Machine Learning Approach to Recovering Metabolic Pathways from Metagenomics Sequences

Abdur Rahman M. A. Basher[1] and Steven J. Hallam[1,2]

[1]Graduate Program in Bioinformatics, [2]Department of Microbiology and Immunology, University of British Columbia

1 INTRODUCTION AND MOTIVATION

Microbial communities facilitate the majority of the biochemical activity on Earth, playing integral roles in energy and matter transformations in natural and engineered ecosystems. Metagenomics is used to analyze the genetic material of microbial communities directly from an environmental sample.

To estimate the metabolic potential of a metagenomic sample we devise a novel approach to reconstructing biological pathways from enzyme annotations and environmental parameters. Our approach enjoys a modular, flexible strategy based on statistical hierarchical Bayesian deep framework that encodes emergent information represented in MetaCyc, a highly curated database of enzyme sequences, reactions, and pathways. The model is based on graphical modeling techniques to infer latent pathways represented as mixture components in a sample. For the training, we adopt a collapsed Gibbs sampling technique to examine the genetic content of metagenomic datasets. Further, the model is well defined mathematically and aligns with the biological interpretations. Based on our preliminary analysis, we anticipate that our model can outperform the PathoLogic algorithm on a single organism task.

2 APPROACH

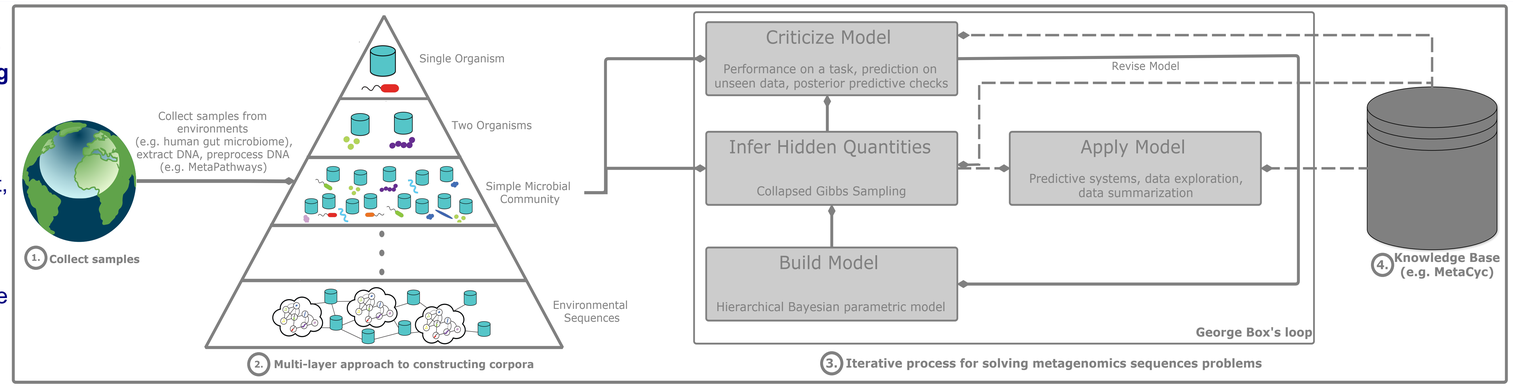

Figure 1. A modern interpretation of the perspective of George Box's loop to iterative process for solving metagenomics data analysis problems. Metagenomic sequence information is assembled and ORFs are predicted and annotated (1). Next, the enzymes and their associated pathways are curated based on a multilayer approach where the top layer dataset is synthesized using one organism (positive) mixed with non-overlapping pathways-enzymes (negative) (2). The next layer comprises of two organisms with partially overlapping pathways-enzymes (positive) and non-overlapping pathways-enzymes (negative). Continuing to add pathway information from more organisms approaches the metabolic potential of a microbial community. Afterward, Box's loop comes into play in which an iterative cycle of experimental design, model formulation, model criticism, and application refines the model (3). In the first step of the loop, a probability model is built with a well defined mathematical object. Observed data enter the picture in the second step of Box's loop. Here, the computational aspects are applied to infer the pathways using an inference algorithm to compute the posterior distribution (e.g. collapsed Gibbs sampling) and a knowledge base (e.g. MetaCyc) (4). Finally, the trained model is tested against real data, identifying the important ways that it succeeds and fails in extracting pathways.

3 HIERARCHICAL BAYESIAN PARAMETRIC MODEL

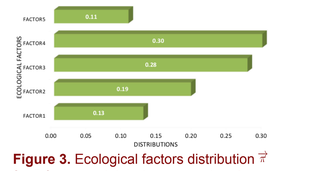

Figure 3. Ecological factors distribution $\vec{\pi}$ for 5 factors exhibited in a sample.

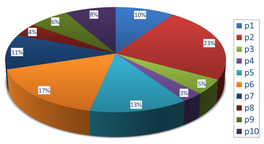

Figure 7. Pathways distribution $\vec{\Omega}$ for 10 pathways in a sample.

PATHWAYS-REACTIONS RELATION

	r_1	r_2	r_3	r_4	r_5
p_1	1.00	0.00	0.00	0.00	1.00
p_2	0.00	0.00	0.00	1.00	0.00
p_3	0.00	1.00	1.00	0.00	0.00
p_4	0.00	0.00	0.00	1.00	1.00
p_5	1.00	0.00	1.00	0.00	0.00
p_6	0.00	1.00	1.00	0.00	0.00
p_7	1.00	1.00	0.00	1.00	1.00
p_8	1.00	1.00	0.00	0.00	0.00
p_9	0.00	1.00	1.00	1.00	0.00
p_{10}	1.00	0.00	0.00	0.00	1.00

Figure 8. A sparse binary Λ_P for 10 pathways with 5 reactions filled with 0s and 1s, randomly.

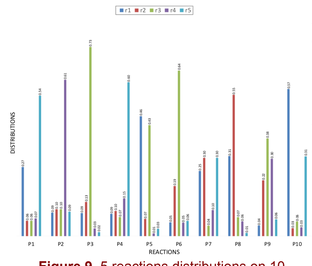

Figure 9. 5 reactions distributions on 10 pathways defined in Φ.

REACTIONS-ENZYMES RELATION

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
r_1	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
r_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00
r_3	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00
r_4	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00
r_5	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00

Figure 10. A sparse binary matrix Λ_R for 10 enzymes allocated randomly on 5 reactions.

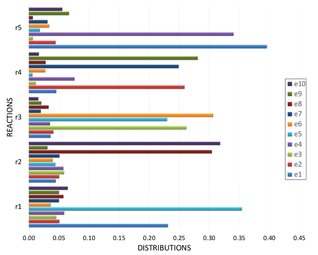

Figure 11. 10 enzymes distributions over 5 reactions defined in Ψ.

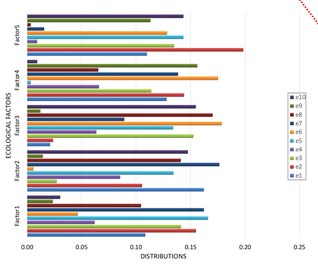

Figure 4. 10 enzymes distribution over 5 factors encoded in Θ.

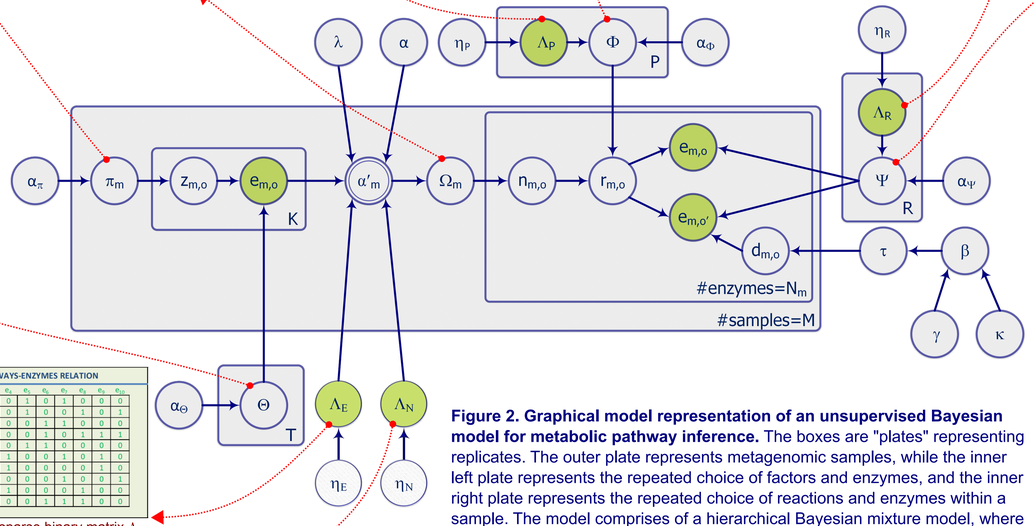

Figure 2. Graphical model representation of an unsupervised Bayesian model for metabolic pathway inference. The boxes are "plates" representing replicates. The outer plate represents metagenomic samples, while the inner left plate represents the repeated choice of factors and enzymes, and the inner right plate represents the repeated choice of reactions and enzymes within a sample. The model comprises of a hierarchical Bayesian mixture model, where enzymes constitute reactions, reactions are mixed to form pathways, pathways are determined by reactions and environmental parameters, and each sample is treated as a vector of environmental parameter distributions.

PATHWAYS-PATHWAYS RELATION

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
p_1	0.00	0.38	0.00	0.00	0.15	0.00	0.00	0.38	0.00	0.00
p_2	0.29	0.00	0.41	0.08	0.00	0.00	0.00	0.22	0.00	0.00
p_3	0.00	0.31	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.31
p_4	0.00	0.33	0.00	0.00	0.33	0.11	0.00	0.00	0.00	0.00
p_5	0.26	0.00	0.00	0.26	0.00	0.00	0.00	0.26	0.00	0.18
p_6	0.00	0.00	0.00	0.20	0.00	0.00	0.32	0.49	0.00	0.00
p_7	0.00	0.00	0.14	0.00	0.00	0.48	0.00	0.19	0.00	0.19
p_8	0.07	0.22	0.00	0.00	0.00	0.12	0.14	0.00	0.22	0.22
p_9	0.00	0.00	0.17	0.00	0.28	0.00	0.00	0.26	0.28	0.00
p_{10}	0.00	0.00	0.18	0.00	0.29	0.00	0.16	0.36	0.00	0.00

Figure 6. An adjacency matrix Λ_N for 10 pathways with random transition probabilities drawn from a Bernoulli prior distribution.

PATHWAYS-ENZYMES RELATION

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
p_1	0	1	1	0	1	0	1	0	0	0
p_2	0	0	0	0	1	0	0	1	0	1
p_3	1	1	0	0	0	1	1	0	0	0
p_4	0	1	0	0	0	1	0	1	1	1
p_5	1	0	0	0	1	1	0	0	0	0
p_6	0	0	0	1	0	0	1	0	1	0
p_7	0	0	1	1	0	0	0	1	0	0
p_8	0	0	1	0	0	0	1	0	0	1
p_9	0	1	0	1	0	0	0	1	0	0
p_{10}	1	0	0	0	0	1	1	1	0	0

Figure 5. A sparse binary matrix Λ_E for 10 pathways with 10 enzymes filled with 0s and 1s, randomly.

4 COLLAPSED GIBBS SAMPLING ALGORITHM

The overall generative process is summarized below:

(1) Sample a distribution $\tau \sim Beta(.|\gamma, \kappa)$

(2) For each ecological factors $t \in \{1, ..., T\}$:

  a. Sample a distribution over enzymes $\Theta_t = (\theta_{t,1}, ..., \theta_{t,P})^\intercal \sim Dirichlet(.|\overrightarrow{\alpha_\Theta})$

(3) For each pathway $i \in \{1, ..., P\}$:

  a. Sample labels from $\Lambda_P^i \in \{0,1\}_1^R \sim Bernoulli(.|\overrightarrow{\eta_P}) : \forall j(\in i) = 1$ where $j \in \{1, ..., R\}$

  b. Sample a distribution over reactions $\Phi_i = (\phi_{i,1}, ..., \phi_{i,R})^\intercal \sim Dirichlet(.|\overrightarrow{\alpha_\Phi}), \Lambda_P^i$

(4) For each reaction $j \in \{1, ..., R\}$:

  a. Sample labels from $\Lambda_R^j \in \{0,1\}_1^E \sim Bernoulli(.|\overrightarrow{\eta_R}) : \forall e(\in j) = 1$ where $e \in \{1, ..., E\}$

  b. Sample a distribution over enzymes $\Psi_j = (\psi_{j,1}, ..., \psi_{j,E})^\intercal \sim Dirichlet(.|\overrightarrow{\alpha_\Psi}), \Lambda_R^j$

(5) For each sample $m \in \{1, ..., M\}$:

  a. Sample a distribution over ecological factors $\overrightarrow{\pi_m} = (\pi_{m,1}, ..., \pi_{m,T})^\intercal \sim Dirichlet(.|\overrightarrow{\alpha_\pi})$

  b. For each enzyme $o \in \{1, ..., K_m\}$:

    1. Sample a factor label $z_{m,o} \sim Multinomial(.|\overrightarrow{\pi_m})$

    2. Sample an enzyme label $e_{m,o} \sim Multinomial(.|\overrightarrow{\Theta_{z_{m,o}}})$

  c. Compute the Dirichlet prior $\overrightarrow{\alpha'_m}$ for m using $\overrightarrow{\alpha'_m} = \overrightarrow{\lambda}.\overrightarrow{\rho'_m} + \alpha$
  $\overrightarrow{\rho'_m} = (\vec{s}.\mathbf{A}_E^\intercal).\mathbf{A}_N$ and $\vec{s} = (\frac{C_{m,i}}{K_m}, ..., \frac{C_{m,E}}{K_m})$

  d. Sample a distribution over pathways $\overrightarrow{\Omega_m} = (\Omega_{m,1}, ..., \Omega_{m,P})^\intercal \sim Dirichlet(.|\overrightarrow{\alpha'_m})$

  e. For each observed enzyme $o \in \{1, ..., N_m\}$:

    1. Sample a pathway label $n_{m,o} \sim Multinomial(.|\overrightarrow{\Omega_m})$

    2. Sample a reaction label $r_{m,o} \sim Multinomial(.|\overrightarrow{\Phi_{n_{m,o}}})$

    3. Sample an enzyme $e_{m,o} \sim Multinomial(.|\overrightarrow{\Psi_{r_{m,o}}})$

    4. Sample a binary label $d_{m,o} \sim Bernoulli(.|\tau)$

    5. Sample an additional enzyme $e_{m,o'}$ according to:
    $(e_{m,o} \sim Multinomial(.|\overrightarrow{\Psi r_{m,o}})$ if $d_{m,o} = 0) \vee (\{e_{m,o}, e_{m,o'}\} \sim \{Multinomial(.|\overrightarrow{\Psi r_{m,o}})\}_1^2$ if $d_{m,o} = 1)$

Terminologies

M	Number of samples to generate.
T	Number of ecological factors.
P	Number of metabolic pathways.
R	Number of reactions.
E	Number of enzymes including dummy enzymes.
K_m	Number of sampling process for the mth sample.
$C_{m,i}$	Number of times an enzyme i was sampled for the mth sample.
N_m	Number of enzymes in sample m.
$z_{m,o}$	Mixture indicator that chooses the factor for the oth enzyme in sample m.
$e_{m,o}$	Enzyme indicator for the oth enzyme in sample m.
$n_{m,o}$	Mixture indicator that chooses the pathway for the oth enzyme in sample m.
$r_{m,o}$	Reaction indicator for the oth enzyme in sample m.
$d_{m,o}$	Dummy variable that specifies whether to choose enable for additional o'th enzyme or not in sample m.
$e_{m,o'}$	Additional Enzyme indicator for the o'th enzyme in sample m.
$\overrightarrow{\pi_m}$	Ecological factors proportion for sample m.
Θ	Multinomial distribution over E enzymes along T ecological properties (T × E matrix).
Φ	Multinomial distribution on R reactions along P pathways (P × R matrix).
Ψ	Multinomial distribution on E enzymes along R reactions (R × E matrix).
$\overrightarrow{\Omega_m}$	Pathways proportion for sample m.
$\overrightarrow{\alpha_m}$	Scaled, smoothed, normalized parameter for pathways in sample m (P-vector).
$\Lambda_E\ \Lambda_P\ \Lambda_R$	Sparse binary indicator matrices of sizes (P × E), (P × R), and (R × E), respectively.
Λ_N	Stochastic matrix of size P × P.
$\vec{\lambda}$	Hyperparameter that specifies the weight contributed to pathways (P-vector).
$\overrightarrow{\alpha_\pi}\ \overrightarrow{\alpha_\Theta}\ \overrightarrow{\alpha_\Phi}\ \overrightarrow{\alpha_\Psi}$	Hyperparameters on π, Θ, Φ, and Ψ, respectively.
α	Symmetric hyperparameter on $\overrightarrow{\alpha'_m}$.
$\overrightarrow{\eta_E}\ \overrightarrow{\eta_N}\ \overrightarrow{\eta_P}\ \overrightarrow{\eta_R}$	Bernoulli prior distribution on $\Lambda_E, \Lambda_N, \Lambda_P$, and Λ_R respectively.
τ	Bernoulli prior distribution on $d_{m,o}$.
β	Beta distribution.
$\gamma\ \ \kappa$	Hyperparameters for β.

5 CONCLUSION AND DISCUSSION

Inspired by Box's loop, an unsupervised hierarchical deep Bayesian architecture is developed to detect pathways that are present in ecological sequences, which are constructed in a multi-layer approach.

Further work includes adopting a supervised strategy to recovering pathways, studying the correlated pathways and enzymes to better understanding the microbial interactions, examining pathways abundances to assist in capturing the global metabolic network in samples, constructing a well-defined set of ecological factors contributing to pathways inference, learning mixtures of taxa in micobiomes, suggesting ways to capture super-pathways in MetaCyc, and proposing ways to minimize computational intricacies in a sparse metagenomics sequences.

6 REFERENCES

[1]- Karpe, Peter D., et al. (2011). "The pathway tools pathway prediction algorithm." Standards in genomic sciences 5.3: 424.

[2]- Abubucker, Sahar, et al. (2012). "Metabolic reconstruction for metagenomic data and its application to the human microbiome." PLoS Comput Biol 8.6.

[3]- Shafiei, Mahdi, et al. (2014). "BiomeNet: A Bayesian model for inference of metabolic divergence among microbial communities." PLoS Comput Biol 10.11.

[4]- Hanson, Niels W., et al. (2014). "Metabolic pathways for the whole community." BMC genomics 15.1.

[5]- Blei, David M. (2014). "Build, compute, critique, repeat: Data analysis with latent variable models." Annual Review of Statistics and Its Application 1: 203-232.

[6]- Konwar, Kishori M., et al. (2015). "MetaPathways v2. 5: quantitative functional, taxonomic and usability improvements." Bioinformatics 31.20: 3345-3347.