

Chapter 1

De Novo Signaling Pathway Reconstruction From Multiple Data Sources

Dongxiao Zhu[†], Michael G. Rabbat[‡], Alfred O. Hero,
III^{††}, Robert Nowak[‡] and Mario Figueiredo^{‡‡}

[†] *Stowers Institute for Medical Research,
1000 East 50th Street, Kansas City, MO 64110, USA*

e-mail: `doz@stowers-institute.org`

[‡] *Department of Electronic Engineering, University of Wisconsin, Madison
3610 Engineering Hall, 1415 Engineering Drive Madison, WI 53706, USA*

e-mail: `rabbat@cae.wisc.edu`, `nowak@engr.wisc.edu`

^{††} *Department of Electronic Engineering and Computer Science, University
of Michigan, Ann Arbor,
1301 Beal Avenue, Ann Arbor, MI 48105, USA*

e-mail: `hero@umich.edu`

^{‡‡} *Instituto de Telecomunicacoes, and Instituto Superior Tecnico,
1049-001 Lisboa, PORTUGAL*

e-mail: `mario.figueiredo@lx.it.pt`

ABSTRACT

Signaling pathways are the primary means of regulating cell growth, metabolism, differentiation and apoptosis. The *de novo* signaling pathway reconstruction problem can be divided into two sub-problems: discovery of pathway components and ordering the pathway components. While the literature abounds with computational and biological approaches for discovering pathway components, there has only been limited research on ordering pathway components, despite its importance. The main biological approach, genetic epistasis analysis, is limited by the cost and unavailability of mutants. Existing computational approaches reconstruct the network from numerical data (e.g., microarray gene expression profiles) which may be unreliable. Consequently, these approaches are sensitive to data selection. Here we describe a new statistical approach to signaling network reconstruction exploiting information about the set of genes belonging to each pathway to reconstruct the “gene regulation network topology” in the form of a first-order Markov chain transition matrix. The approach naturally integrates information from multiple data sources such as text literature and biological expert knowledge and is not limited to the numerical or categorical data used by previous approaches. We demonstrate the advantages of this approach over previous approaches by reconstructing three well known signaling pathways from information reported in public cell signaling papers and databases.

Keywords: Signaling Pathway, Network, Microarray, Proteomics, EM Algorithm, Markov chain, importance sampling

1.1 Introduction

In this chapter, we focus on estimating the order of genes along a pathway assuming the terminals and unordered intermediate pathway components are known. Signaling pathways are the primary means of regulating cell growth, metabolism, differentiation, and apoptosis. The sensing and processing of extracellular stimuli are mediated by *signal transduction cascades*: molecular circuits that seek to detect, amplify, and integrate biochemical signals to generate responses such as changes in enzyme activity, activation/deactivation of transcription factors, gene expression, or ion-channel activity [1]. Biochemically, the extracellular signal is transmitted through a series of molecular modifications (e.g. phosphorylation, dephosphorylation, acetylation, methylation) and interactions (e.g. protein-protein interaction, protein-DNA interaction).

Recent bioinformatics research efforts have shifted from single gene or pairwise gene analysis to signaling pathway and network analysis. With the evolution of signaling pathway discovery methods, the definition of such pathways has evolved. In earlier decades, when genetic epistatic experiments were the predominant approach to reconstructing signaling pathways, a signaling pathway was defined as

“the cascade of processes by which an extracellular signal (typically a hormone or neurotransmitter) interacts with a receptor at the cell surface, causing a change in the level of a second messenger for example calcium or cyclic AMP, and ultimately effects a change in the cells functioning” [1]. In the post-genomic era, simultaneously quantifying the abundance levels of thousands of biomolecules enables “high throughput” signaling pathway reconstruction. Lu et al. define a signaling pathway as a specified group of genes that have coordinated association with a phenotype of interest [2]. Subramanian et al. give a more general definition of signaling pathways as groups of genes that share common biological function, chromosomal location, or regulation [3]. Subramanian’s approach looks at a hypothetical set of genes and detects significant enrichment toward the top of a rank-ordered list. Both of these studies give the analyst the power to solve the first sub-problem in signaling pathway reconstruction: the discovery of pathway components. However, in the past, epistatic relationships among pathway components have been ignored. These relationships are the key to understanding the underlying biological mechanism of gene interactions. In this chapter we describe a framework for addressing the second signaling pathway reconstruction sub-problem – that of ordering the pathway components. We propose a new definition of signaling pathway as *a series of gene interactions that lead to an endpoint biological function from a membrane receptor*.

Many biological and/or computational approaches to discovering signaling pathway components have been proposed. Biological approaches include traditional low throughput protein-protein interaction analysis such as immunoprecipitation, western blot and pull-down assay and high throughput protein-protein interaction analysis such as yeast two-hybrid assay. Computational approaches mainly focus on clustering genes according to function. Examples include network constrained clustering [4] and other methods such as hierarchical clustering [5], *K*-means type clustering [6], Model-based clustering [7][8][9]. These analyses have led to discovery of many signaling pathway components. The ultimate goal of pathway reconstruction analysis is to decipher the order in which the signal is transmitted. However, despite its importance, there has only been limited research on estimating the order of pathway components.

In the classical approach to pathway discovery, called genetic epistasis analysis, a pair of genes are mutated in the same strain and the phenotype of the double mutant is compared with those of the corresponding single mutants. The predominant phenotype defines the epistatic relationship between genes [10]. The success of this approach is contingent on the measured phenotype, and therefore, the analysis of different pathways requires a large variety of experiments. For example, satisfactory answers to the following questions are prerequisites to effective epistasis analysis: *What kind of phenotype should be measured? How to quantify this phenotype (e.g., morphology)?* In addition, as pointed out by Van Driessche et al. [11], “the rules of epistasis cannot be applied consistently if the experimental procedures are not identical for all pairs of genes in a certain pathway.”

In recent work, Van Driessche et al. [11] propose a new epistasis analysis using microarray gene expression profiles as a more objective phenotype. Their approach greatly relaxes the stringent requirements of classical epistasis analysis because the knowledge of relationships between gene function and phenotype is not essential. They reconstruct part of the Protein Kinase A Pathway by making ten combinations of single or double mutations in six genes. The approach is limited to reconstructing very small size pathways due to the combinatorial explosion of the number of mutations needed. Additional mutations are either prohibited by cost or by possibly lethal effects. In addition, the approach of [11] implicitly requires that the mutations induce significant gene expression variation so that the epistatic relationship can be determined without requiring replicated experiments.

In the last decade we have witnessed a rapid accumulation of high throughput genomic data. However, techniques for reliable knowledge extraction from this data are lacking. Instead of acquiring new data, Liu and Zhao propose a purely computational approach to reconstructing the order of pathway components from existing genomic and proteomic data [12]. Assuming all terminal and intermediate components (unordered) are known, each permutation of the pathway components is scored according to the sum of a function based on gene expression data and a separate function based on protein-protein interaction data. The gene expression score tests whether the correlation between adjacent gene pairs is significantly higher than random gene pairs in the pathway using a hypergeometric distribution model [12]. The protein-protein interaction score is based on a binomial model for interaction, or not, of adjacent proteins. The parameter (false negative rate) was estimated from protein-protein interactions in the Database of Interacting Proteins (DIP). Using the simplified Mitogen Activated Protein Kinase (MAPK) pathway as an example, they report that the “known” MAPK pathway has the second highest score among all the possible pathway permutations, which is much better than that obtainable using genomic data or proteomic data alone.

Being probably the first pure computational approach of its kind, the advantage of Liu and Zhao’s approach is that it exploits existing data. The approach also provides compelling evidence for the advantages of integrating multiple data sources. However, their approach also has a number of limitations:

- It heavily relies on the availability of high throughput data.
- It integrates only numeric data sources. Many kinds of non-numerical meta information, e.g. published literature, public databases, and biologist’s expert knowledge, are difficult to include in their probability model.
- Similar to the classical epistasis analysis, the approach is limited to reconstructing nonlethal signaling pathways.
- The approach is also limited to short pathways due to the computational complexity introduced by the permutation search in their algorithm.

Here we describe a new maximum likelihood approach that exploits information about which genes are in each pathway to reconstruct a “signal transduction network topology” in the form of a first-order Markov chain transition matrix. A graphical depiction of the system is shown in Figure 1.1. Information on the genes composing a pathway can be integrated from multiple data sources (solid errors in Fig. 1.1). The network displayed at bottom of Fig. 1.1 is reconstructed by estimating the directed pairwise interaction probabilities between components (proteins) of the network. A Markov model of is used to express these probabilities as a state transition probability matrix. Non-zero entries of this matrix corresponds to edges, i.e. interactions, between a pair of components. The proposed technique naturally combines pathway information (both composition information and epistasis information) that are derived from multiple data sources. The Markov model estimation technique, termed *network inference from co-occurrences* (NICO), was originally developed by Rabbat et al. [13] for tomographic reconstruction of telecommunications networks from untimed packets arrivals at monitored links of the network. We refer the interested reader to [13] for the complete technical details.

To summarize, the features of our proposed techniques are:

- The unordered pathway composition information can be either integrated from high throughput experiments or from meta-information, as shown in Fig. 1.1.
- Prior information on pathway epistasis can be easily integrated into the first-order Markov model in the form of a prior distribution on the transition matrix. For example, we can easily take advantage of well-known biological constraints such as the fact that kinase and phosphatase appear upstream of their substrate in the pathways.
- Our approach can be scaled to large pathways using Monte Carlo importance sampling methods.

It is often the case that available pathway composition information and prior epistatic information are not sufficient to resolve ambiguous epistasis relationships among a subset of genes. Our method also provides a measure of confidence for each potential ordering of genes in a pathway, and these confidences can be used to suggest future experiments to the biologist to resolve the ambiguity. For example, several pathway orderings may have the same confidence as measured by the likelihood score. Comparing these equally likely candidate pathways may allow biologists to identify the non-redundant set of genetic experiments which can resolve the ambiguity (represented by dotted curves in Fig. 1.1). In this sense, the proposed technique may be incorporated into a sequential design of experiments, possibly resulting in significant savings in experimental effort.

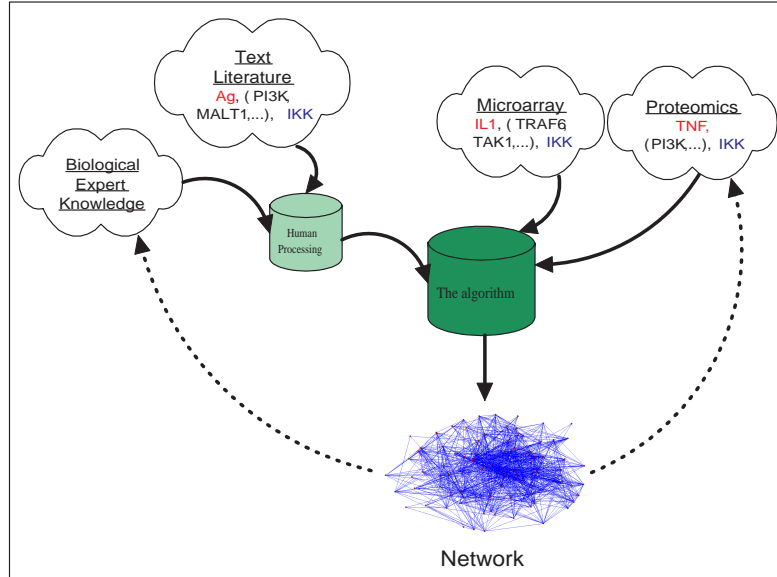


Figure 1.1: A schematic representation of the signaling pathway reconstruction algorithm. The starting pathway component is in red (left), and the ending pathway component is in blue (right). Pathway components in the parenthesis are intermediate and unordered. The solid lines represent the inputs to the algorithm (different sources of pathway information). The dotted lines represent the outputs from the algorithm (the maximum likelihood pathway(s)).

1.2 Methods

1.2.1 Mathematical Formulation of the Problem

This section outlines the *network inference from co-occurrences* (NICO) framework for inferring network structure from incomplete observations. We assume the true biological signaling pathway is an ordered path $\mathbf{z} = (z_1, z_2, \dots, z_N)$ where each z_i corresponds to a signaling protein. Mathematically, we model each ordered path as a sample from a first-order Markov chain. As described above, most existing gene expression analysis techniques identify clusters of pathway components without any information about their order within the pathway. To account for the fact that we do not directly observe order information, we model an observation \mathbf{y} as a sample \mathbf{z} from the Markov chain, subjected to a random permutation $\boldsymbol{\tau}$, so that $z_t = y_{\tau_t}$. Thus, the permutation “shuffles” the elements of each pathway, obscuring the ordering. We refer to such shuffled observations as *co-occurrences* because the observation \mathbf{x} reflects which signalling proteins “occur” in the pathway, without any order information.

The true signaling network which we are trying to elucidate consists of an ensemble of signaling pathways which can be viewed as a collection $\mathcal{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(T)}\}$ of T independent samples of the Markov chain. The Markov chain is parameterized by an initial state distribution, $\boldsymbol{\pi}$, and a transition matrix, \mathbf{A} . Each initial state distribution parameter $\pi_i = P[z_1 = i]$ for $i \in S$, where $S = \{1, \dots, |S|\}$ is the set of distinct pathway components (the states of the Markov chain) indexed using the natural numbers and $|S|$ is the number of components in the signalling network. Similarly, the transition matrix parameters are $A_{i,j} = P[z_t = j | z_{t-1} = i]$ for $i, j \in S$. These parameters must satisfy the constraints

$$\sum_{i \in S} \pi_i = 1 \quad \text{and} \quad \sum_{j \in S} A_{i,j} = 1. \quad (1.1)$$

A collection of co-occurrences, \mathcal{Y} may be obtained from multiple data sources; *e.g.*, cluster analysis of high throughput data, text literature mining, or biological expert knowledge. To recover the signaling network topology from \mathcal{Y} , we treat the corresponding unobserved permutations, $\{\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}, \dots, \boldsymbol{\tau}^{(T)}\}$ as hidden variables and describe an *expectation-maximization* (EM) algorithm for computing maximum likelihood estimates of the Markov chain parameters. This approach, NICO, was initially developed for solving a similar problem of inferring the topology of a telecommunications network from co-occurrences [13]. In section 1.2.2 we introduce further notation and review the standard approach to estimating parameters of a Markov chain when fully ordered pathways are available. In section 1.2.3 we present the EM algorithm for estimating Markov chain parameters from unordered pathways. For relatively large pathways, we describe a Monte Carlo E-step, based on importance sampling, that approximates the E-step computation (section 1.2.4). Finally, we discuss how to incorporate prior pathway information (section 1.2.5).

1.2.2 Estimating a Markov Chain from Direct Observations

The sections 1.2.2, 1.2.3, 1.2.4, 1.2.5 are adapted in large part from [13], to which we refer the reader for further technical details and bibliographic references on Markov chain models, the EM algorithm, and importance sampling. Our goal is to estimate the Markov chain parameters $\boldsymbol{\pi}$ and \mathbf{A} . It is convenient to introduce an alternative representation for an ordered pathway: instead of $\mathbf{z} = (z_1, z_2, \dots, z_N)$ we write $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$ where each \mathbf{w}_i is a length- $|S|$ binary vector such that $(w_{i,j} = 1) \Leftrightarrow (z_i = j)$. With this representation, we can write the log-likelihood of an observation \mathbf{w} as

$$\log P[\mathbf{w}|\mathbf{A}, \boldsymbol{\pi}] = \sum_{i \in S} w_{1,i} \log \pi_i + \sum_{t=2}^N \sum_{i,j \in S} w_{t-1,i} w_{t,j} \log A_{i,j}. \quad (1.2)$$

Note that most of the terms in these sums are zero, because, by construction, most of the $w_{i,j}$ are zero.

Next, suppose we observe a collection of independent *ordered* pathways $\mathcal{W} = \{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(T)}\}$, where $\mathbf{w}^{(m)}$ corresponds to a pathway of N_m elements. Maximizing the log-likelihood $\log P[\mathcal{W}|\mathbf{A}, \boldsymbol{\pi}] = \sum_{m=1}^T \log P[\mathbf{w}^{(m)}|\mathbf{A}, \boldsymbol{\pi}]$ under the constraints in Eqn. 1.1 leads to estimates

$$\hat{A}_{i,j} = \frac{\sum_{m=1}^T \sum_{t=2}^{N_m} w_{t-1,i}^{(m)} w_{t,j}^{(m)}}{\sum_{j \in S} \sum_{m=1}^T \sum_{t=2}^{N_m} w_{t-1,i}^{(m)} w_{t,j}^{(m)}} \quad (1.3)$$

$$\hat{\pi}_i = \frac{1}{T} \sum_{m=1}^T w_{1,i}^{(m)}. \quad (1.4)$$

1.2.3 Estimating a Markov Chain from Shuffled Observations via the EM Algorithm

Now, we would like to compute maximum likelihood estimates of the Markov chain transition matrix \mathbf{A} and initial state distribution $\boldsymbol{\pi}$ from a collection of unordered co-occurrences $\mathcal{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(T)}\}$. Corresponding to each co-occurrence $\mathbf{y}^{(m)}$ is a permutation $\boldsymbol{\tau}^{(m)}$. Again, it is convenient to introduce binary representations for co-occurrences and permutations. Similar to before, instead of the co-occurrence $\mathbf{y} = (y_1, y_2, \dots, y_N)$ we write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ where each \mathbf{x}_i is a binary vector such that $(x_{i,j} = 1) \Leftrightarrow (y_i = j)$. Also, we represent each shuffling $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_N)$ in terms of a N -by- N permutation matrix \mathbf{r} so that $(r_{t,t'} = 1) \Leftrightarrow (\tau_t = t')$. With this notation the binary co-occurrence \mathbf{x} and ordered path \mathbf{w} are related via the permutation matrix \mathbf{r} via the expression

$$w_{t,i} = \prod_{t'=1}^N (x_{t',i})^{r_{t,t'}}, \quad (1.5)$$

where we adopt the convention $0^0 = 1$.

If, in addition to the collection co-occurrences $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)}\}$, we were also given the corresponding permutation matrices $\mathcal{R} = \{\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \dots, \mathbf{r}^{(T)}\}$ then we could undo the shufflings and apply the direct maximum likelihood procedure outlined in the previous section. Since, in practice, we do not know the corresponding permutations, we treat them as hidden variables and derive an EM algorithm, modelling each permutation as being drawn from an equally likely distribution on all permutations of the appropriate length; *i.e.*, if $\mathbf{x}^{(m)}$ corresponds to a path of N_m elements then $\mathbf{r}^{(m)}$ is modelled as a random permutation matrix drawn uniformly from the collection of all permutations of N_m elements, denoted by Ψ_{N_m} .

The EM algorithm alternates between the *expectation* or E-step, which amounts to estimating expected permutations for each path conditioned on the current parameter estimates, and the *maximization* or M-step, where the parameter estimates are updated based on the expected permutations computed in the E-step. More precisely, in the E-step we compute sufficient statistics

$$\bar{\alpha}_{t',t''}^{(m)} = E \left[\sum_{t=2}^{N_m} r_{t,t'} r_{t-1,t''} \middle| \mathbf{x}^{(m)}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}} \right] \quad (1.6)$$

$$= \frac{\sum_{\mathbf{r} \in \Psi_{N_m}} r_{t,t'} r_{t-1,t''} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]}{\sum_{\mathbf{r} \in \Psi_{N_m}} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]}, \quad (1.7)$$

and

$$\bar{r}_{1,t'}^{(m)} = E \left[r_{1,t'} \middle| \mathbf{x}^{(m)}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}} \right] \quad (1.8)$$

$$= \frac{\sum_{\mathbf{r} \in \Psi_{N_m}} r_{1,t'} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]}{\sum_{\mathbf{r} \in \Psi_{N_m}} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]}, \quad (1.9)$$

where each term $P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]$ is easily computed after using \mathbf{r} to unshuffle $\mathbf{x}^{(m)}$:

$$P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}] = P[\mathbf{y}^{(m)} | \boldsymbol{\tau}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}] = \hat{\pi}_{y_{\tau_1}^{(m)}} \prod_{t=2}^{N_m} \hat{A}_{y_{\tau_{t-1}}^{(m)}, y_{\tau_t}^{(m)}}. \quad (1.10)$$

Closed form expressions for the M-step updates are then given by

$$\left(\hat{A}_{i,j} \right)_{\text{new}} = \frac{\sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)}}{\sum_{j=1}^{|S|} \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)}} \quad (1.11)$$

and

$$\left(\bar{\pi}_i \right)_{\text{new}} = \frac{\sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}{\sum_{i=1}^{|S|} \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}. \quad (1.12)$$

This procedure, along with the Monte Carlo and MAP variations described below, are easily modified to handle the case when the endpoints of each pathway are known, *e.g.*, when membrane receptors and transcription factors are known.

1.2.4 Monte Carlo E-Step by Important Sampling

For a large pathway, the combinatorial nature of the equations (1.7) and (1.9), the summation over all permutations in the pathway, may render exact computation impractical. We describe an importance sampling-based approximation version of the E-step which avoids this issue. Without loss of generality, we focus on a particular length- N co-occurrence $\mathbf{y} = (y_1, y_2, \dots, y_N)$, dropping the superscript (m) to lighten the notation. We also drop the hats from $(\hat{\mathbf{A}}, \hat{\boldsymbol{\pi}})$ and use simply $(\mathbf{A}, \boldsymbol{\pi})$ to denote the current Markov chain parameter estimates in the EM algorithm.

Intuitively, there are a large number of permutations, but typically only a few of these permutations will contribute a non-negligible conditional probability $P[\mathbf{t}|\boldsymbol{\tau}, \mathbf{A}, \boldsymbol{\pi}]$. The idea behind importance sampling is to sample high probability permutations based on the current parameter estimates, and then combine these in a statistically sound fashion.

The importance sampling procedure we propose sequentially samples a permutation in the following fashion, making intuitive use of the current parameter estimates, \mathbf{A} and $\boldsymbol{\pi}$ and the co-occurrence \mathbf{y} .

Step 0: Let $\mathcal{U} = \{1, 2, \dots, N\}$ denote the set of elements remaining to be placed in the permutation.

Step 1: For each $j \in \mathcal{U}$ set $p_j \propto \pi_j$, normalized so that $\sum_{j \in \mathcal{U}} p_j = 1$. Sample an element j from \mathcal{U} according to the distribution $\{p_j\}$ just defined. Set $\tau_1 = j$ and update $\mathcal{U} \leftarrow \mathcal{U} \setminus j$. Let $i = 2$ denote the next position to be filled in the permutation.

Step 2: For each remaining $j \in \mathcal{U}$ set $p_j \propto A_{\tau_{i-1}, j}$, normalized so that $\sum_{j \in \mathcal{U}} p_j = 1$. Sample an element j from \mathcal{U} according to the distribution $\{p_j\}$ just defined. Set $\tau_i = j$. Update $\mathcal{U} \leftarrow \mathcal{U} \setminus j$ and $i \leftarrow i + 1$.

Step 3: Repeat step 2 until $i = N + 1$, at which the entire permutation has been sampled.

Repeating this procedure L times yields L independent sample permutations, $\boldsymbol{\tau}^1, \boldsymbol{\tau}^2, \dots, \boldsymbol{\tau}^L$, or in equivalent binary notation, $\mathbf{r}^1, \dots, \mathbf{r}^L$. Associate the reweighting factor

$$z_\ell = \prod_{j=2}^N \sum_{k=j}^N A_{y_{\tau_{j-1}^\ell}, y_{\tau_k^\ell}} \quad (1.13)$$

with the ℓ th permutation. Then the Monte Carlo E-step approximations are given

by

$$\hat{\alpha}_{t',t''} = \frac{\sum_{\ell=1}^L \sum_{t=2}^N r_{t-1,t'}^\ell r_{t,t''}^\ell z_\ell}{\sum_{\ell=1}^L \sum_{t=2}^N z_\ell}, \quad (1.14)$$

$$\hat{r}_{1,t'} = \frac{\sum_{\ell=1}^L r_{1,t'}^\ell z_\ell}{\sum_{\ell=1}^L z_\ell}. \quad (1.15)$$

Computing approximations in this fashion, with the reweighting terms, ensures that the approximations $\hat{\alpha}_{t',t''}$ converge to the exact values $\bar{\alpha}_{t',t''}$ almost surely as $L \rightarrow \infty$. In fact, it can be shown that desirable properties of the EM algorithm are preserved when $L \propto N^4$; *i.e.*, when a polynomial complexity approximation scheme is used, in contrast to exponential complexity required to perform the exact E-step computation. See [13] for further details.

1.2.5 Incorporating Prior Information

Prior information about the Markov chain parameters \mathbf{A} and $\boldsymbol{\pi}$ can easily be incorporated into the algorithm by applying independent Dirichlet priors to each row of the transition matrix and to the initial state distribution. Hence, we have

$$P[\boldsymbol{\pi}|\mathbf{u}] \propto \prod_{i=1}^{|S|} \pi_i^{u_i-1} \quad (1.16)$$

$$P[\mathbf{A}|\mathbf{v}] \propto \prod_{i=1}^{|S|} \prod_{j=1}^{|S|} A_{i,j}^{v_{i,j}-1}, \quad (1.17)$$

where the parameter u_i and $v_{i,j}$ should be non-negative in order to have proper priors. The larger that u_i is relative to the other $u_{i'}$, $i' \neq i$, the greater our prior belief that pathway component i is a starting component of the pathway rather than the others. Similarly, the larger $v_{i,j}$ relative to other $v_{i,j'}$ for $j' \neq j$, the more likely we expect that, *a priori*, the signal is transmitted from pathway component i to pathway component j relative to the transmissions from i to the other pathway components.

Incorporating these priors into our model only results in a change to the M-step of the EM algorithm. Instead of equations (1.11) and (1.12), which lead to maximum likelihood estimates, we have

$$(\hat{\pi}_i)_{\text{new}} = \frac{u_i + \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}{\sum_{i=1}^{|S|} \left(u_i + \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)} \right)}, \quad (1.18)$$

and

$$(\hat{A}_{i,j})_{\text{new}} = \frac{v_{i,j} + \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)}}{\sum_{j=1}^{|S|} \left(v_{i,j} + \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)} \right)}, \quad (1.19)$$

leading to maximum *a posteriori* (MAP) parameter estimates.

1.3 Results

Using three representative signaling pathways, we intend to show three useful properties of our approach: reconstruction of the order of genes in the pathway assuming the intermediate and terminal components are known; ease of incorporating prior knowledge in the form of a prior on the transition matrix; identifying the most important missing information that could improve confidence in the path order reconstruction. The latter will be useful for specifying the most informative future experiment if needed (Fig. 1.1).

1.3.1 Protein Kinase A Pathway

The protein kinase A (PKA) pathway is an essential signaling pathway for cell development. The central component, cyclic AMP (cAMP)-dependent protein kinase A is able to phosphorylate a variety of proteins and thereby effect their activity. Malfunction of this pathway leads to developmental arrest or attenuation, precocious development and aberrant sporulation and germination [14], [11]. Van Driessche et al. use this pathway to demonstrate a microarray based epistasis approach [11]. They reconstruct an incomplete pathway by making ten combinations of single or double mutations in six genes. However, the relationships between several pairs of genes cannot be determined from such an analysis. For example, the level of interaction between *acaA* and *pkaR* is not tested because the corresponding mutations are not analyzed or are difficult to make. Despite this missing information, our approach is able to reconstruct the reported pathway based only on the information about terminal components and the unordered intermediate components in each pathway (Fig. 1.2, Fig. 1.3). This suggests that our techniques may enable biologists to reconstruct pathways without having to perform exhaustive experiments on all pairwise interactions.

Since the protein kinase A pathway is a relatively small pathway, it is perhaps not surprising that we are able to reconstruct it in a straightforward manner. For larger pathways, available prior pathway composition information often only allows the pathway be reconstructed up to a certain “low resolution” i.e., up to certain ambiguities in relative ordering within the pathway. Incorporating prior knowledge can often help to reveal the order of the whole pathway or an ensemble of pathways; i.e., a signaling network. In the next subsection we illustrate our methods on the more complicated SAPK/JNK pathway.

1.3.2 SAPK/JNK Pathway

Stress-activated protein kinases (SAPK)/Jun N-terminal kinases (JNK) are members of the MAPK family and are activated by a variety of environmental stresses,

PKA Pathway

acaA, (pkaC, pkaR), Development

regA, (pkaR, pkaC), Development

yakA, (pufA, pkaC), Development

Figure 1.2: The (unordered) protein kinase A signaling pathway information that is used by the NICO algorithm to reconstruct the ordered pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. The pathway is mainly adapted from Van Driessche et al. [11].

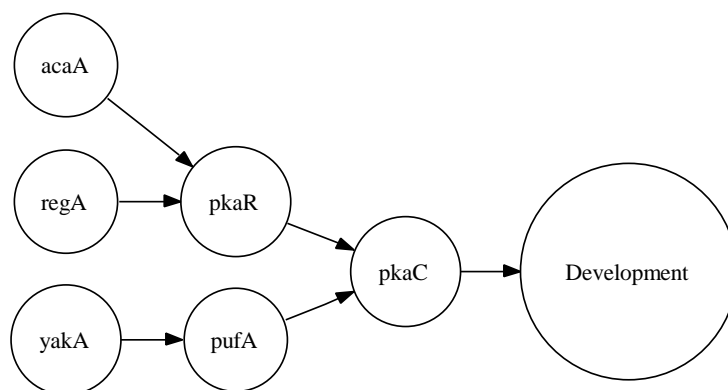


Figure 1.3: The protein kinase A signaling network topology reconstructed from unordered pathway composition data using the methodology outlined in Sec. 1.2 (Fig. 1.2).

SAPK/JNK Pathway
GF, (HPK, MEKK, MKK), **JNK**
GF, (EKK, HPK, MKK), **JNK**
GF, (RAC, RAS, MEKK, MKK), **JNK**
GF, (RAS, CDC42, RAC, MKK, MEKK), **JNK**
GF, (RAS, RAC), **RHO**
CS1, (RAC, MEKK, MKK, CDC42), **JNK**
CS2, (MEKK, MKK, RAC), **JNK**
FASL, (GCKs, MKK, MEKK), **JNK**
OS, (ASK1, MEKK, MKK), **JNK**

Figure 1.4: The (unordered) SAPK/JNK signaling pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “GF” stands for Growth Factor, “CS” stands for Cellular Stress, “FASL stands for Fas Ligand”, “OS” stands for Oxidation Stress. The pathway is adapted from <http://www.cellsignal.com/>.

inflammatory cytokines, growth factors and GPCR agonists. Stress signals are delivered to this cascade by small GTPases of the Rho family (Rac, Rho, Cdc42) [15]. Similar to our study of the protein kinase A pathway, we attempt to reconstruct pathway order based only on the terminal components and on unordered lists of co-occurring intermediate pathway components (Fig. 1.4).

In the NICO framework, epistasis relationships of the pathway components are fully defined by the probability transition matrix \mathbf{A} . For the observed unordered SAPK/JNK pathway, there may be multiple maxima in the likelihood function. For example, the two estimates of \mathbf{A} in Eq. 1.20 and Eq. 1.21 corresponding to two possible epistasis relationships between MEKK and MKK that are equally likely. The ordered row names are: “GF”, “RAS”, “CDC42”, “MEKK”, “MKK”, “JNK”, “RAC”, “Rho”, “HPK”, “CS1”, “CS2”, “FASL”, “GCKs”, “OS”, “ASK1”. All-zero rows correspond to the end-of-pathway components “JNK” and “RHO” (these terminals do not emit signals), and probabilities in non-zero rows sum up to 1. We incorporated prior information that the MEKK protein phosphorylates the MEK protein [15] by setting parameter $v_{4,5} = 1$ in the Dirichlet prior $p[\mathbf{A}|\mathbf{v}]$ on the transition matrix \mathbf{A} . Recall that the larger we make this probability value, the more confidence we have in prior belief. With this prior, the algorithm reconstructs the whole pathway correctly (Fig. 1.5).

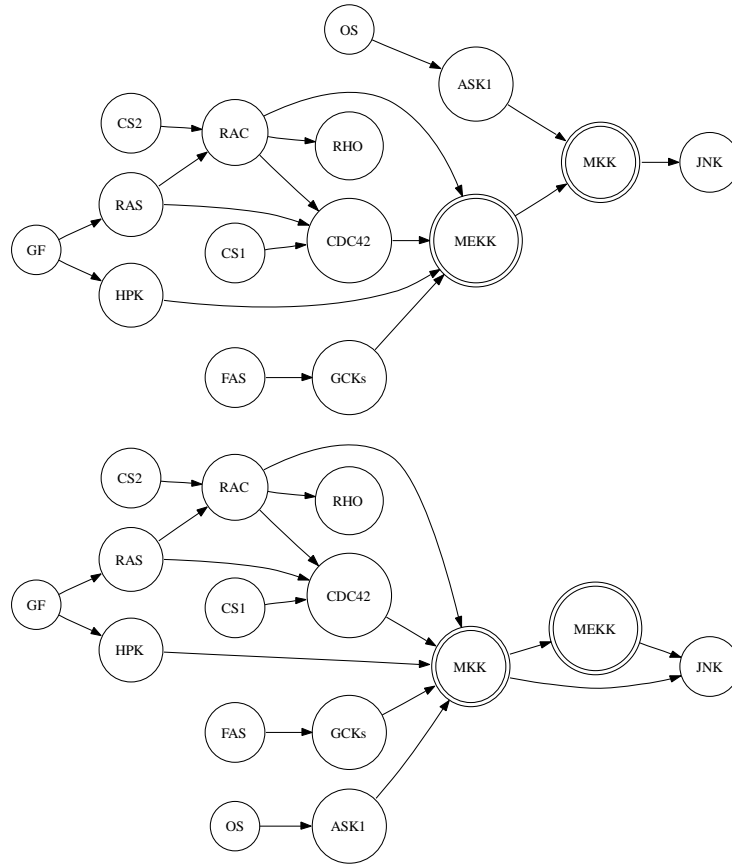


Figure 1.5: Upper panel: The correct SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 1.20 estimated from unordered pathway composition data (Fig. 1.4) improved by incorporating a prior information on gene-gene interactions, in particular the interactions between the two double-circled components. Lower panel: The incorrect SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 1.21 estimated from unordered pathway composition data without incorporating prior information.

$$\hat{\mathbf{A}} = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.333 & 0 & 0 & 0.667 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (1.20)$$

$$\hat{\mathbf{A}}' = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.333 & 0 & 0.667 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.875 & 0 & 0.125 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (1.21)$$

Often prior epistasis information or pathway composition information may not suffice to resolve all ordering ambiguity in the pathway. In such cases it would be useful to predict the crucial pieces of information necessary to resolve remaining ambiguity. We next show how our method can be applied to perform such a prediction for the Nuclear Factor κ B (NF κ B) pathway.

1.3.3 NF κ B Pathway

NF κ B proteins function as dimeric transcription factors that control genes regulating a broad range of biological processes including innate and adaptive immunity, inflammation, apoptosis, stress responses, B cell development and lymphoid

organogenesis [16]. $\text{NF}\kappa\text{B}$ pathways mediate the signal transduction from extracellular stimuli to these transcription factors including controlled cytoplasmic-nuclear shuttling and modulation of transcriptional activity [17].

We specified the terminal components of different stimuli receptors (start) and $\text{NF}\kappa\text{B}$ (end), and pathway components corresponding to each stimuli (Fig. 1.6). The latter can often be derived from a combination of computational approaches (e.g. clustering) and the biologist’s expert knowledge. Biological expert knowledge is acquired gradually over years from multiple sources such as the literature, science seminars, and experimental results. We also incorporated several pieces of prior biological information including the epistasis relationships between PI(3)K and PLC γ 2 [18], between PLC γ 2 and PKC [18], between PKC and MALT1 [19], between MALT1 and TRAF6 (TNF-receptor-associated factor 6) [20], between TRAF6 and TAK1 (TGF β -activated kinase 1) [21], between TAK1 and IKK [20], between PI(3)K and Akt/Cot complex [22], and between JNK and β TrCP (β Transducin Repeat-Containing Protein) [23]. The biology background is as follows: Upon PI(3)K activation the Akt/Cot complex is likely recruited to the membrane through the Akt PH domain, which binds the phospholipid PIP3 [22]. JNK induces β TrCP to activate $\text{NF}\kappa\text{B}$ pathway [23]. Tyrosine phosphorylation of phospholipase PLC γ 2 is a crucial activation switch that initiates and maintains intracellular calcium mobilization in response to extracellular stimuli [18]. PKC was reported to be able to activate MALT1 upon receiving extracellular stimuli [19]. MALT1 binds and activates TRAF6 [20]. TRAF6 activates TAK1 through the adaptor protein TAB2 [21] and TAK1 activates IKK [20].

The NICO algorithm successfully reconstructs most of the pathway component orders [18-26], with the sole ambiguity being between $\text{NF}\kappa\text{B}$ complex1 and complex2 (Fig. 1.7). Indeed, in this case the ambiguity can be detected by investigating the relative maxima of the likelihood function, $P[\mathcal{X}|\hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}]$. A relative maximum that is approximately equal to the global maximum indicates an ambiguity that is localized by the positions of the relative maxima over the space of transition matrices \mathbf{A} (Eqs. 1.22, 1.23, Appendix. A). To resolve this ambiguity, our analysis indicates that biologists should focus on investigating the epistatic relationship between these two complexes.

1.3.4 Assembling Signaling Pathways into Signaling Networks

Biological signaling pathways tend to share a fair amount of common signal components, and we often define these as signaling networks. The latter provides a more complete view of cellular regulatory mechanisms. Fig. 1.8 presents a signaling network assembled from SNK/JNK and $\text{NF}\kappa\text{B}$ pathways.

NF κ B Pathway

Ag, (PKC, PI(3)K, PLC γ 2, MALTI, TAK.TAB1/2, IKK, TRAF6, NF κ BC2, NF κ BC1), **NF κ B**
Ag.MHC, (TRAF6, PLC γ 2, MALTI, TAK.TAB1/2, PKC, IKK, NF κ BC1, NF κ BC2), **NF κ B**
IL-1, (IKK, TRAF6, TAK.TAB1/2, NF κ BC1, NF κ BC2), **NF κ B**
dsRNA, (NF κ BC1, PKR, IKK, NF κ BC2), **NF κ B**
TNF, (IKK, MEKK, NF κ BC1, NF κ BC2), **NF κ B**
GF, (AKT.COT, IKK, PI(3)K, NF κ BC2, NF κ BC1), **NF κ B**
LT, (PI(3)K, IKK, NF κ BC1, NF κ BC2, AKT.COT), **NF κ B**
LT, (IKK, NIK, NF κ BC1, NF κ BC2), **NF κ B**
UV, (bTrCP, NF κ BC1, NF κ BC2, JNK), **NF κ B**

Figure 1.6: The (unordered) NF κ B signaling pathways. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “Ag” stands for Antigen, “Ag-MHC” stands for Major Histocompatibility Complex (MHC) Antigen, “IL-1” stands for Interleukemia-1, “dsRNA” stands for double stranded RNA, TNF stands for Tumor Necrosis Factor, “GF” stands for Growth Factor, “LT” stands for heat-labile enterotoxin. “NF κ BC1” and “NF κ BC2” stand for NF κ B complexes 1 and 2. The pathway is adapted from <http://www.cellsignal.com/>.

1.4 Discussion

In this chapter, we presented a model based approach to inferring the order of an unordered list of pathway components connecting known terminal genes. Compared to previous genetic and computational approaches, our approach does not directly depend on the numeric format of the data, thus it enjoys the features of versatility, flexibility and a high level of data abstraction. The knowledge of intermediate pathway components and terminal components can be derived either from numeric data using computational/statistical methods or from meta-data using biological expertise, e.g., terminal genes of a pathway are often specified as membrane receptor (start) and transcription factor (end). In this sense, the approach represents progress in data integration for gene pathway discovery. Moreover, the MAP variant of the NICO algorithm permits seamless incorporation of prior epistatic knowledge in the form of a prior on the transition matrix. When ambiguities do exist our algorithm can identify them and provide information on the most fruitful set of future experiments needed to resolve the ambiguities.

Many researchers have found the topology of networks of signaling pathways to be scale-free and sparse. In such topologies a small number of nodes (hub nodes) are highly connected while the remaining nodes are not. The hub nodes may form interaction motifs (functional modules) that are often shared by multiple pathways. Our

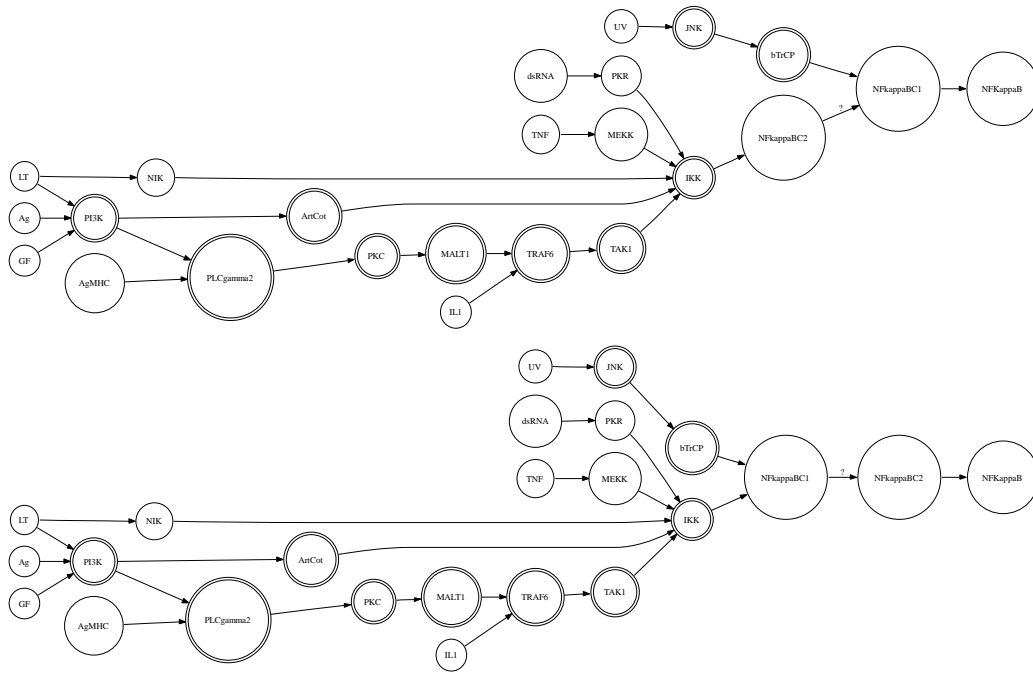


Figure 1.7: The two possible NF κ B signaling network topologies defined by the probability transition matrix Eq.A.16 and Eq.A.17 estimated from unordered pathway composition data (Fig. 1.6) after incorporating prior information. The relationships between the two double-circled components are disambiguated from prior information. The epistasis relationship labeled with “?” remains ambiguous and deserves further investigation.

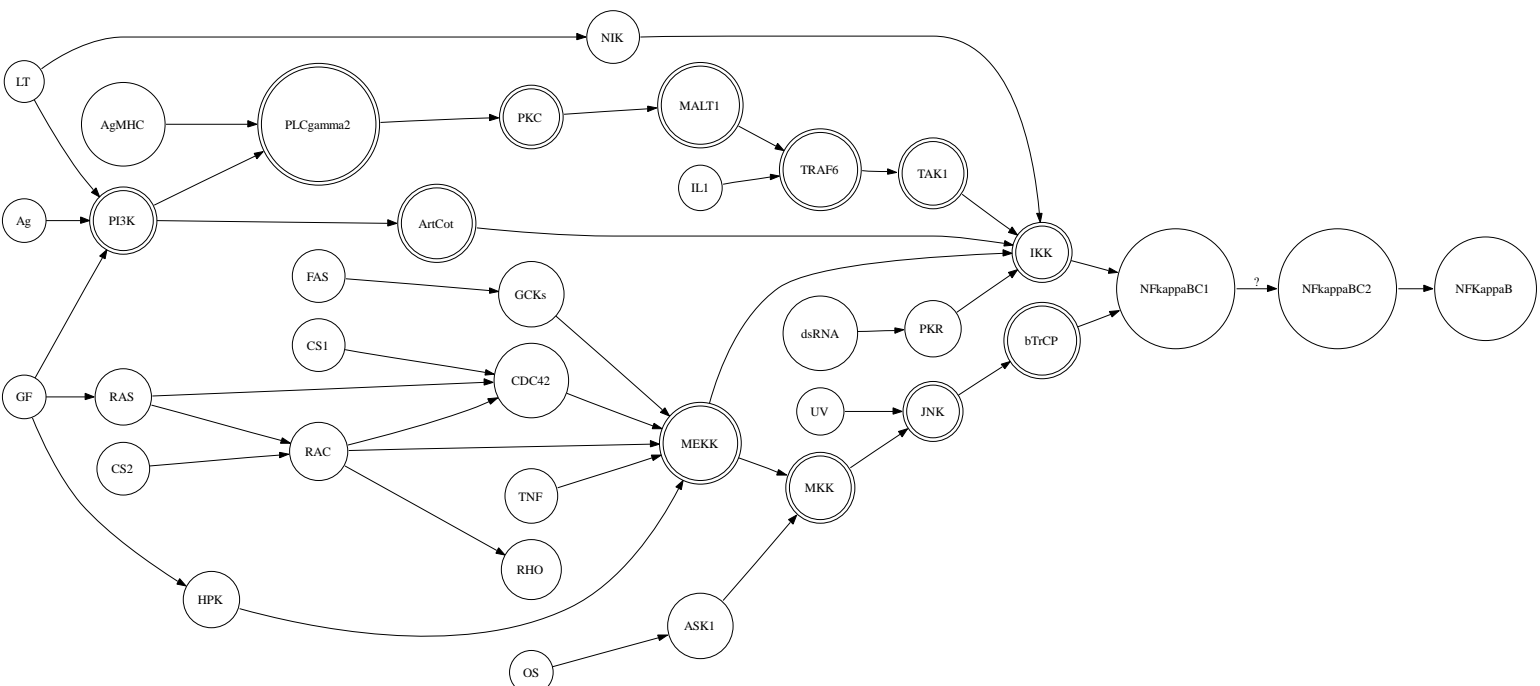


Figure 1.8: The signaling networks assembled from SNK/JNK and NF-κB pathways.

Bibliography

- [1] Berg, J.M., Tymoczko, J.L. and Stryer, L. *Biochemistry*, W. H. Freeman, New York, 2006.
- [2] Lu, T., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. *Proc. Natl. Acad. Sci. USA*, **12**, 13544-13549, 2005.
- [3] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. *Proc. Natl. Acad. Sci. USA*, **102**(43), 15545-15550, 2005.
- [4] Zhu, D., Hero, A.O., Cheng, H., Khanna, R. and Swaroop, A. *Bioinformatics*, **21**(21), 4014-4021, 2005.
- [5] Eisen, M., Spellman, P., Brown, P.O. and Botstein, D. *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868, 1998.
- [6] Hartigan, J.A. and Wong, M.A. A k -means clustering algorithm. *Applied Statistics*, **28**, 100-108, 1979.
- [7] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. *Bioinformatics*, **10**, 977-987, 2001.
- [8] McLachlan, G., Bean, R. and Peel, D. *Bioinformatics*, **18**, 413-422, 2002.
- [9] Schliep, A., Schonhuth, A. and Steinhoff, C. *Bioinformatics*, **19**, i255-i263, 2003.
- [10] Avery, L. and Wasserman, S. *Trends. Genet.*, **8**, 312-316, 1992.
- [11] Van Driessche, N., Demsar, J., Booth, E.O., Hill, P., Juvan, P., Zupan, B., Kuspa, A. and Shaulsky, G. Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.*, **37**(5), 471-477, 2005.
- [12] Liu, Y., and Zhao, H. *BMC Bioinformatics*, **5**:158, 2004.
- [13] Rabbat, M.G., Figueiredo, A.T. and Nowak, R.D. University of Wisconsin - Madison Technical Report ECE-06-2, April, 2006.

- [14] Loomis, W.F. *Microbiol. Mol. Biol. Rev.*, **62**(3), 684-694, 1998
- [15] Weston, C.R. and Davis, R.J. *Curr Opin. Genet. Dev.*, **12**, 14-21, 2002.
- [16] Pomerantz, J.L. and Baltimore, D. *Mol. Cell*, **10**, 693-695, 2002.
- [17] Ghosh, S. and Karin, M. *Cell*, **109**, S81-S96, 2002.
- [18] Humphries, L.A., Dangelmaier, C., Sommer, K., Kipp, K., Kato, R.M., Griffith, N., Bakman, I., Turk, C.W., Daniel, J.L. and Rawlings, D.J. *J. Biol. Chem.*, **279**, 37651-37661, 2004.
- [19] Che, T., You, Y., Wang, D., Tanner, M.J., Dixit, V.M. and Lin, X. *J. Biol. Chem.*, **279**(16), 15870 -15876, 2004.
- [20] Sun, L., Deng, L., Ea, C.K., Xia, Z.P. and Chen, Z.J. *Cell*, **14**(3), 289-301, 2004.
- [21] Morlon, A., Munnich, A. and Smahi A. *Human Molecular Genetics*, **14**(23), 3751-3757, 2005.
- [22] Kane, L.P., Mollenauer, M.N., Xu, Z., Turck, C.W. and Weiss, A. *Mol. Cell Biol.*, **22**(16), 5962-5974, 2002.
- [23] Spiegelman, V.S., Stavropoulos, P., Latres, E., Pagano, M., Ronai, Z., Slaga, T.J. and Fuchs, S.Y. *J. Biol. Chem*, **276**(29), 27152-27158, 2001.
- [24] Hu, W.H., Pendergast, J.S., Mo, X.M., Brambilla, R., Bracchi-Ricard, V., Li, F., Walters, W.M., Blits, B., He, L., Schaal, S.M. and Bethea, J.R. *J. Biol. Chem*, **280**(32), 29233-29241, 2005.
- [25] Panta, G.R., Kaur, S., Cavin, L.G., Cortès, M.L., Mercurio, F., Lothstein, L., Sweatman, T.W., Israel, M. and Arsura, M. *Mol. Cell. Biol.*, **24**(5), 1823-1835, 2004.
- [26] Bonnet, M.C., Daurat, C., Ottone, C. and Meurs, E.F. *Cell Signal.*, 2006 Feb 28; [Epub ahead of print].
- [27] Vogelstein, B., Lane, D. and Levine, A.J. *Nature*, **408**, 307-310, 2000.