Identifying Therapeutic Targets for Joint Specific Rheumatoid Arthritis Using Bayesian Network Models

Melissa Wei, Richard Ainsworth, and Wei Wang

The Department of Chemistry & Biochemistry, University of California San Diego, 9500 Gilman drive, La Jolla, California 92037

Abstract

Rheumatoid arthritis has a diverse pathogenesis, and its complex reactions to treatments may contribute to difficulty in finding treatments. Here we searched for therapeutic targets for joint-specific rheumatoid arthritis in several gene perturbation recipes that were more effective in certain joints. We used single cell RNA seq data from rheumatoid arthritis patients to train 20 directed acyclic graphs and conducted Bayesian inference on them. In order to find the most efficient reprogramming recipes, we simulated the knockdown and overexpression of joint-specific rheumatoid arthritis state genes in the network and found the resulting correlation to the osteoarthritis state. The identification of these therapeutic targets could help treat joint-specific rheumatoid arthritis.

Introduction

Rheumatoid arthritis, an autoimmune disease, is the most common type of inflammatory arthritis. Although advances in rheumatoid arthritis have been made, the complex pathways and interactions between cell states and genes makes it particularly hard to treat. The disease is primarily symmetrical and typically affects diarthrodial joints, such as hips and knees, as the disease progresses.[1,2] Recently it was proposed that epigenetic differences in fibroblast-like synoviocytes (FLS) in synovial joints contribute to the complexity of rheumatoid arthritis and its pathogenic pathways.[3,4] This was supported by the findings of differentially methylated genes in RA knee and hip joint FLS.[5] The top 25 most differentially expressed OA genes identified from Ai et al.[6] and top 25 most differentially expressed RA knee and RA hip genes from data set were selected as a 50 gene list. These genes were selected as they are most likely to be the greater causes for joint-specific rheumatoid arthritis state genes in the network and found the resulting correlation to OA. We discretized the expression levels of each gene, shown in Figure 1a, and learned a Bayesian network model purely from quantitative data and did not incorporate prior knowledge.

Next we conducted Bayesian inference for each Directed Acyclic Graph using the Monte Carlo Markov Chain (MCMC) method using the initial probabilities P(E), where E is a given condition for each cell state. The probabilities obtained from the wild type (unperturbed state) of each of the three cell states corresponded to a potential minimum. To simulate the knockdown or overexpression of certain genes, we clamped the activity levels of these genes. Gene overexpression was simulated by clamping gene activity levels to 1, and knockdowns to 0. The influence would calculate the probabilities of the other genes present in the Perturbation Genes and the inference for each Directed Acyclic Graph using the Monte Carlo Markov Chain (MCMC) method using the initial probabilities P(E), where E is a given condition for each cell state. The probabilities obtained from the wild type (unperturbed state) of each of the three cell states corresponded to a potential minimum. To simulate the knockdown or overexpression of certain genes, we clamped the activity levels of these genes. Gene overexpression was simulated by clamping gene activity levels to 1, and knockdowns to 0. The influence would calculate the probabilities of the other genes present in the network, which is the joint-specific FLS expression level that is the joint-specific expression level that can be used to calculate the probability of each of the other FLS expression levels. Further research is needed to biologically validate the recipes, to find therapeutic targets for different FLS that are not from knees and hips, and to find more efficient recipes with the current method (for example, more than 3 genes could be perturbed as this will likely improve efficiency of the recipes), or by a different method. This may also be useful for joint-specific rheumatoid arthritis as realization of joint-specific gene recipes may treat patients with joint-specific rheumatoid arthritis.

Figure 3. Table of the top 5 knee and hip recipes, different recipes are identified for each joint, FGF10, LRP1B, and SHISA5 for knee, and PLXNC1, LRP1B, and FAM135B for hip.

We noticed that the recipes for knee and hip RA are different. Most of the common genes in the top recipes also happened to be major nodes in genetic network (Figure 2a). Expectedly, similar genes (FGF10, PLXNC1, LRP1B) are seen in both joint recipes, however the combinations of genes are quite unique for each joint.

Acknowledgments

I want to thank Dr. Komvis for directing the Research Scholars program, Dr. Wang for welcoming me into his lab, and Dr. Ainsworth for being such a great mentor.

References


DAG:

A directional acyclic graph (DAG) is a directed graph with no directed cycles. In other words, there is a directed path from any vertex to any other vertex. DAGs are often used to represent causal relationships between events. In a DAG, the nodes represent events and the directed edges represent the causal relationships between them. DAGs are often used in biological systems to represent regulatory networks, where the nodes represent genes and the edges represent regulatory interactions. For example, in the network shown in Figure 1c, the node labeled “Gene A” has an edge directed towards the node labeled “Gene B,” indicating that Gene A has a regulatory effect on Gene B.

In order to train the highest-scoring Directed Acyclic Graph (DAG) which is a necessary component for the Bayesian network, we used a greedy algorithm which maximizes the BIC (Bayesian Information Criterion) score in order to improve running complexity. BIC is calculated:

$$BIC(G) = -2 \log(P(D|G)) - \frac{d}{2} \log(n),$$

where BIC(G) is the probability of the data given the DAG, P(D|G) is the probability of the data, and d is the number of parameters.

Figure 1. Data processing was done by discirnetizing over log10 TPM (transcripts per million) on the median. Genes are considered expressed (1) or not expressed (0).

In order to train the highest-scoring Directed Acyclic Graph (DAG) which is a necessary component for the Bayesian network, we used a greedy algorithm which maximizes the BIC (Bayesian Information Criterion) score in order to improve running complexity. BIC is calculated:

$$BIC(G) = -2 \log(P(D|G)) - \frac{d}{2} \log(n),$$

where BIC(G) is the probability of the data given the DAG, P(D|G) is the probability of the data, and d is the number of parameters.

In order to train the highest-scoring Directed Acyclic Graph (DAG) which is a necessary component for the Bayesian network, we used a greedy algorithm which maximizes the BIC (Bayesian Information Criterion) score in order to improve running complexity. BIC is calculated:

$$BIC(G) = -2 \log(P(D|G)) - \frac{d}{2} \log(n),$$

where BIC(G) is the probability of the data given the DAG, P(D|G) is the probability of the data, and d is the number of parameters.

Figure 2. Table showing top edges found in the Directed Acyclic Graphs.

<table>
<thead>
<tr>
<th>Source</th>
<th>Node</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LRP1B</td>
<td>FAM135B</td>
</tr>
<tr>
<td>2</td>
<td>SHISA5</td>
<td>LRP1B</td>
</tr>
<tr>
<td>3</td>
<td>LRP1B</td>
<td>FAM135B</td>
</tr>
<tr>
<td>4</td>
<td>SHISA5</td>
<td>LRP1B</td>
</tr>
<tr>
<td>5</td>
<td>LRP1B</td>
<td>FAM135B</td>
</tr>
</tbody>
</table>

Figure 4. Pre-Perturbation and Post-Perturbation RA vs OA expression level correlation.

The top two graphs compare RA knee to OA expression levels, while the bottom two compare RA hip to OA expression levels. It can be seen that there is a strong negative correlation in the left two graphs, which show the pre perturbation relations between the RA and OA states. The right graph show the new correlations for the two joint types after being perturbed by their respective top recipe.

Discussion and Conclusions

In this study we used dynamic Bayesian networks (DBN) in order to model the complex interactions between selected genes that were shown to be differentially expressed between RA and OA. The top 25 differentially expressed OA genes identified from Ai et al.[6] and top 25 most differentially expressed RA knee and RA hip genes from data set were selected as a 50 gene list. These genes were selected as they are most likely to be the greater causes for joint-specific rheumatoid arthritis state genes in the network and found the resulting correlation to OA.

We used single cell normalized RNA seq data set from Ai et al.[6] which includes 5 samples of rheumatoid arthritis (RA) knee, 5 samples of RA hip, and 10 samples of osteoarthritis (OA), which we regarded as the normal state. The top 25 differentially expressed OA genes identified from Ai et al.[6] and top 25 most differentially expressed RA knee and RA hip genes from data set were selected as a 50 gene list. These genes were selected as they are most likely to be the greater causes for joint-specific rheumatoid arthritis state genes in the network and found the resulting correlation to OA. We discretized the expression levels of each gene, shown in Figure 1a, and learned a Bayesian network model purely from quantitative data and did not incorporate prior knowledge.

FIGURE 1

(a) Data processing was done by discirnetizing over log10 TPM (transcripts per million) on the median. Genes are considered expressed (1) or not expressed (0).

(b) In order to train the highest-scoring Directed Acyclic Graph (DAG) which is a necessary component for the Bayesian network, we used a greedy algorithm which maximizes the BIC (Bayesian Information Criterion) score in order to improve running complexity. BIC is calculated:

$$BIC(G) = -2 \log(P(D|G)) - \frac{d}{2} \log(n),$$

where BIC(G) is the probability of the data given the DAG, P(D|G) is the probability of the data and d is the number of parameters.

(c) Table of the top 5 knee and hip recipes, different recipes are identified for each joint, FGF10, LRP1B, and SHISA5 for knee, and PLXNC1, LRP1B, and FAM135B for hip.

We noticed that the recipes for knee and hip RA are different. Most of the common genes in the top recipes also happened to be major nodes in genetic network (Figure 2a). Expectedly, similar genes (FGF10, PLXNC1, LRP1B) are seen in both joint recipes, however the combinations of genes are quite unique for each joint.