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 microarrway 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 microarrway 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 microarrway 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 → mRNA → protein
- Mircoarray 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

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

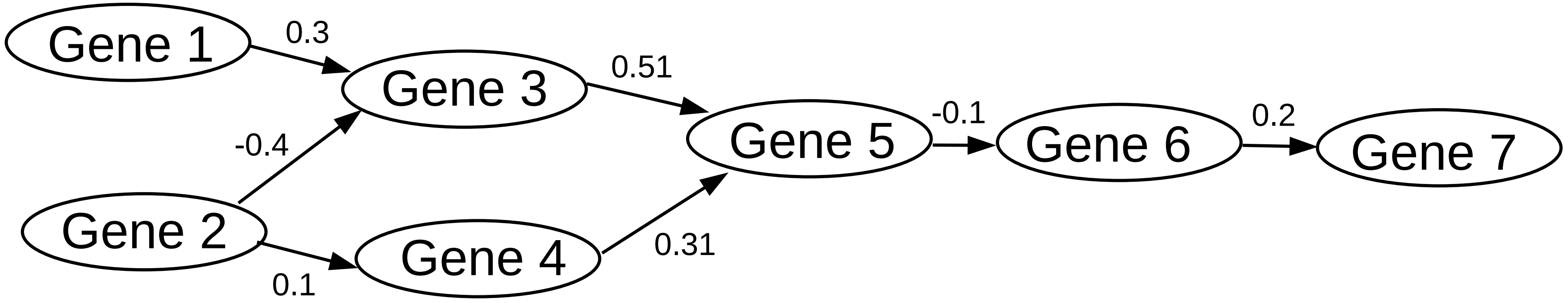
Genetic Algorithm: Approach

- 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



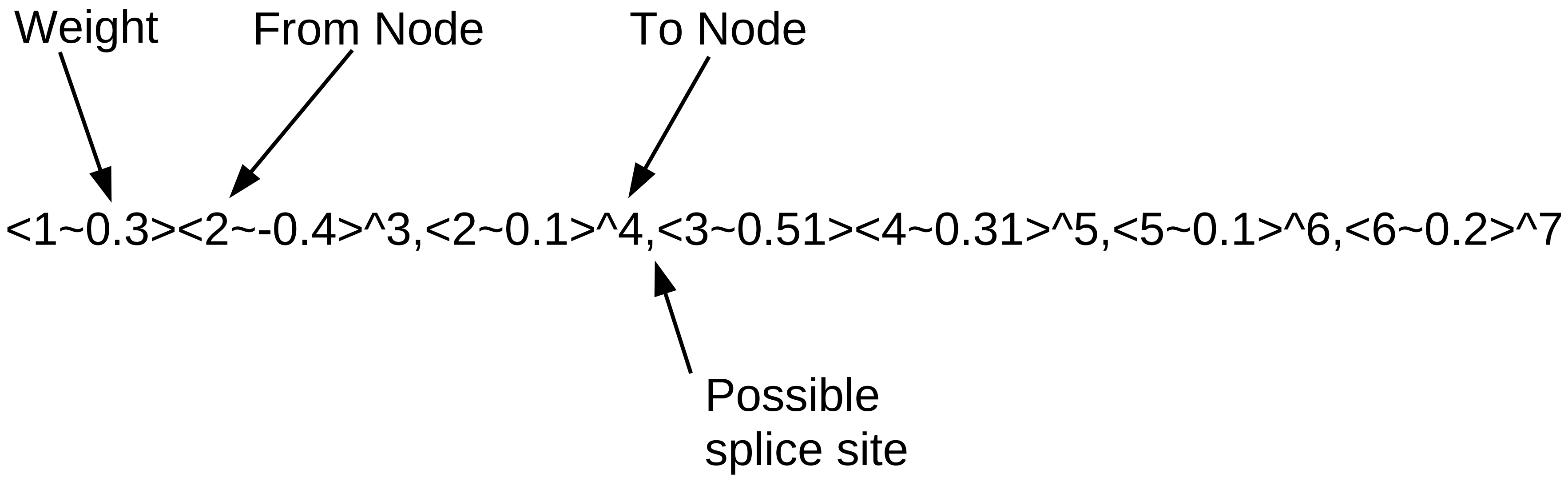
Representing Gene Networks

- Matrix representation:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.51 & 0.31 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Representing Gene Networks

- String representation:



Weight

From Node

To Node

<1~0.3><2~-0.4>^3,<2~0.1>^4,<3~0.51><4~0.31>^5,<5~0.1>^6,<6~0.2>^7

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:

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

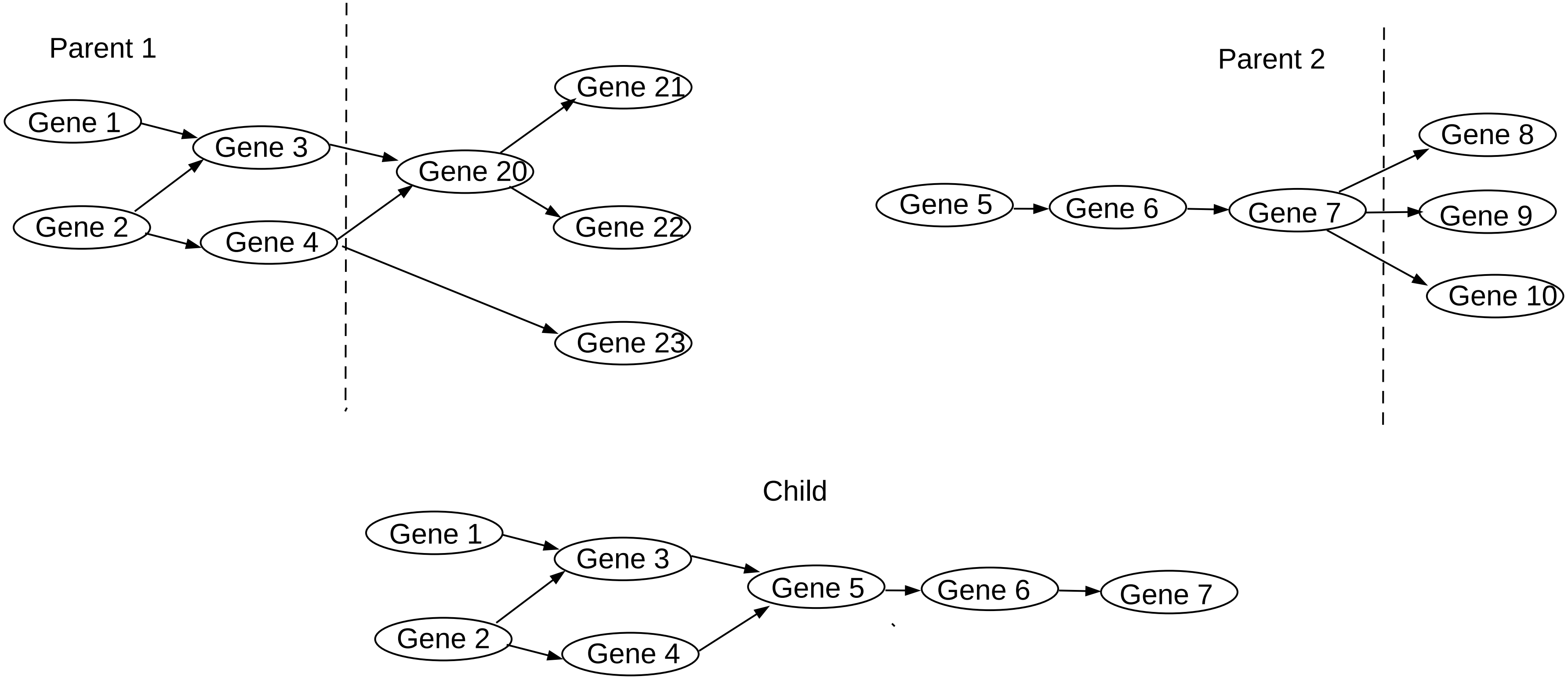
Evaluating Network Fitness

- Overall network fitness:

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

- Imposed bias towards smaller networks

Single Point Network Crossover



Single Machine Results

Initial Pop	Net Size	GA Fitness	BP Error	Markov matches	Time
2000	50	186.60	2.032	6	30 min
2000	20	257.16	1.200	0	2 mins
5000	50	180.43	1.200	8	35 min
5000	20	221.40	0.8323	6	5 min

Single Machine Results

Initial Pop	Net Size	GA Fitness	BP Error	Markov matches	Time
2000	50	186.60	2.032	6	30 min
2000	20	257.16	1.200	0	2 mins
5000	50	180.43	1.200	8	35 min
5000	20	221.40	0.8323	6	5 min

- Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000

Single Machine Results

Initial Pop	Net Size	GA Fitness	BP Error	Markov matches	Time
2000	50	186.60	2.032	6	30 min
2000	20	257.16	1.200	0	2 mins
5000	50	180.43	1.200	8	35 min
5000	20	221.40	0.8323	6	5 min

- 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

Initial Pop	Net Size	GA Fitness	BP Error	Markov matches	Time
2000	50	186.60	2.032	6	30 min
2000	20	257.16	1.200	0	2 mins
5000	50	180.43	1.200	8	35 min
5000	20	221.40	0.8323	6	5 min

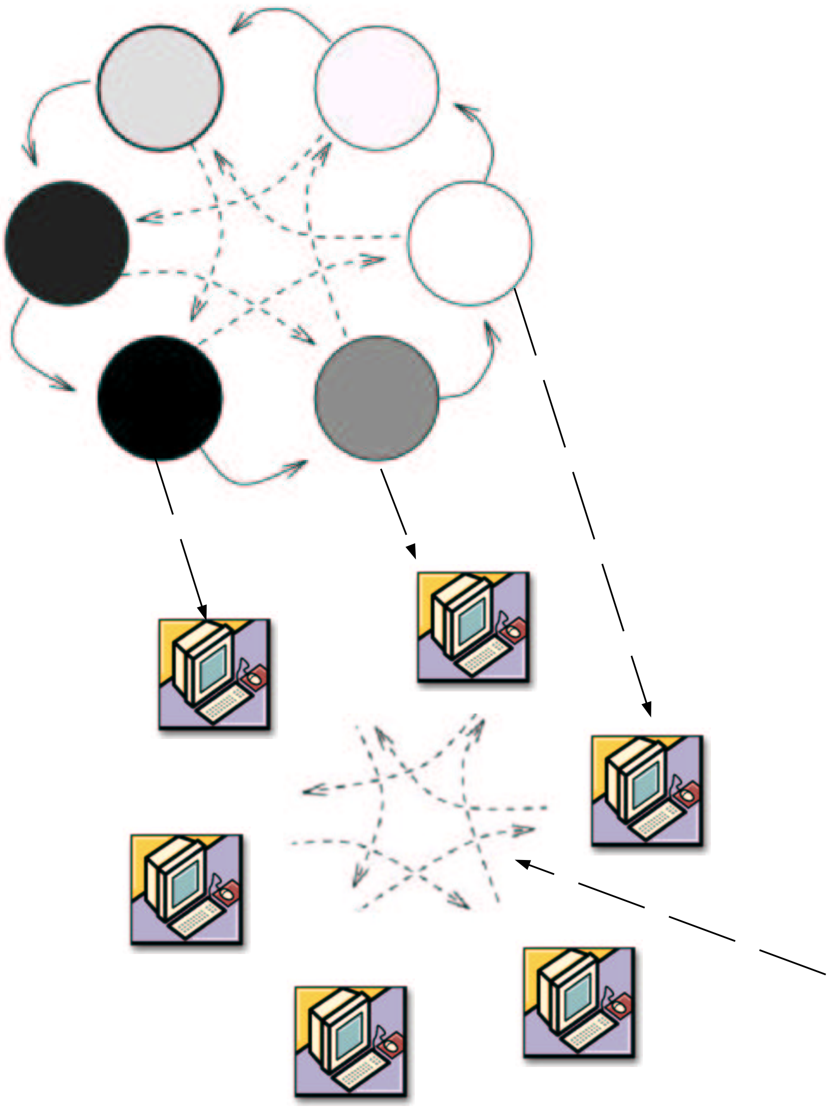
- Best random network of initial population of 5000 with 50 nodes has a fitness of over 15000
- Very good fitness networks found. . .
- But a different (sub-)network every time

Single Machine Results

Initial Pop	Net Size	GA Fitness	BP Error	Markov matches	Time
2000	50	186.60	2.032	6	30 min
2000	20	257.16	1.200	0	2 mins
5000	50	180.43	1.200	8	35 min
5000	20	221.40	0.8323	6	5 min

- 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 it best individual to all the other islands

Island Model Results

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

Island Model Results

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

- Better fitness, but same problem as before

Island Model Results

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

- 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

Conclusions

- 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