Inferring Gene Networks from Microarray Data using a Hybrid GA

Mark Cumiskey, John Levine and Douglas Armstrong

johnl@inf.ed.ac.uk
http://www.aiai.ed.ac.uk/~johnl

Institute for Adaptive and Neural Computation and Centre for Intelligent Systems and their Applications, School of Informatics, University of Edinburgh
Introduction

Many genome sequencing efforts are now complete
Introduction

- Many genome sequencing efforts are now complete
- Focus shifts to function of genes and their interactions
Introduction

- Many genome sequencing efforts are now complete
- Focus shifts to function of genes and their interactions
- Interactions shown as gene expression networks
Introduction

- Many genome sequencing efforts are now complete
- Focus shifts to function of genes and their interactions
- Interactions shown as gene expression networks
- Applications: cancer research, drug discovery, etc.
Introduction

- Many genome sequencing efforts are now complete
- Focus shifts to function of genes and their interactions
- Interactions shown as gene expression networks
- Applications: cancer research, drug discovery, etc.
- Microrarray data: parallel snapshot of gene activity
Introduction

Many genome sequencing efforts are now complete.
Focus shifts to function of genes and their interactions.
Interactions shown as gene expression networks.
Applications: cancer research, drug discovery, etc.
Microrarray data: parallel snapshot of gene activity.
Multiple microarray snapshots allow gene expression networks to be inferred.
Introduction

Many genome sequencing efforts are now complete
Focus shifts to function of genes and their interactions
Interactions shown as gene expression networks
Applications: cancer research, drug discovery, etc.
Microrarray data: parallel snapshot of gene activity
Multiple microarray snapshots allow gene expression networks to be inferred
Our aim is to infer gene expression network topologies and weights using a hybrid genetic algorithm
Introduction

Many genome sequencing efforts are now complete
Focus shifts to function of genes and their interactions
Interactions shown as gene expression networks
Applications: cancer research, drug discovery, etc.
Microrarray data: parallel snapshot of gene activity
Multiple microarray snapshots allow gene expression networks to be inferred
Our aim is to infer gene expression network topologies and weights using a hybrid genetic algorithm
We combine the GA with a back-propagation local search
Microarray Data

- Goal: to decipher the connections of the genetic network
- Pathway: DNA $\rightarrow$ mRNA $\rightarrow$ protein
- Microarray technology provides a snapshot of mRNA levels
- mRNA levels are an indirect measurement of gene activity
- Multiple mRNA snapshots over time reveal the gene interactions
- Massive data sets: 6,000 genes for the yeast cell
- Too large to infer anything meaningful by hand
Genetic Algorithm: Approach

- Each individual is a valid gene network

Network is a set of binary links between genes with weights on each link. Fitness is judged by how well the network predicts the microarray data. A specialist crossover operator is used to combine two networks. The Genetic Algorithm (GA) uses coarse-grained weights on the links. Refinement of the weights is done after the GA is finished using a back-propagation local search algorithm. The single population GA is compared with an island model. The approach is compared with Friedman's results on the Rosetta data set.
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
Genetic Algorithm: Approach

- Each individual is a valid gene network

- Network is a set of binary links between genes with weights on each link

- Fitness judged by how well network predicts the microarray data
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
- Fitness judged by how well network predicts the microarray data
- Specialist crossover operator to combine two networks
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
- Fitness judged by how well network predicts the microarray data
- Specialist crossover operator to combine two networks
- GA uses coarse-grained weights on the links
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
- Fitness judged by how well network predicts the microarray data
- Specialist crossover operator to combine two networks
- GA uses coarse-grained weights on the links
- Refine the weights after the GA is finished using a back-propagation local search algorithm
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
- Fitness judged by how well network predicts the microarray data
- Specialist crossover operator to combine two networks
- GA uses coarse-grained weights on the links
- Refine the weights after the GA is finished using a back-propagation local search algorithm
- Compare single population GA with an island model
Genetic Algorithm: Approach

- Each individual is a valid gene network
- Network is a set of binary links between genes with weights on each link
- Fitness judged by how well network predicts the microarray data
- Specialist crossover operator to combine two networks
- GA uses coarse-grained weights on the links
- Refine the weights after the GA is finished using a back-propagation local search algorithm
- Compare single population GA with an island model
- Compare with Friedman’s results on Rosetta data set
Example Gene Network
Representing Gene Networks

Matrix representation:

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.3 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.51 & 0.31 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]
Representing Gene Networks

String representation:

\[ <1 \sim 0.3> <2 \sim -0.4> ^3, <2 \sim 0.1> ^4, <3 \sim 0.51> <4 \sim 0.31> ^5, <5 \sim 0.1> ^6, <6 \sim 0.2> ^7 \]

Weight From Node To Node

Possible splice site
Evaluating Network Fitness

Estimate gene expression levels at time $t + 1$ given levels at time $t$:

$$s_i(t + 1) = \sum_{j=0}^{n} w_{ji} x_j(t)$$
Evaluating Network Fitness

- Estimate gene expression levels at time $t + 1$ given levels at time $t$:

$$s_i(t + 1) = \sum_{j=0}^{n} w_{ji} x_j(t)$$

- Pass estimate through a sigmoid function for biological realism:

$$x_i(t + 1) = \frac{n_j}{1 + e^{-n_j(s_i(t+1))}}$$
Evaluating Network Fitness

Overall network fitness:

\[ \text{fitness} = \sum_{i=0}^{n} \sum_{t=0}^{T} |y_i(t) - x_i(t)| + \text{num\_nodes}/b \]

Imposed bias towards smaller networks
Single Point Network Crossover

Parent 1

Gene 1 → Gene 3
Gene 2 → Gene 4
Gene 3 → Gene 20
Gene 4 → Gene 21
Gene 20 → Gene 22
Gene 21 → Gene 23
Gene 22 → Gene 23

Parent 2

Gene 5 → Gene 6
Gene 6 → Gene 7
Gene 7 → Gene 8
Gene 7 → Gene 9
Gene 7 → Gene 10

Child

Gene 1 → Gene 3
Gene 2 → Gene 4
Gene 3 → Gene 5
Gene 4 → Gene 5
Gene 5 → Gene 6
Gene 6 → Gene 7
## Single Machine Results

<table>
<thead>
<tr>
<th>Initial Pop</th>
<th>Net Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>50</td>
<td>186.60</td>
<td>2.032</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.200</td>
<td>0</td>
<td>2 mins</td>
</tr>
<tr>
<td>5000</td>
<td>50</td>
<td>180.43</td>
<td>1.200</td>
<td>8</td>
<td>35 min</td>
</tr>
<tr>
<td>5000</td>
<td>20</td>
<td>221.40</td>
<td>0.8323</td>
<td>6</td>
<td>5 min</td>
</tr>
</tbody>
</table>
## Single Machine Results

<table>
<thead>
<tr>
<th>Initial Pop</th>
<th>Net Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>50</td>
<td>186.60</td>
<td>2.032</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.200</td>
<td>0</td>
<td>2 mins</td>
</tr>
<tr>
<td>5000</td>
<td>50</td>
<td>180.43</td>
<td>1.200</td>
<td>8</td>
<td>35 min</td>
</tr>
<tr>
<td>5000</td>
<td>20</td>
<td>221.40</td>
<td>0.8323</td>
<td>6</td>
<td>5 min</td>
</tr>
</tbody>
</table>

Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000
Single Machine Results

<table>
<thead>
<tr>
<th>Initial Pop</th>
<th>Net Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>50</td>
<td>186.60</td>
<td>2.032</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.200</td>
<td>0</td>
<td>2 mins</td>
</tr>
<tr>
<td>5000</td>
<td>50</td>
<td>180.43</td>
<td>1.200</td>
<td>8</td>
<td>35 min</td>
</tr>
<tr>
<td>5000</td>
<td>20</td>
<td>221.40</td>
<td>0.8323</td>
<td>6</td>
<td>5 min</td>
</tr>
</tbody>
</table>

- Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000
- Very good fitness networks found...
Single Machine Results

<table>
<thead>
<tr>
<th>Initial Pop</th>
<th>Net Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>50</td>
<td>186.60</td>
<td>2.032</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.200</td>
<td>0</td>
<td>2 mins</td>
</tr>
<tr>
<td>5000</td>
<td>50</td>
<td>180.43</td>
<td>1.200</td>
<td>8</td>
<td>35 min</td>
</tr>
<tr>
<td>5000</td>
<td>20</td>
<td>221.40</td>
<td>0.8323</td>
<td>6</td>
<td>5 min</td>
</tr>
</tbody>
</table>

- Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000
- Very good fitness networks found...
- But a different (sub-)network every time
### Single Machine Results

<table>
<thead>
<tr>
<th>Initial Pop</th>
<th>Net Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>50</td>
<td>186.60</td>
<td>2.032</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.200</td>
<td>0</td>
<td>2 mins</td>
</tr>
<tr>
<td>5000</td>
<td>50</td>
<td>180.43</td>
<td>1.200</td>
<td>8</td>
<td>35 min</td>
</tr>
<tr>
<td>5000</td>
<td>20</td>
<td>221.40</td>
<td>0.8323</td>
<td>6</td>
<td>5 min</td>
</tr>
</tbody>
</table>

- Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000
- Very good fitness networks found...
- But a different (sub-)network every time
- Impossible to validate results
Island Model GA

after each generation each island sends its best individual to all the other islands
## Island Model Results

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Pop</th>
<th>Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2000</td>
<td>50</td>
<td>225.60</td>
<td>2.214</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.334</td>
<td>2</td>
<td>7 mins</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>50</td>
<td>223.43</td>
<td>2.200</td>
<td>4</td>
<td>42 min</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>20</td>
<td>227.30</td>
<td>1.542</td>
<td>7</td>
<td>9 min</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>50</td>
<td>122.60</td>
<td>2.635</td>
<td>9</td>
<td>30 min</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>20</td>
<td>117.16</td>
<td>1.986</td>
<td>5</td>
<td>6 mins</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>50</td>
<td>116.22</td>
<td>1.256</td>
<td>7</td>
<td>40 min</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>20</td>
<td>132.30</td>
<td>0.623</td>
<td>5</td>
<td>11 min</td>
</tr>
</tbody>
</table>
## Island Model Results

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Pop</th>
<th>Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2000</td>
<td>50</td>
<td>225.60</td>
<td>2.214</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.334</td>
<td>2</td>
<td>7 mins</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>50</td>
<td>223.43</td>
<td>2.200</td>
<td>4</td>
<td>42 min</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>20</td>
<td>227.30</td>
<td>1.542</td>
<td>7</td>
<td>9 min</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>50</td>
<td>122.60</td>
<td>2.635</td>
<td>9</td>
<td>30 min</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>20</td>
<td>117.16</td>
<td>1.986</td>
<td>5</td>
<td>6 mins</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>50</td>
<td>116.22</td>
<td>1.256</td>
<td>7</td>
<td>40 min</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>20</td>
<td>132.30</td>
<td>0.623</td>
<td>5</td>
<td>11 min</td>
</tr>
</tbody>
</table>

Better fitness, but same problem as before
## Island Model Results

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Pop</th>
<th>Size</th>
<th>GA Fitness</th>
<th>BP Error</th>
<th>Markov matches</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2000</td>
<td>50</td>
<td>225.60</td>
<td>2.214</td>
<td>6</td>
<td>30 min</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>20</td>
<td>257.16</td>
<td>1.334</td>
<td>2</td>
<td>7 mins</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>50</td>
<td>223.43</td>
<td>2.200</td>
<td>4</td>
<td>42 min</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>20</td>
<td>227.30</td>
<td>1.542</td>
<td>7</td>
<td>9 mins</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>50</td>
<td>122.60</td>
<td>2.635</td>
<td>9</td>
<td>30 min</td>
</tr>
<tr>
<td>8</td>
<td>2000</td>
<td>20</td>
<td>117.16</td>
<td>1.986</td>
<td>5</td>
<td>6 mins</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>50</td>
<td>116.22</td>
<td>1.256</td>
<td>7</td>
<td>40 min</td>
</tr>
<tr>
<td>8</td>
<td>5000</td>
<td>20</td>
<td>132.30</td>
<td>0.623</td>
<td>5</td>
<td>11 min</td>
</tr>
</tbody>
</table>

- Better fitness, but same problem as before
- Results with simulated data demonstrate validity of the technique
Conclusions

Techniques for reconstruction of gene networks are still in their infancy
Techniques for reconstruction of gene networks are still in their infancy

We lack suitable benchmarks to validate techniques
Conclusions

- Techniques for reconstruction of gene networks are still in their infancy
- We lack suitable benchmarks to validate techniques
- Current techniques can produce plausible networks to pass to a biologist for verification
Conclusions

- Techniques for reconstruction of gene networks are still in their infancy
- We lack suitable benchmarks to validate techniques
- Current techniques can produce plausible networks to pass to a biologist for verification
- The GA can find highly fit networks that explain the test data
Conclusions

- Techniques for reconstruction of gene networks are still in their infancy
- We lack suitable benchmarks to validate techniques
- Current techniques can produce plausible networks to pass to a biologist for verification
- The GA can find highly fit networks that explain the test data
- Back-propagation can be used to fine tune the network weights
Conclusions

- Techniques for reconstruction of gene networks are still in their infancy
- We lack suitable benchmarks to validate techniques
- Current techniques can produce plausible networks to pass to a biologist for verification
- The GA can find highly fit networks that explain the test data
- Back-propagation can be used to fine tune the network weights
- The island model markedly improves the fitness level achieved
Conclusions

- Techniques for reconstruction of gene networks are still in their infancy
- We lack suitable benchmarks to validate techniques
- Current techniques can produce plausible networks to pass to a biologist for verification
- The GA can find highly fit networks that explain the test data
- Back-propagation can be used to fine tune the network weights
- The island model markedly improves the fitness level achieved
- Simulation may be able to provide benchmark data