Interactively exploring distributed computational models of biology

James Watson & Janet Wiles

ARC Centre for Complex Systems

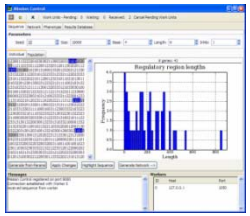
ARC Centre of Excellence in Bioinformatics

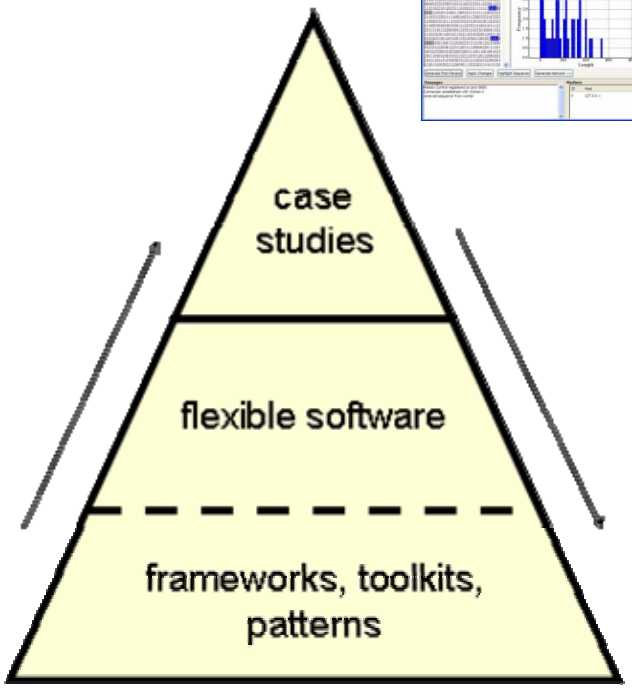
ARC Centre for Complex Systems
School of ITEE | The University of Queensland | ST LUCIA QLD 4069 | AUSTRALIA

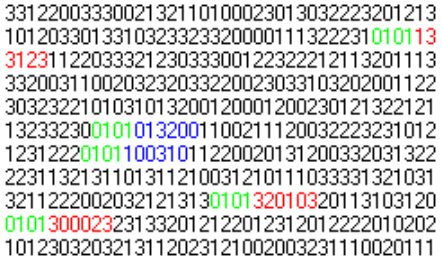
T: +61 7 3365 1003 | F: +61 7 3365 1533 | E: outreach@accs.edu.au

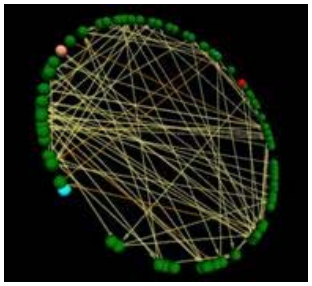
Biological modeling techniques

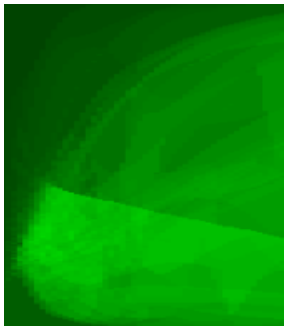
Small team sizes
Transient models
Reused components
Reinvention
Team members move on

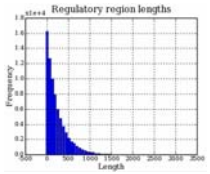


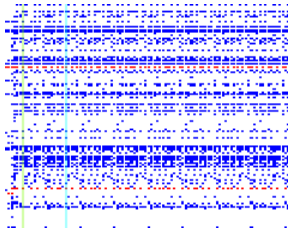


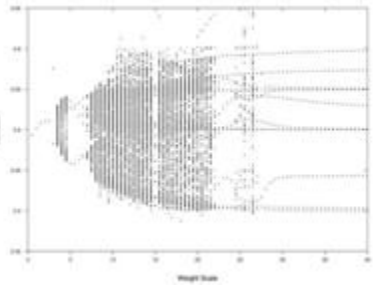


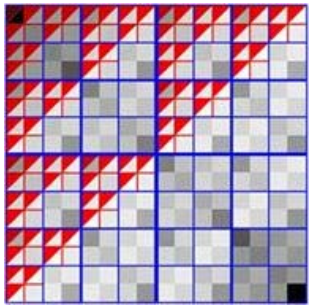












Interactive modeling

- The ACCS GRN group has found dynamic visualization and interaction particularly useful in biological modeling preferable to batched simulations
  - Cross-disciplinary makes model shared language
  - Projects (e.g., Neurosphere Lab) have been successful when collaborator interacts / critiques through visualizations / GUI's
  - User exploration facilitates insight into dynamics
  - Quickly hone in on points of interest by dynamically prioritizing (steering) computations to those regions

Simulations take time

Speed is necessary for interactive modeling

Options

- Improve software
- Faster hardware
  - becoming multi-core
- More hardware

Why distributed?

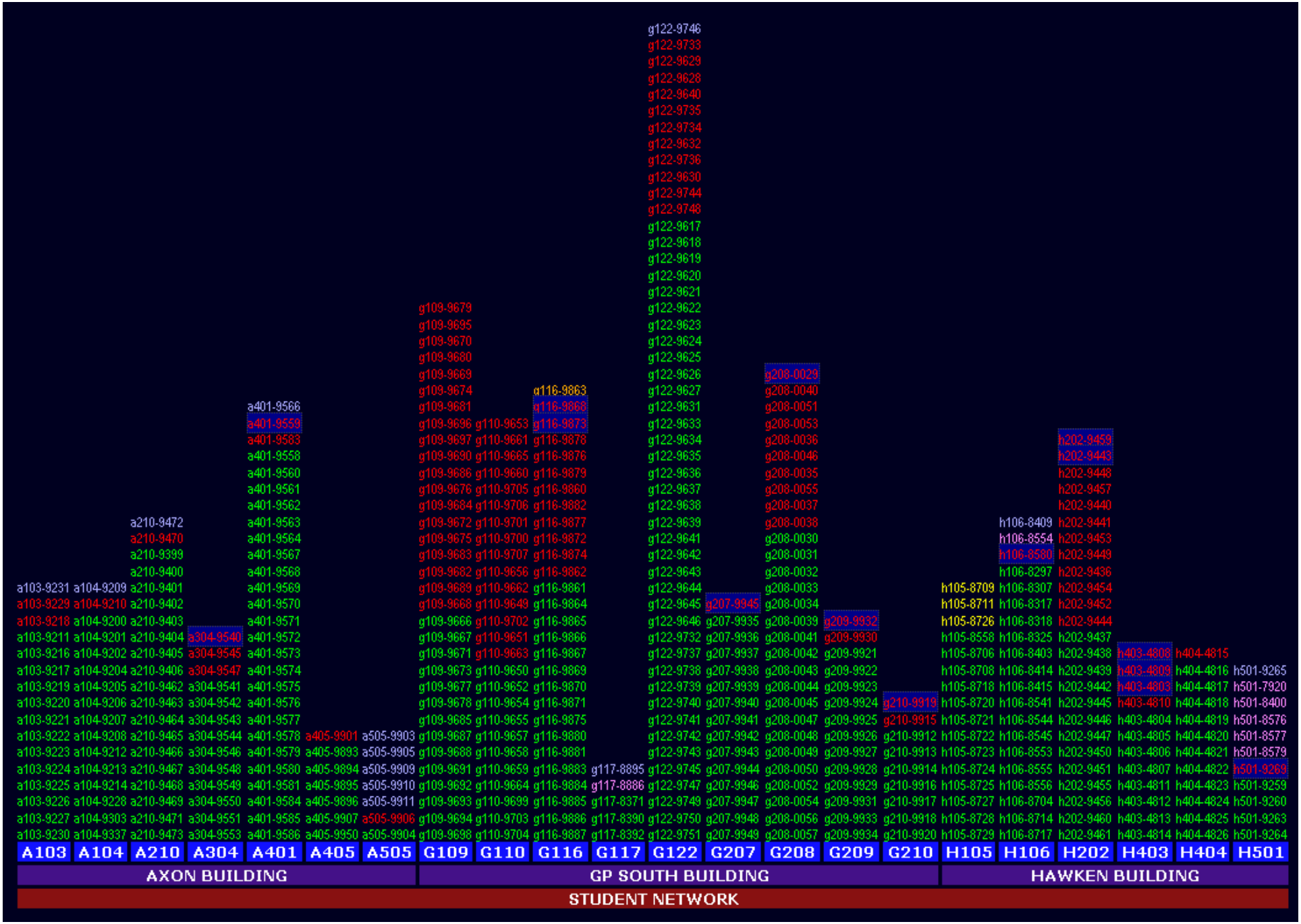
- Supercomputing is not ideal for interactivity
- The Grid
  - Many machines currently available
- Idle CPU's in many research institutes
- This project: explore feasibility of interactive simulations using distributed computing

Available hardware

- ~400 Desktop machines in student labs
- Fast local network

GX150SD - P3 1GHz, 256MB RAM
GX270SD - P4 2.8GHz, 512MB RAM
GX280SD - P4 3GHZ HT, 512MB RAM
GX280DT - P4 3GHZ HT, 512MB RAM
GX520DT - P4 3GHZ HT, 1024MB RAM
OP745DT - Core2 Duo 2.13GHz, 2048MB RAM

“All hardware types have either 100mbit or 1000mbit network cards, all of them are connected at 100mbit speed with gigabit backbones.”

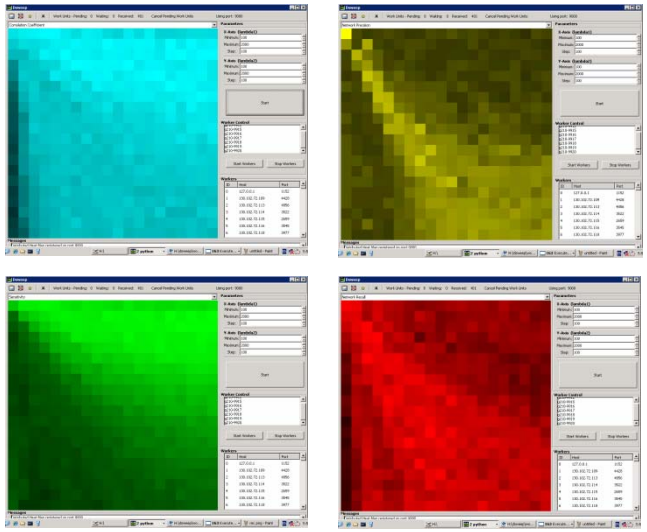

A103 A104 A210 A304 A401 A405 A505 G109 G110 G116 G117 G122 G207 G208 G209 G210 H105 H106 H202 H403 H404 H501
AXON BUILDING
GP SOUTH BUILDING
HAWKEN BUILDING
STUDENT NETWORK

Lab Characteristics

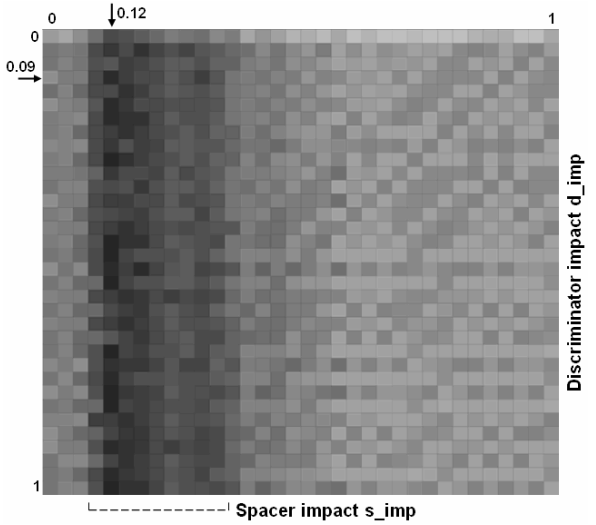
- Constantly changing usage
- Re-imaged on boot
- Have a ‘failure’ rate of ~5 machines an hour

Distributed speed-up

- Search for regularisation parameter values in multi-purpose machine learning
  - Up to 20 minutes per run

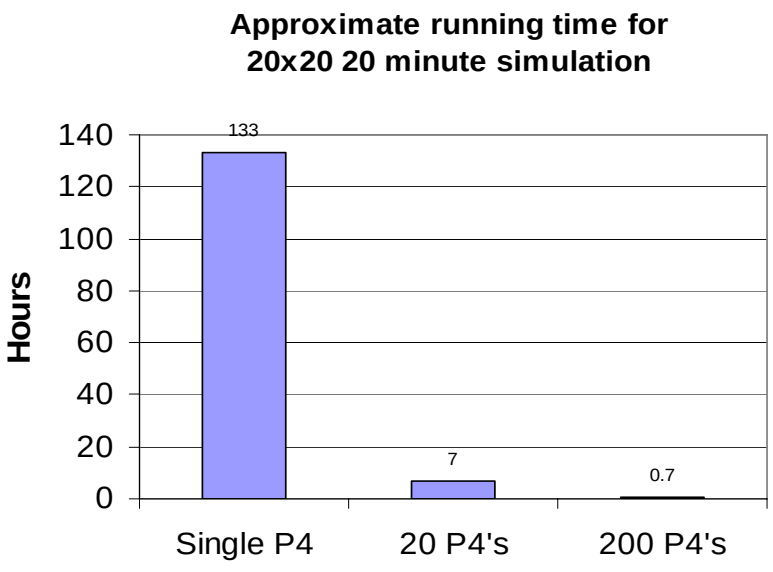


- Promoter model of *Bacillus subtilis* for transcription start site prediction
  - ~3 minutes per run
  - We performed 1000's of runs for a recent submission

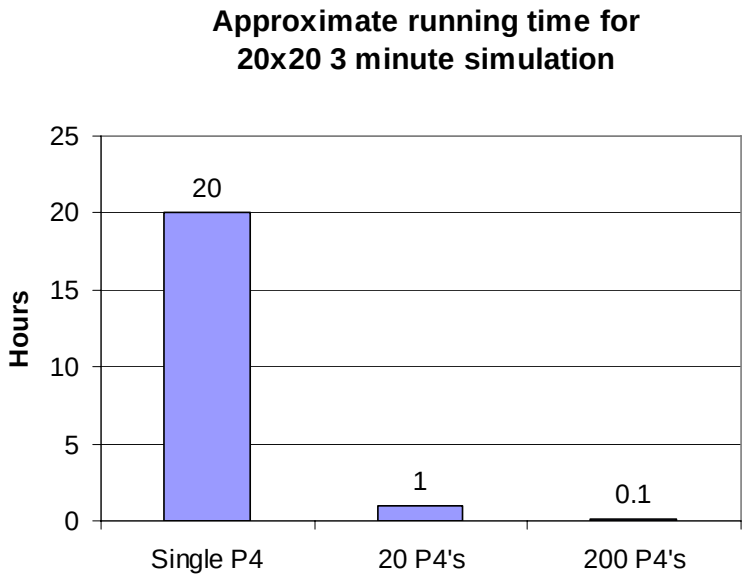


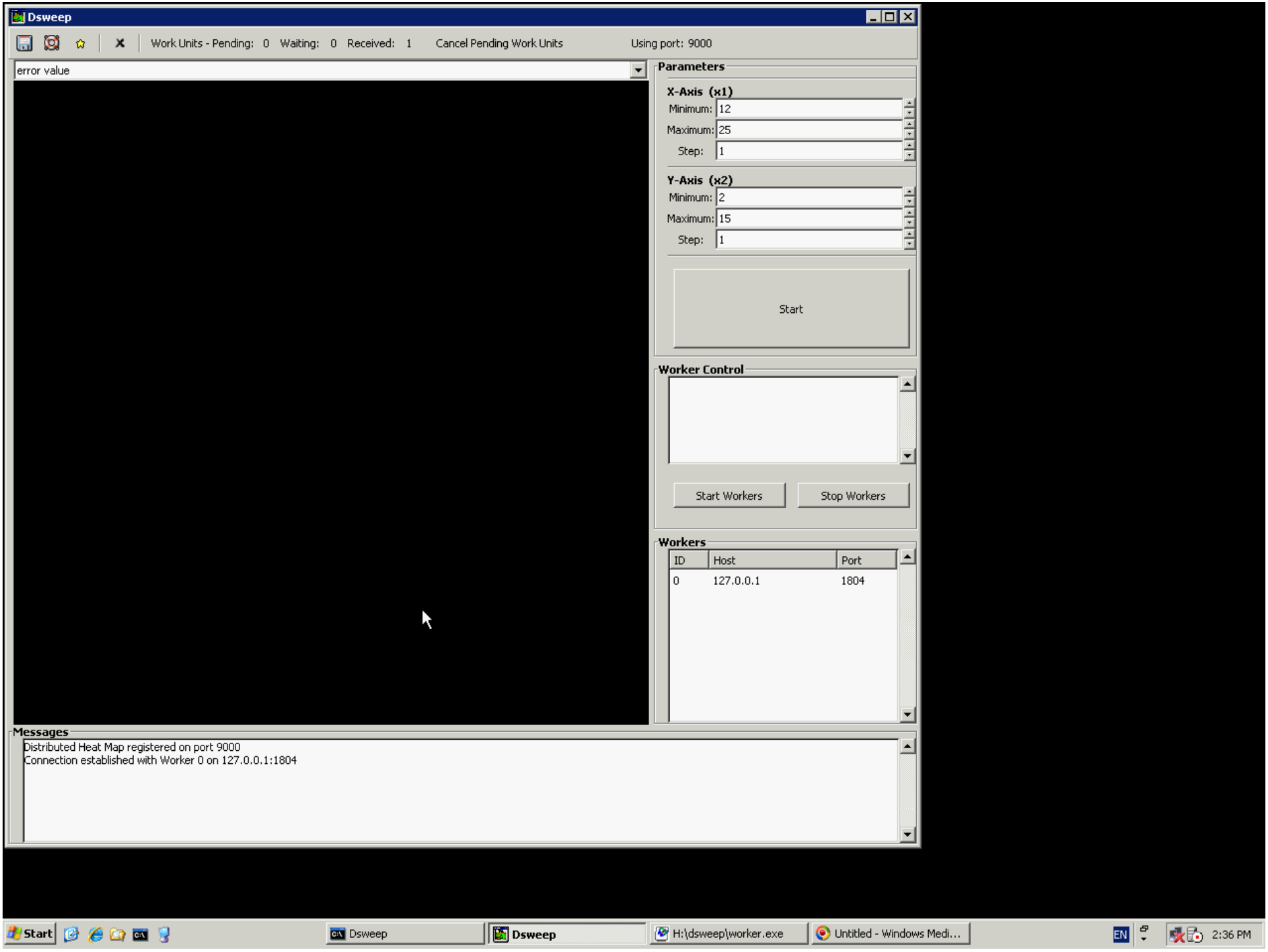
Distributed speed-up

- Search for regularisation parameter values in multi-purpose machine learning
  - Up to 20 minutes per run



- Promoter model of *Bacillus subtilis* for transcription start site prediction
  - ~3 minutes per run
  - We performed 1000's of runs for a recent submission




Dsweep
Work Units - Pending: 0 Waiting: 0 Received: 1
Cancel Pending Work Units
Using port: 9000
error value
Parameters
X-Axis (x1)
Minimum: 12
Maximum: 25
Step: 1
Y-Axis (x2)
Minimum: 2
Maximum: 15
Step: 1
Start
Worker Control
Start Workers
Stop Workers
Workers
ID
Host
Port
0
127.0.0.1
1804
Messages
Distributed Heat Map registered on port 9000
Connection established with Worker 0 on 127.0.0.1:1804
Start
Dsweep
Dsweep
H:\dsweep\worker.exe
Untitled - Windows Medi...
EN
2:36 PM

Beyond parameter sweeps

- Many biological models are “embarrassingly parallel”
  - Individuals / populations / etc. can be easily divided into distinct units of work
    - Generate regulatory network from genetic sequence
  - Work units often small, but many

Evolving the Artificial Genome 2.0

Model Help

Explore Mapping | Evolution

Genome

Seed: 10 Size: 10000

Base: 4 Length: 6

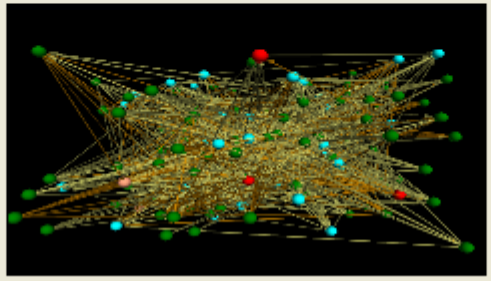
Inhibition: 1 Time Steps: 100

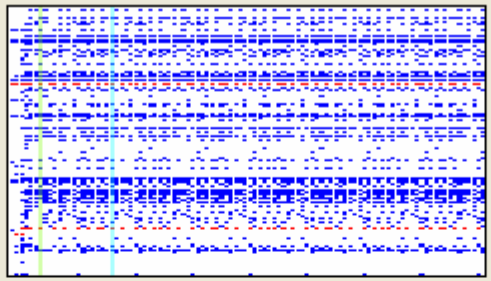
Generate

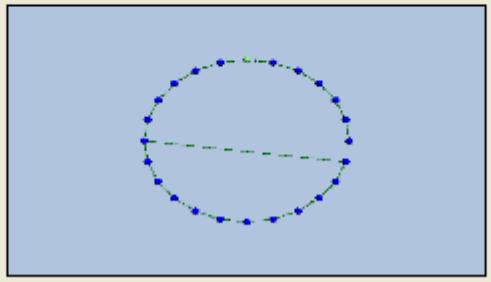
```
01220113233320330303133323302031200011 30
30210100032211331200311101203210003000 31
30002113333322000032221133201201330303 22
13303202200310213312031331130131101322 11
33122003330021321101000230130322320121 3
10120330133103233233200001113222310101 13
31231122033321230330012232221211320111 3
33200311002032320332200230331032020011 22
30323221010310132001200012002301213221 21
13233230101013200110021112003222323101 2
12312220101100310112200201312003320313 22
22311321311013112100312101110333313210 31
32112220020321213130101320103201131031 20
01013000232313320121220123120122220102 02
10123032032131120231210020032311100201 11
31101021130000032023103012031301121220 31
13033010001322121233030122020100111023 11
10312033031003113003333231210031022003 11
31120031020220010010222013321121121200 13
20023110303310301110003203012210121213 00
23112303001031230220111201320331101122 30
30020202332203211303312210323121111000 21
21202111203321012220313211021122032221 03
22021030012331201331223001230221301111 2
```

Network

nodes: 200 # links: 2336 K: 11.680







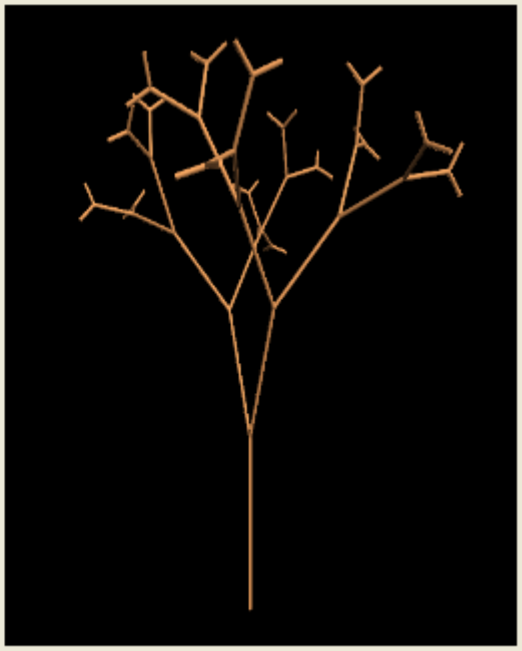
Phenotype

☑ Light Interception: 71.57 %

☑ Reproductive Success: 76.93 %

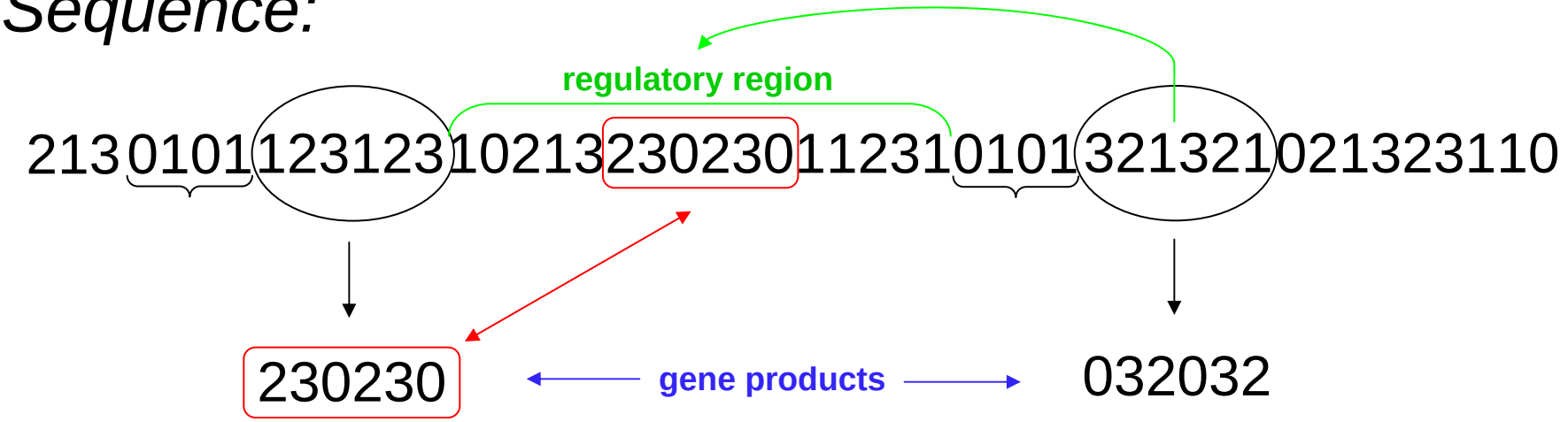
☑ Surface Area: 100.00 %

Total Relative Fitness: **3.80**

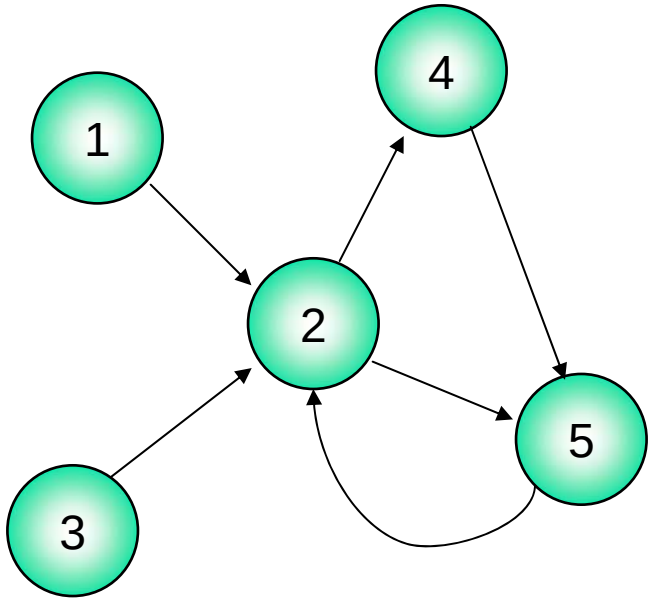


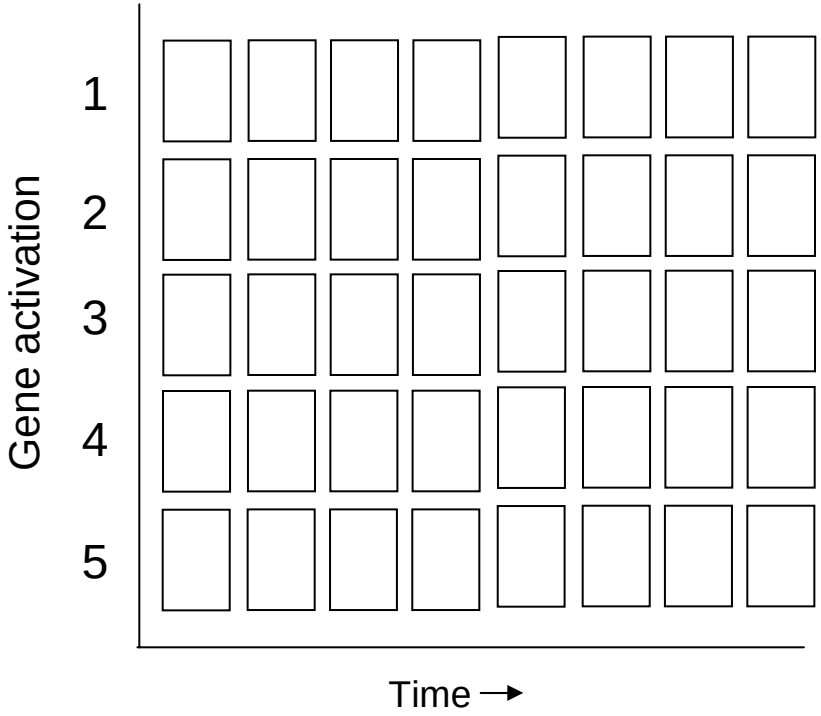
The Artificial Genome

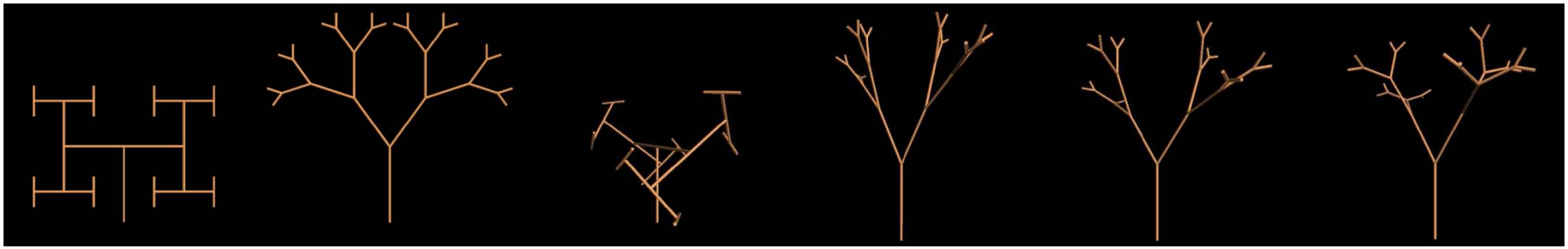
Sequence:

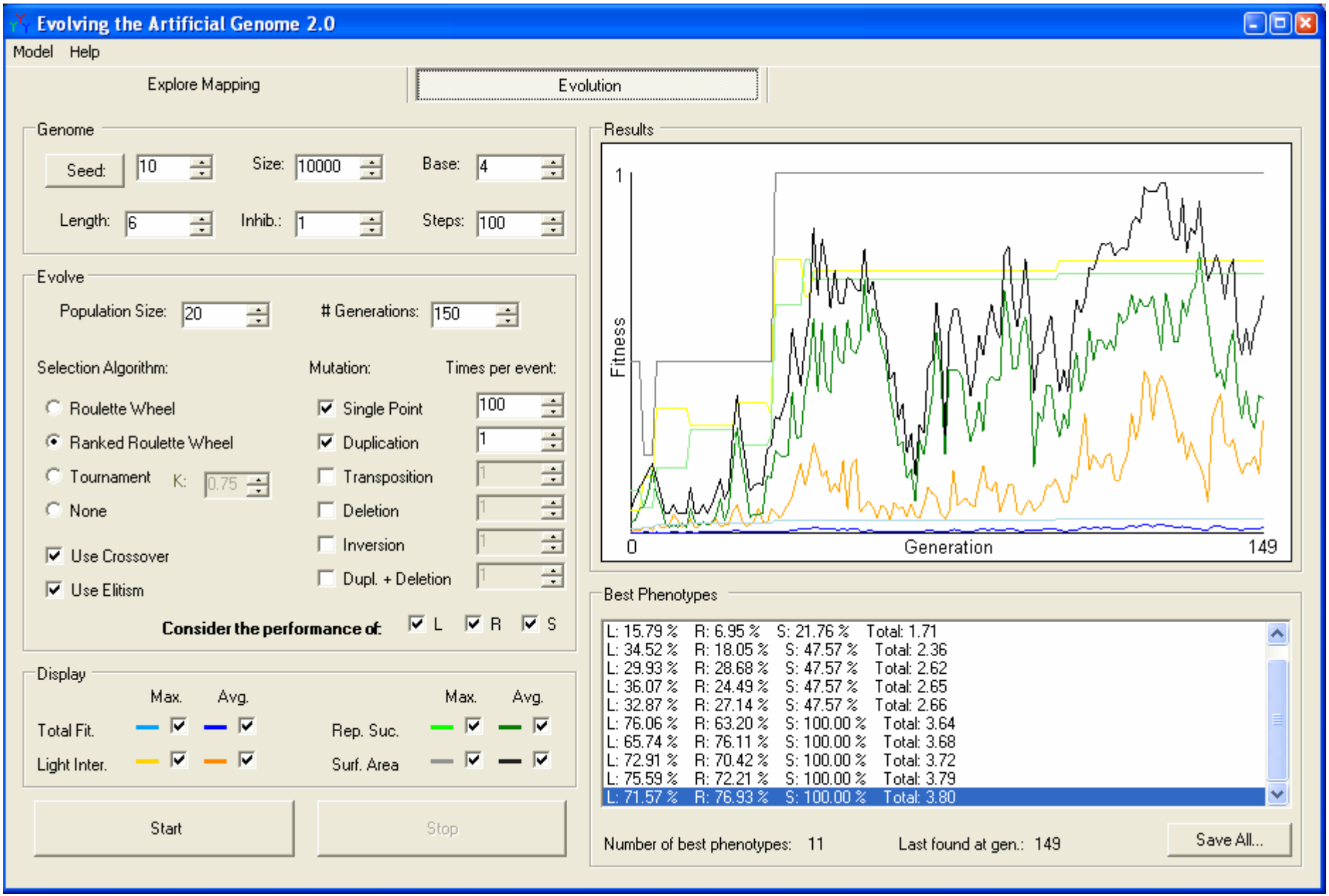


Defines network:


1
2
3
4
5


Gene activation
1
2
3
4
5
Time




Evolving the Artificial Genome 2.0
Model
Help
Explore Mapping
Evolution
Genome
Seed:
10
Size:
10000
Base:
4
Length:
6
Inhib.:
1
Steps:
100
Evolve
Population Size:
20
Generations:
150
Selection Algorithm:
Roulette Wheel
Ranked Roulette Wheel
Tournament
K:
0.75
None
Use Crossover
Use Elitism
Mutation:
Times per event:
Single Point
100
Duplication
1
Transposition
Deletion
Inversion
Dupl. + Deletion
Consider the performance of:
L
R
S
Display
Max.
Avg.
Total Fit.
Light Inter.
Rep. Suc.
Surf. Area
Start
Stop
Results
1
Fitness
0
Generation
149
Best Phenotypes
L: 15.79 % R: 6.95 % S: 21.76 % Total: 1.71
L: 34.52 % R: 18.05 % S: 47.57 % Total: 2.36
L: 29.93 % R: 28.68 % S: 47.57 % Total: 2.62
L: 36.07 % R: 24.49 % S: 47.57 % Total: 2.65
L: 32.87 % R: 27.14 % S: 47.57 % Total: 2.66
L: 76.06 % R: 63.20 % S: 100.00 % Total: 3.64
L: 65.74 % R: 76.11 % S: 100.00 % Total: 3.68
L: 72.91 % R: 70.42 % S: 100.00 % Total: 3.72
L: 75.59 % R: 72.21 % S: 100.00 % Total: 3.79
L: 71.57 % R: 76.93 % S: 100.00 % Total: 3.80
Number of best phenotypes: 11
Last found at gen.: 149
Save All...

Distributing the model

Change key method calls to messages

Create_Seq() → request message ‘Create_Seq’

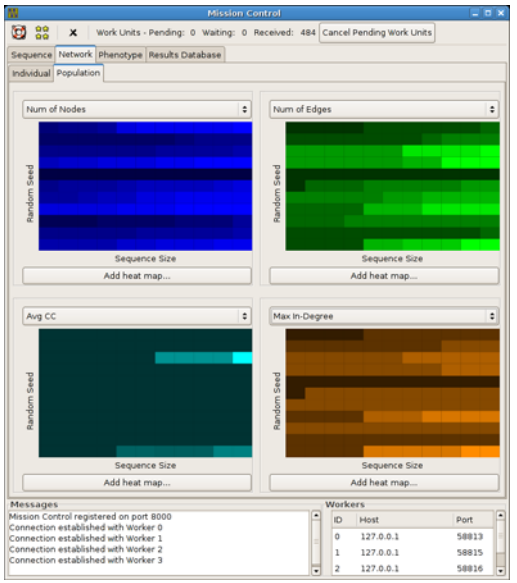
Distribute workers that understand these messages

Maintain list of requests and their parameters

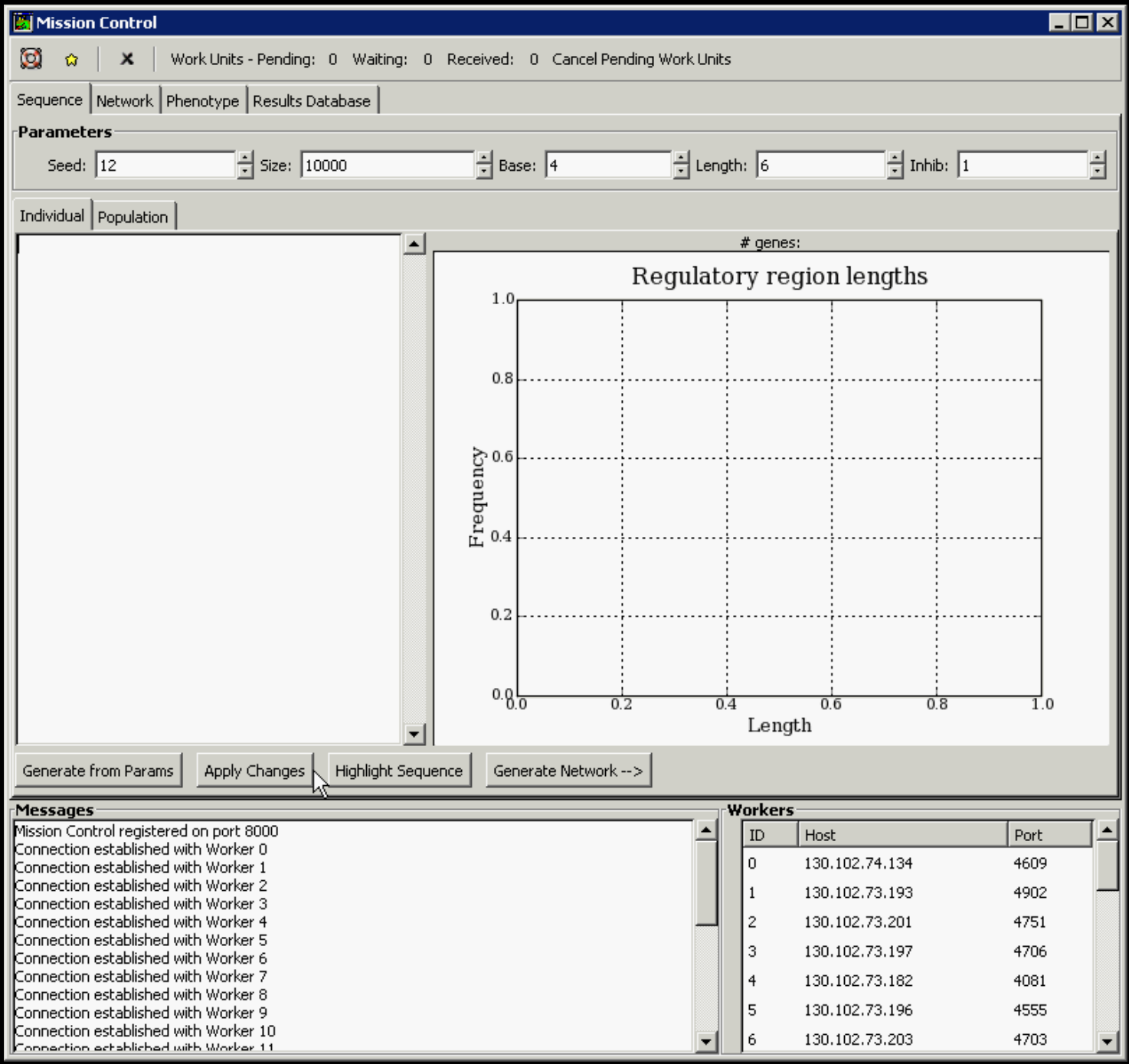
Send requests to idle workers

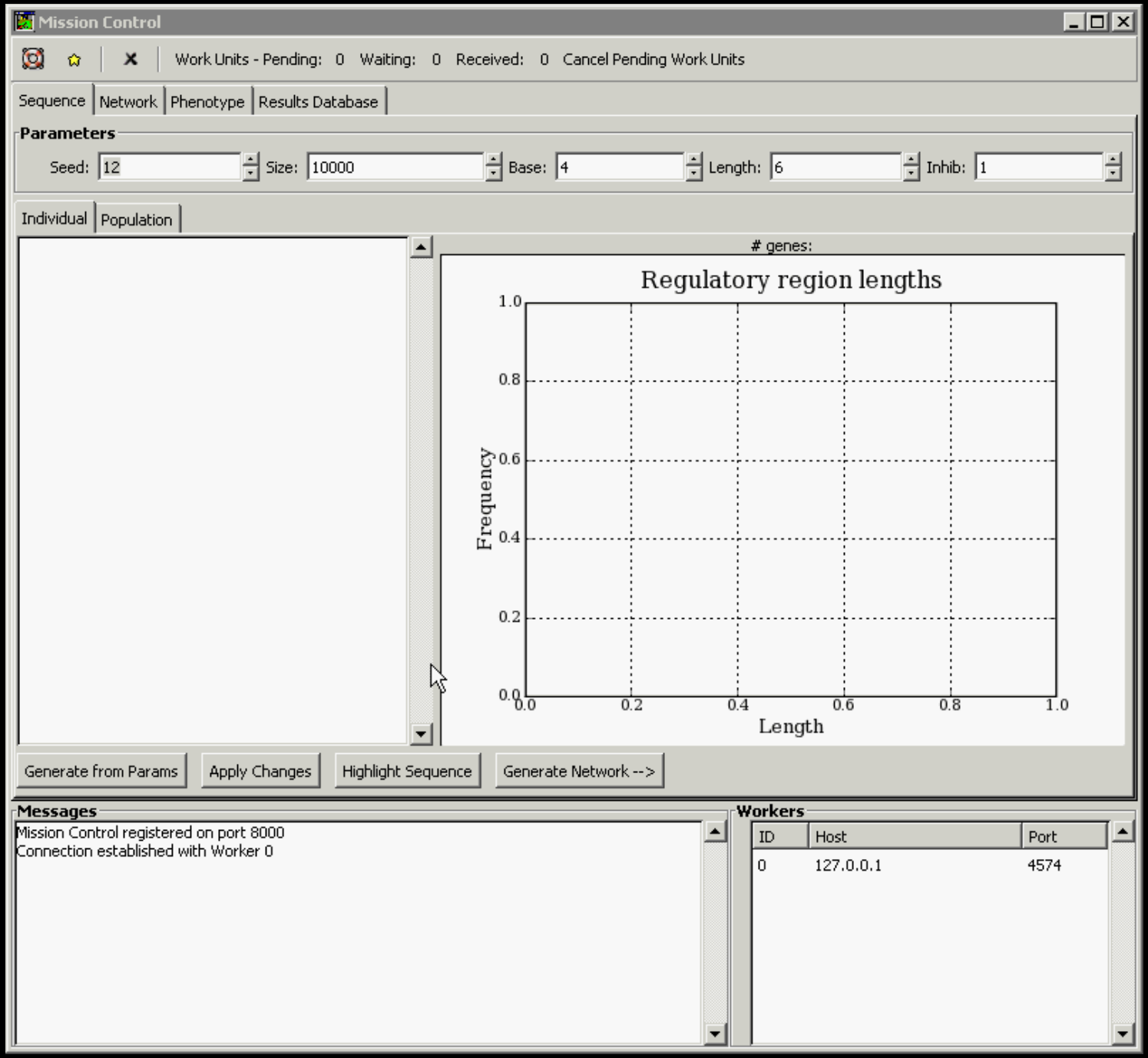
Update GUI as results are returned

Prototype



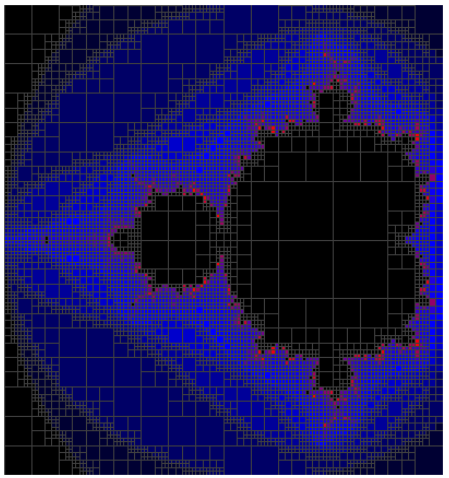
- Minimal image impact
  - Workers are stand-alone binaries
  - Bundle simulation code / data
  - Linux & Windows
- Work requests made dynamically according to GUI interaction
- Worker machines can be dynamically added / removed
- No administrative privileges required
  - no services, frameworks, etc.
  - but need network access to machines
- Practical, but no replacement for existing distributed tools
  - Manual fault recovery
  - No sophisticated scheduling (for this, see Condor / Nimrod)


Mission Control
Work Units - Pending: 0 Waiting: 0 Received: 0 Cancel Pending Work Units
Sequence
Network
Phenotype
Results Database
Parameters
Seed: 12
Size: 10000
Base: 4
Length: 6
Inhib: 1
Individual
Population
genes:
Regulatory region lengths
Frequency
Length
0.0
0.2
0.4
0.6
0.8
1.0
Generate from Params
Apply Changes
Highlight Sequence
Generate Network -->
Messages
Mission Control registered on port 8000
Connection established with Worker 0
Connection established with Worker 1
Connection established with Worker 2
Connection established with Worker 3
Connection established with Worker 4
Connection established with Worker 5
Connection established with Worker 6
Connection established with Worker 7
Connection established with Worker 8
Connection established with Worker 9
Connection established with Worker 10
Workers
ID
Host
Port
0 130.102.74.134 4609
1 130.102.73.193 4902
2 130.102.73.201 4751
3 130.102.73.197 4706
4 130.102.73.182 4081
5 130.102.73.196 4555
6 130.102.73.203 4703


Mission Control
Work Units - Pending: 0 Waiting: 0 Received: 0 Cancel Pending Work Units
Sequence
Network
Phenotype
Results Database
Parameters
Seed: 12
Size: 10000
Base: 4
Length: 6
Inhib: 1
Individual
Population
genes:
Regulatory region lengths
Frequency
Length
0.0
0.2
0.4
0.6
0.8
1.0
Generate from Params
Apply Changes
Highlight Sequence
Generate Network -->
Messages
Mission Control registered on port 8000
Connection established with Worker 0
Workers
ID
Host
Port
0
127.0.0.1
4574

Work in progress

- Integration with Condor
  - Resilient to failure
  - Improved CPU monitoring
- Automated searches of parameter space
  - Nimrod/O integration



Distributed computing is worthwhile!

- Interactivity can be incorporated into distributed models of biology
- Not all models easily distributed
- It's more effort
  - Distribute design from the start
  - Machine failures, shared usage, etc.
  - Getting easier
- But can greatly expand range of feasible runs, collaborative value

Acknowledgements

- ARC Centre for Complex Systems
- ARC Centre in Bioinformatics
- Queensland Cyber Infrastructure Foundation
- Jon Kloske and Rik Taylor (ITIG)

Further Information

James Watson

jwatson@itee.uq.edu.au

http://www.itee.uq.edu.au/~jwatson/