outside given it rained and not rained. This is to ensure that the posterior is the correct probability distribution.

# Bayesian workflow




(Schoot et al. 2021)

# Strength of Bayesian

- Bayesian statistics are particularly useful for modelling networks of biological reactions which are typically modelled by large numbers of parameters.
- If we were to use frequentist methods, we'd require at least as many observations as there are parameters to fit a model (p»n issue)
- In contrast, Bayesian methods incorporate our prior knowledge of the system and use the experimental data to refine the estimates.
- It allows us to group the metabolites according to the pathways they sit under.
- Being a probabilistic model, it takes account uncertainty
- Information sharing amongst groups.
  - when the number of observations for groups varies widely, the groups with smaller numbers of observations will have improved inference about their group parameters by borrowing information

# References


Kamburov, Atanas, Rachel Cavill, Timothy MD Ebbels, Ralf Herwig, and Hector C Keun. 2011. "Integrated Pathway-Level Analysis of Transcriptomics and Metabolomics Data with IMPaLA." *Bioinformatics* 27 (20): 2917–18.

Khatri, Purvesh, Marina Sirota, and Atul J Butte. 2012. "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges." *PLoS Comput Biol* 8 (2): e1002375.

Picart-Armada, Sergio, Francesc Fernández-Albert, Maria Vinaixa, Oscar Yanes, and Alexandre Perera-Lluna. 2018. "FELLA: An r Package to Enrich Metabolomics Data." *BMC Bioinformatics* 19 (1): 1–9.

Ren, Sheng, Anna A Hinzman, Emily L Kang, Rhonda D Szczesniak, and Long Jason Lu. 2015. "Computational and Statistical Analysis of Metabolomics Data." *Metabolomics* 11 (6): 1492–1513.

Schoot, Rens van de, Sarah Depaoli, Ruth King, Bianca Kramer,