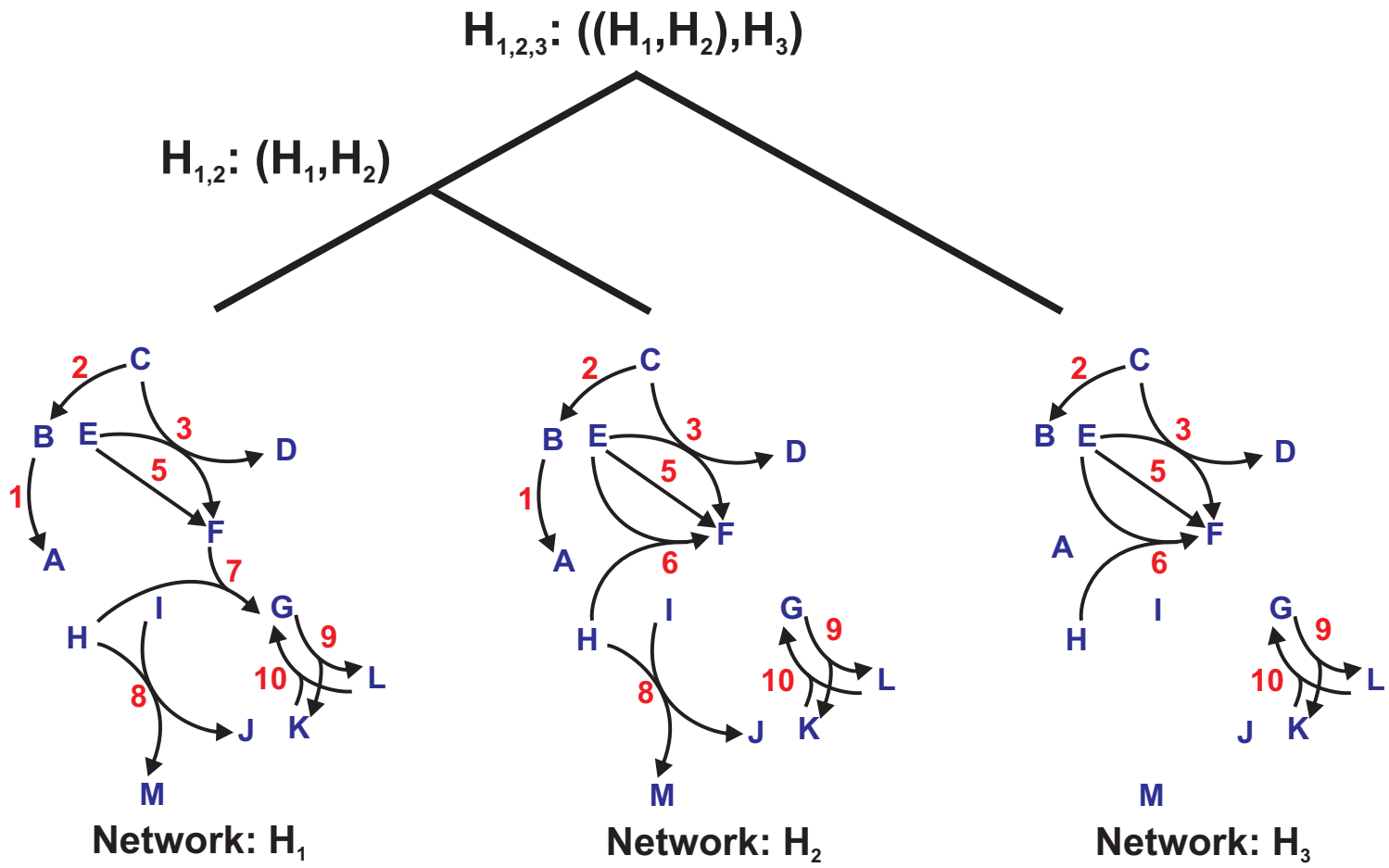
A Bayesian approach to the evolution of metabolic networks on a phylogeny

Aziz Mithani[†], Gail Preston[‡] and Jotun Hein[†]
mithani@stats.ox.ac.uk

[†]Department of Statistics and [‡]Department of Plant Sciences,
University of Oxford, South Parks Road, Oxford, UK

28 August 09

The problem



Contents

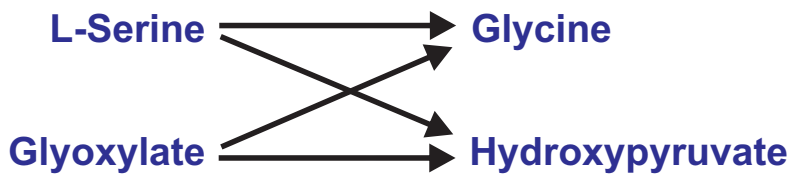
Metabolic networks as hypergraphs

Metabolic Networks

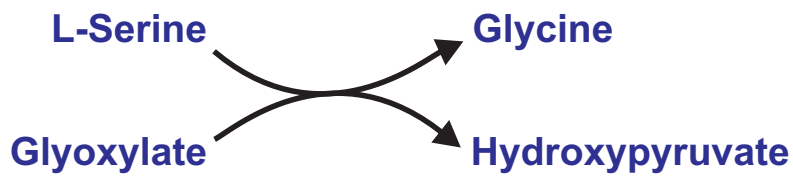
- Set of biochemical reactions connecting two or more metabolites

Representations

Graph:



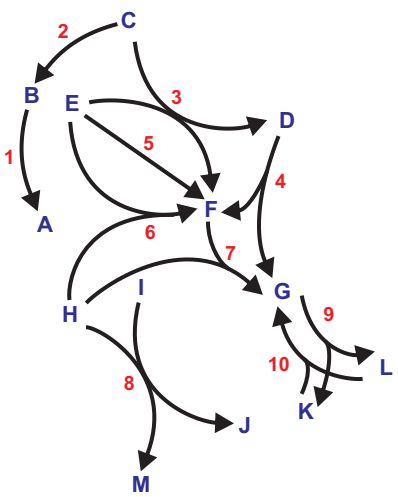
Hypergraph:



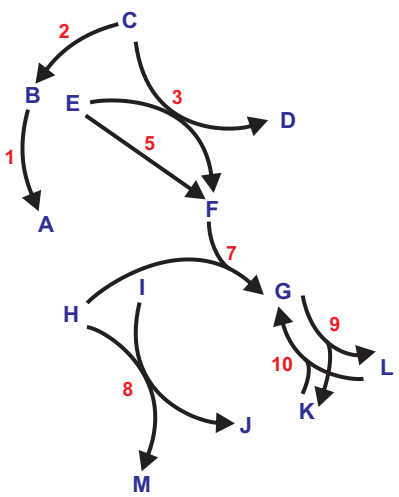
Advantages of Hypergraphs

- Biologically relevant - set of metabolites connected by a reaction
- Captures the relationship between multiple metabolites in a reaction
- Allows to think in terms of gain / loss of reactions

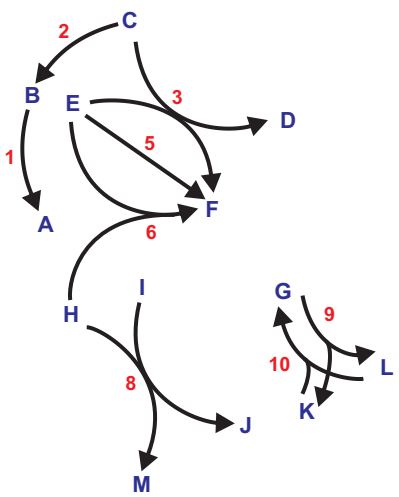
Assumptions



Reference Network (H)



Network H_1: 1110101111



Network H_2:1110110111

- A network containing all hyperedges in $\mathcal{E}$ called a *Reference network* such that $|\mathcal{E}| = M$.
- Hyperedges in the reference network are labelled 1 to M.
  - A network x can be represented as a sequence of 0s and 1s.
- A reversible reaction is represented by two hyperedges.

Network evolution as a continuous time Markov process

State: The set $G \subseteq \mathcal{E}$ of hyperedges present in the network

Next State: Characterised by the insertion of one edge or the deletion of one edge

System Dynamics: Described by the following master equation

$$\frac{\mathrm{d}\mathbb{P}(G)}{\mathrm{d}t} = \mu \sum_{G' \in I(G)} \mathbb{P}(G') + \lambda \sum_{G'' \in D(G)} \mathbb{P}(G'') - \mathbb{P}(G)\Big(\lambda |I(G)| + \mu |D(G)|\Big)$$

where

- $I(G)$: set of networks reachable by insertion of a single hyperedge
- $D(G)$: set of networks reachable by deletion of a single hyperedge
- λ: rate of insertion of a hyperedge
- μ: rate of deletion of a hyperedge

(Mithani et al., *Bioinformatics* 2009)

Neighbour-dependent model of network evolution

- **Neighbours:** Hyperedges sharing at least one node
- The rate from network x to x' depends on x_i, x_i' and the neighbouring hyperedges $\Psi(x_i)$, and is given as follows.

$$\gamma(x_i'; x_i, \Psi(x_i)) = Q(x_i, x_i')F(x_i, \Psi(x_i))$$

where

- Q is the 2×2 base rate matrix:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

- The function F corresponds to neighbourhood component.

- State Space Size: 2^m

(Mithani et al., *Bioinformatics* 2009)

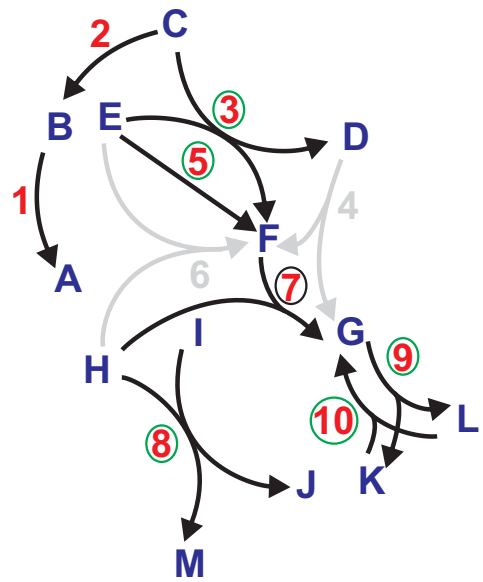
Neighbour-dependent model of network evolution (2)

Neighbourhood Component

$$F(x_i, \Psi(x_i)) = \begin{cases} \frac{|\Psi(x_i)|}{\sum_{i \neq j} x_j} & |\Psi(x_i)| > 0 \\ \frac{1}{M+1} & \text{Otherwise.} \end{cases}$$

Example

$$F(x_7, \Psi(x_7)) = \frac{5}{7} = 0.714$$



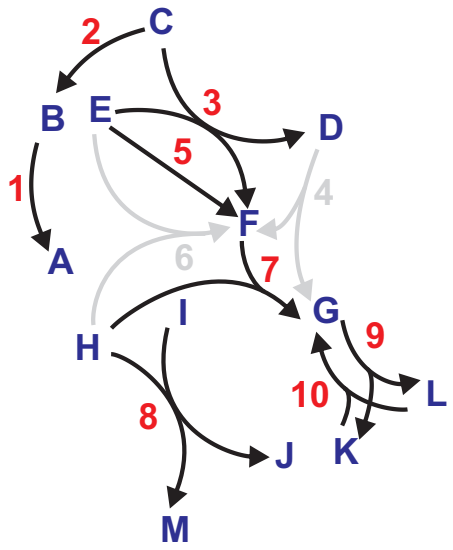
Let $(\lambda, \mu) = (0.05, 0.03)$, then

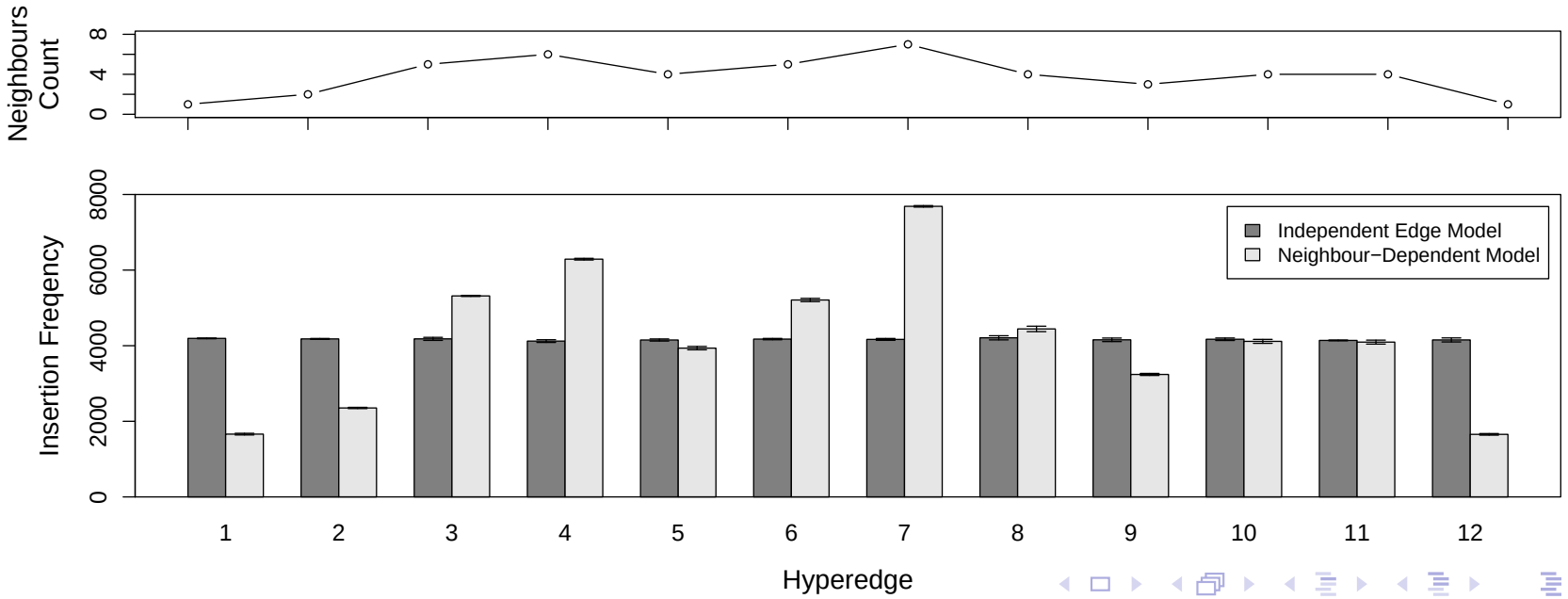
$$Q = \begin{bmatrix} -0.05 & 0.05 \\ 0.03 & -0.03 \end{bmatrix}$$

and

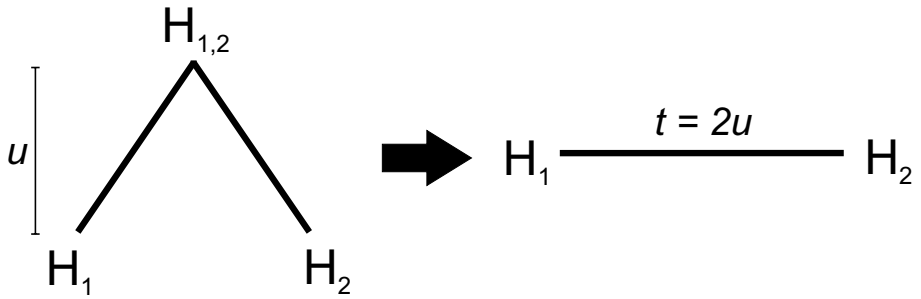
$$\gamma(x_7'; x_7, \Psi(x_7)) = 0.03 \times 0.714 = 0.02142$$

Simulation results





Likelihood of network evolution



- The likelihood of a pair of networks:

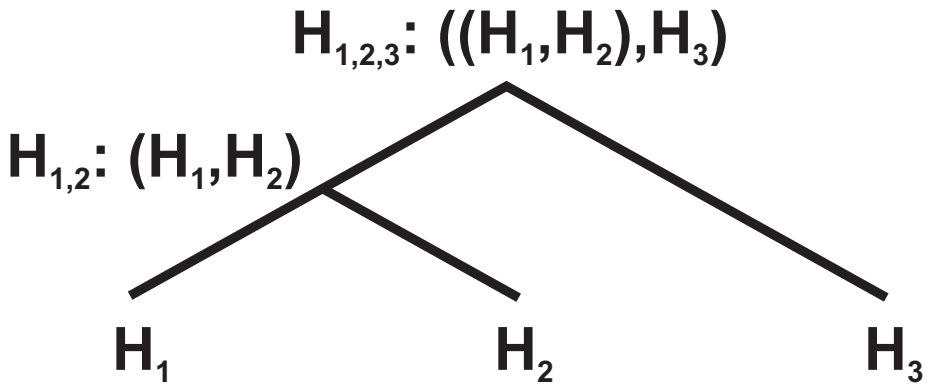
$$\begin{aligned} L_{\Theta,u}(H_1, H_2) &= \sum_{H_{1,2}} P_\infty(H_{1,2}) P_{\Theta,u}(H_1|H_{1,2}) P_{\Theta,u}(H_2|H_{1,2}) \\ &= P_\infty(H_1) P_{\Theta,2u}(H_2|H_1) \qquad \text{(assuming reversibility)} \end{aligned}$$

- The probability $P_{\Theta,t}(N_2|N_1)$ is calculated from the transition matrix $P(t)$ given as

$$P(t) = \exp(tQ) = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}$$

where Q is the rate matrix.

Likelihood of network evolution (2)



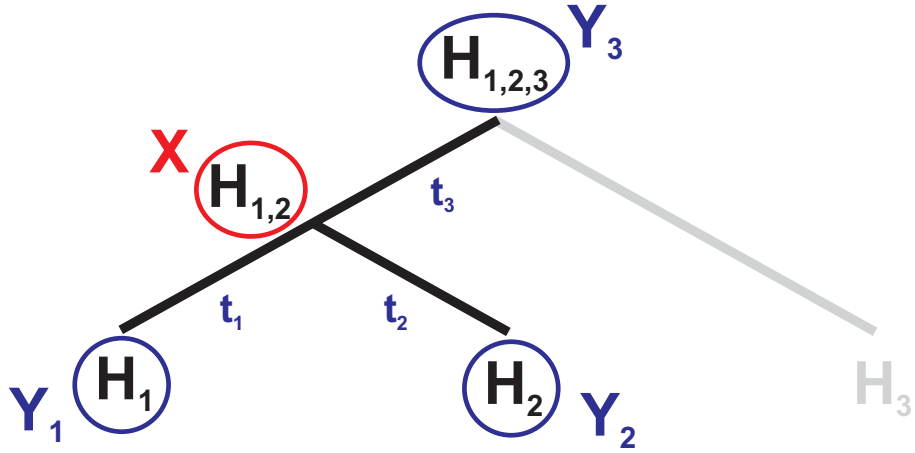
The likelihood of a set of networks connected by a phylogeny:

$$L_\Theta(\mathcal{T}) = \sum_{H_{1,2,3}} \left\{ P_\Theta(H_{1,2,3}) P_{\Theta,t_3}(H_3|H_{1,2,3}) \right.$$
$$\left. \sum_{H_{1,2}} \left\{ P_{\Theta,t_{1,2}}(H_{1,2}|H_{1,2,3}) P_{\Theta,t_1}(H_1|H_{1,2}) P_{\Theta,t_2}(H_3|H_{1,2}) \right\} \right\}$$

Sampling internal nodes

Idea

- Sample each internal network X by conditioning on its three neighbours Y_1, Y_2 and Y_3.



Sampling Procedure

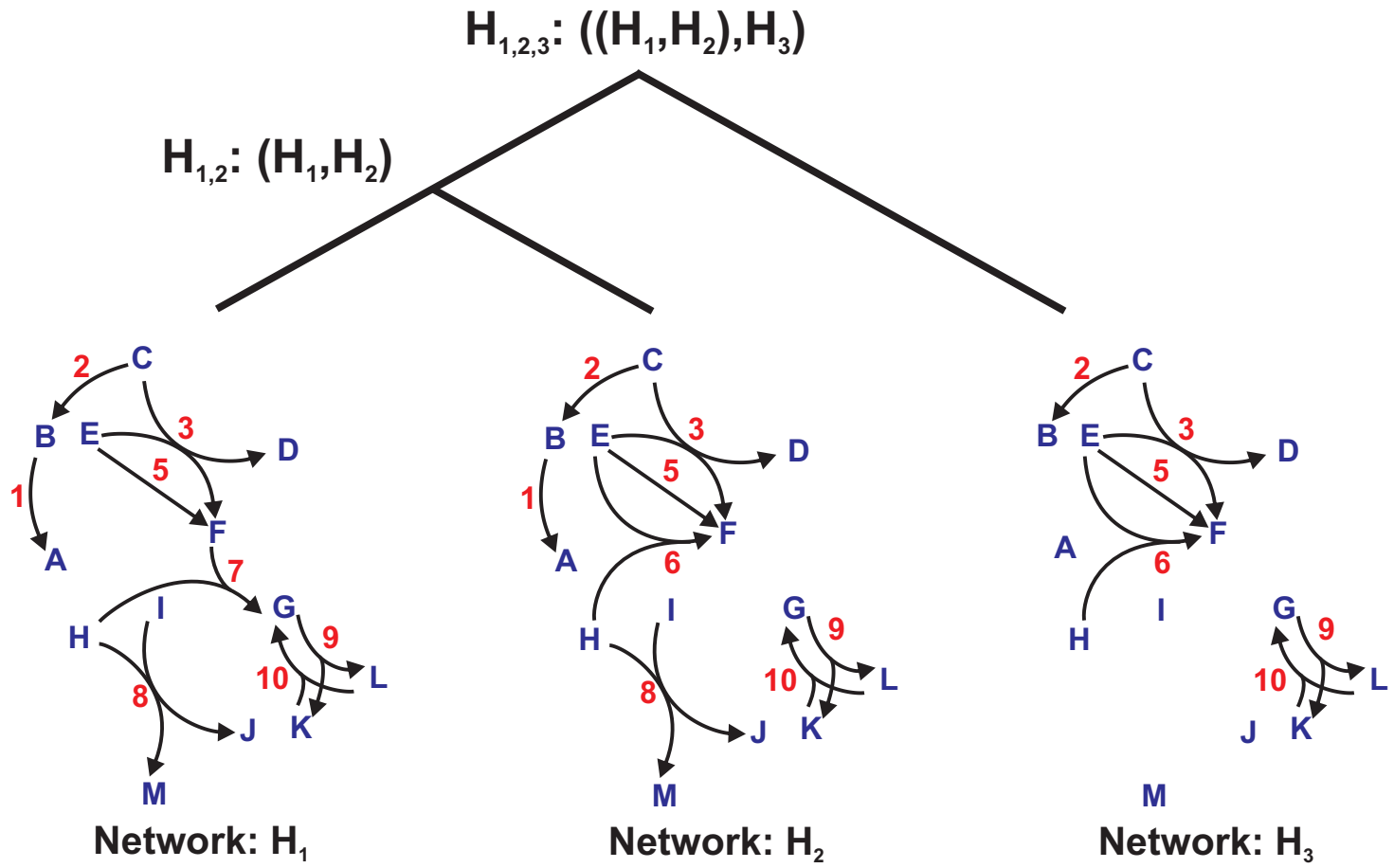
- Calculate the 2×2 rate matrix Γ for each hyperedge.
- For each neighbour, exponentiate Γ to get transition probabilities

$$P_{\Theta,t_k}(Y_k(i)|X(i)) = \exp(t_k \Gamma_X)$$

- Sample the new state $s_i' = \{0, 1\}$ for hyperedge i from the distribution

$$P(s_i) \propto \pi(s_i) \prod_{k=1}^{3} P_{\Theta,t_k}(Y_k(i)|s_i)$$

Likelihood calculation for toy networks



Results: Likelihood calculation for toy networks

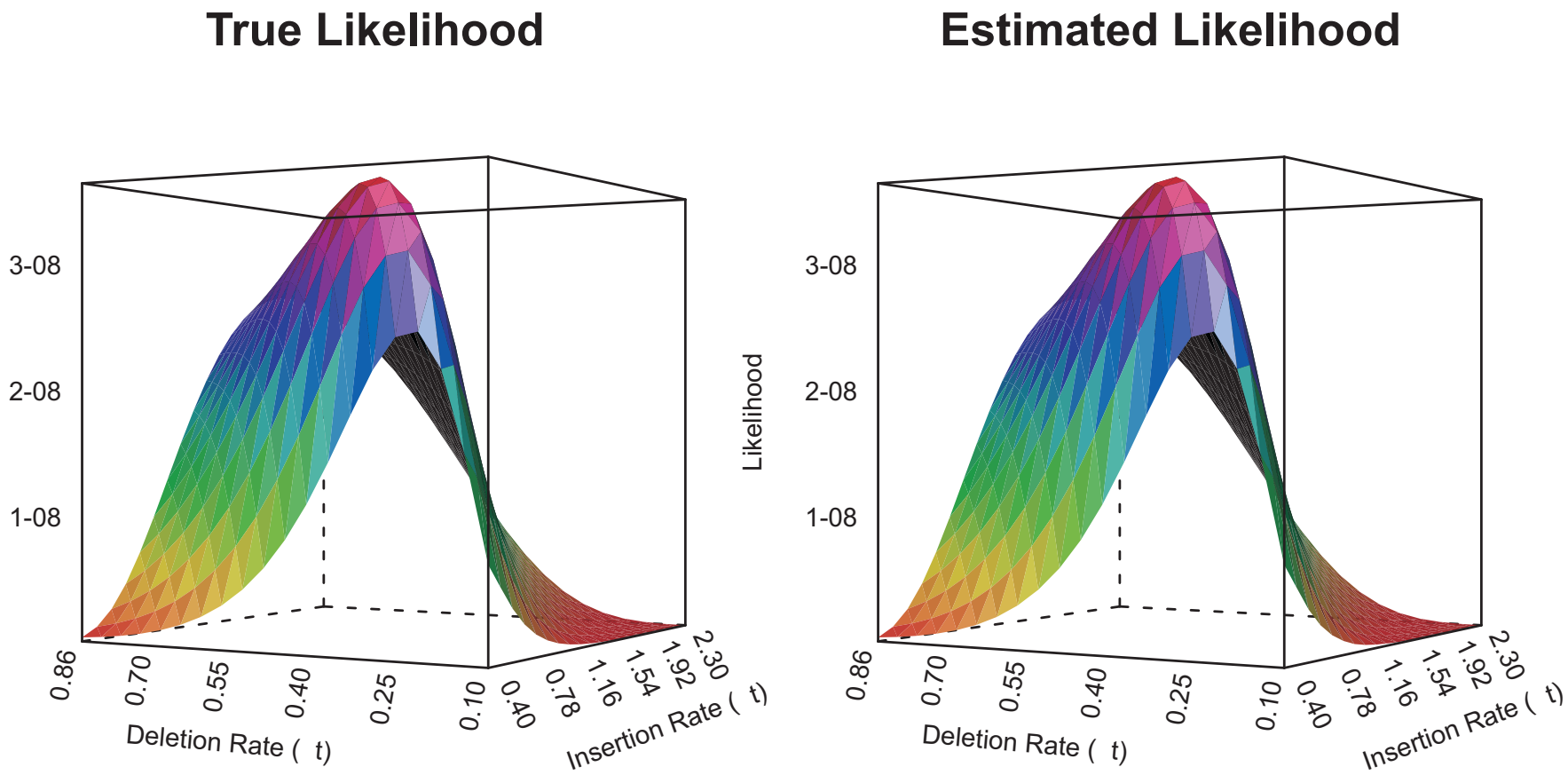


Figure: Likelihood surfaces calculated by matrix exponentiation and by using MCMC for toy networks

Gibbs sampler for parameter estimation

Likelihood function

$$L_\Theta(\mathcal{T}) = \sum_{N_{\text{root}}} P_\Theta(N_{\text{root}}) L_\Theta(N_{\text{root}})$$

$$L_\Theta(N) = \sum_{N_l} P_{\Theta,t_l}(N_l|N) L_\Theta(N_l) \sum_{N_r} P_{\Theta,t_r}(N_r|N) L_\Theta(N_r)$$

Estimation of Parameters

Idea: Easier to calculate conditional distributions $P_t(\mathcal{T}|\theta)$ and $P_t(\theta|\mathcal{T})$ than to obtain the marginal $L_t(\theta)$ by the summation of the joint density.

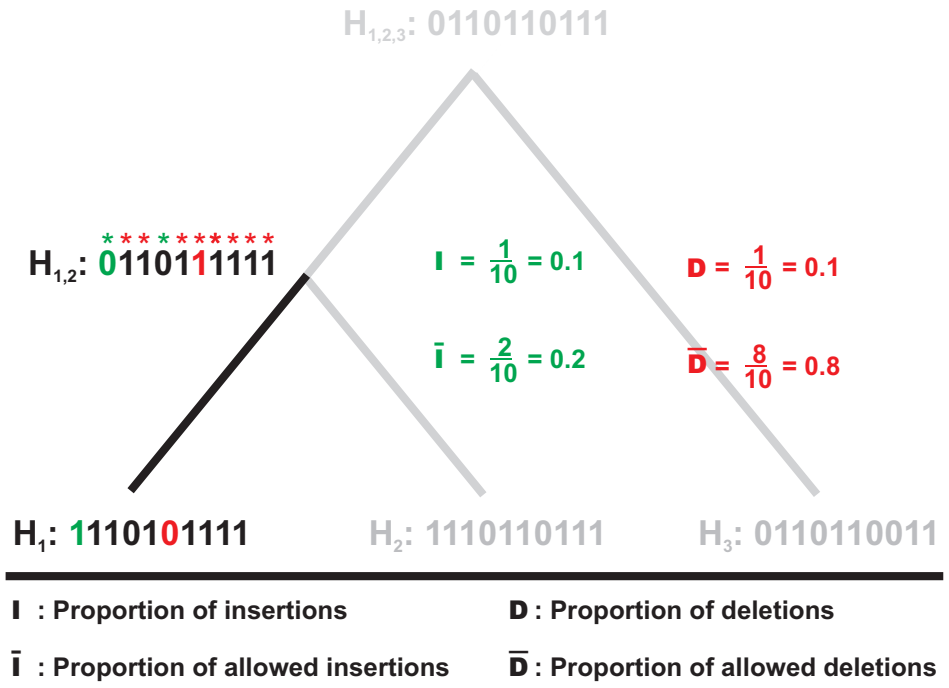
Algorithm

- Choose initial values for the parameters $\Theta^{(0)}$.
- Generate $\mathcal{T}^{(0)}$ by sampling internal nodes using $\Theta^{(0)}$.
- Use $\mathcal{T}^{(0)}$ to generate $\Theta^{(1)}$ by drawing from the distribution $P(\Theta|\mathcal{T})$.
- Repeat n times to generate a Gibbs sequence where a subset of points $(\mathcal{T}^{(i)}, \Theta^{(i)})$, $1 \leq i \leq n$, are the simulated estimates from the joint distribution $P(\mathcal{T}, \Theta)$.

Rates sampling

Idea

- Sample rates based on proportion of insertions and deletions events



Rate Proposal

- Using a gamma distribution

$$\gamma \sim \Gamma(k, \theta); \qquad \gamma = \lambda, \mu$$

where hyper-parameters are calculated from the given tree $\mathcal{T}$.

- **Insertion rate** $k_\lambda = \sum_{i,j} I_{H_i \Rightarrow H_j} + 1, \qquad \theta_\lambda = \sum_{i,j} \bar{I}_{H_i \Rightarrow H_j}$
- **Deletion rate** $k_\mu = \sum_{i,j} D_{H_i \Rightarrow H_j} + 1, \qquad \theta_\mu = \sum_{i,j} \bar{D}_{H_i \Rightarrow H_j}$

Rates sampling (2)

Proposal Probability

- The proposal probability $q(\lambda', \mu' | \lambda, \mu)$ for the rate parameters is given as

$$q(\lambda', \mu' | \lambda, \mu) = \prod_{\gamma=\lambda,\mu} q(\gamma' | \gamma)$$

such that

$$q(\gamma' | \gamma) \propto \gamma^{k_\gamma - 1} \exp\left(-\gamma \theta_\gamma\right).$$

Results: Parameter estimation for toy networks

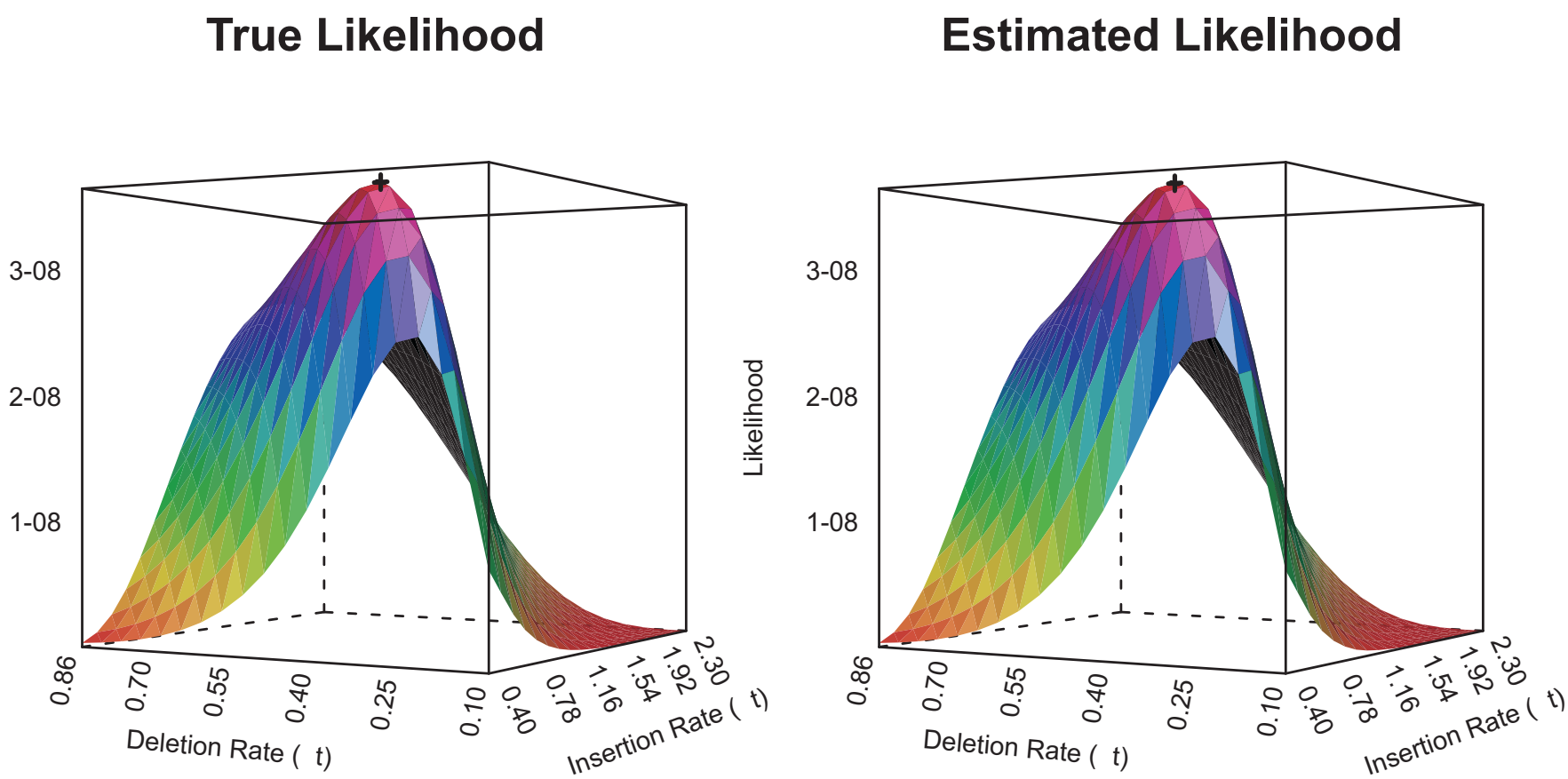
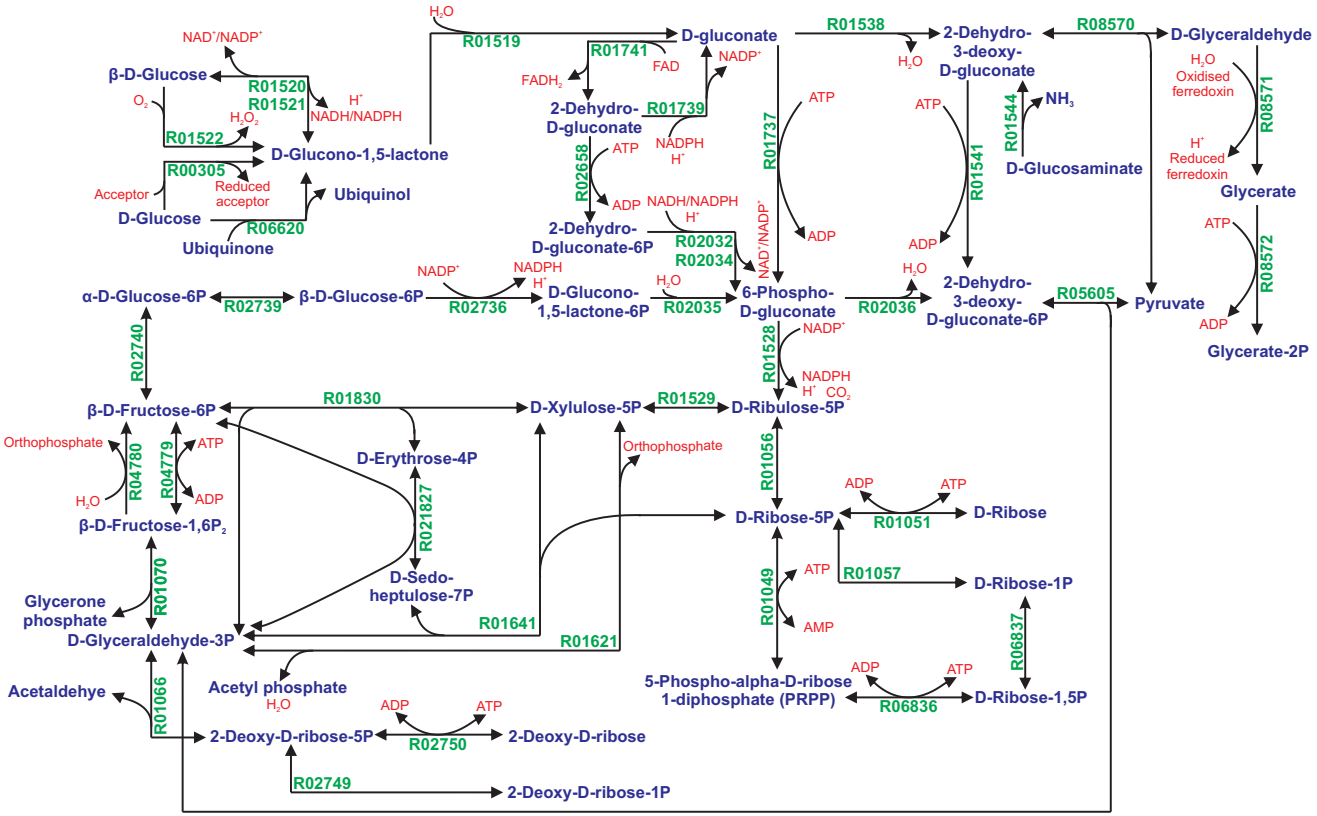


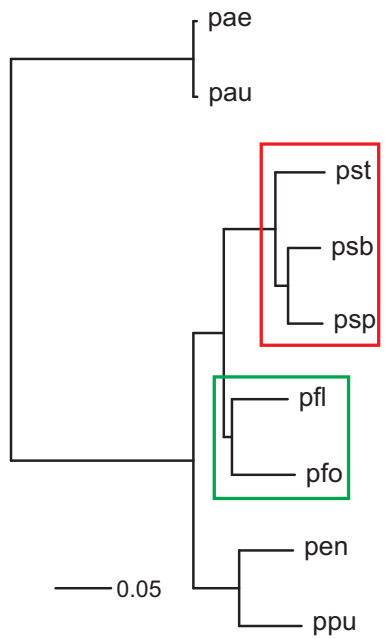
Figure: Parameter estimation using the Gibbs sampler

Parameter estimation for metabolic networks

Pentose Phosphate Pathway



Pseudomonas



Results: Parameter estimation for metabolic networks

Evolution rates

- Insertion rate higher for pathway maps involved in central metabolism and amino acid biosynthesis than those involved in secondary metabolism and amino acid degradation
- Rates higher in *Pseudomonas syringae* compared to *Pseudomonas fluorescens*
  - Supports experimental findings – High number of deletions in *P. syringae* lineage
- Low insertion to deletion ratio (λ/μ) for pathway maps related to amino acids which are poor nutrient sources

Neighbourhood structure

- Pathway maps involved in central metabolism and metabolism of essential amino acid have strong neighbourhood structure.

Summary and further work

Summary

- A stochastic model of network evolution using neighbour dependence
  - hypergraph based - intuitive
  - biologically relevant
- Gibbs sampler for internal nodes sampling
- Gibbs sampler for parameter estimation

Further Work

- Extensions to the model
  - Reaction structure
    - number of metabolites
    - chemical efficiency
  - Ortholog data
    - status of the reaction in closely related species
- Multiple changes per step (lateral gene transfer)

Thank You!