HGC スパコン Web サービス利用法講習会
2011年10月25日 @ ヒトゲノム解析センター

大規模遺伝子ネットワーク推定ソフトウェア SiGN

玉 田 嘉 紀
tamada@is.s.u-tokyo.ac.jp

東京大学 大学院 情報理工学系研究科 コンピュータ科学専攻

理化学研究所 次世代計算科学研究開発プログラム
RIKEN Computational Science Research Program

東京大学
THE UNIVERSITY OF TOKYO

Human Genome Center
Institute of Medical Science, University of Tokyo

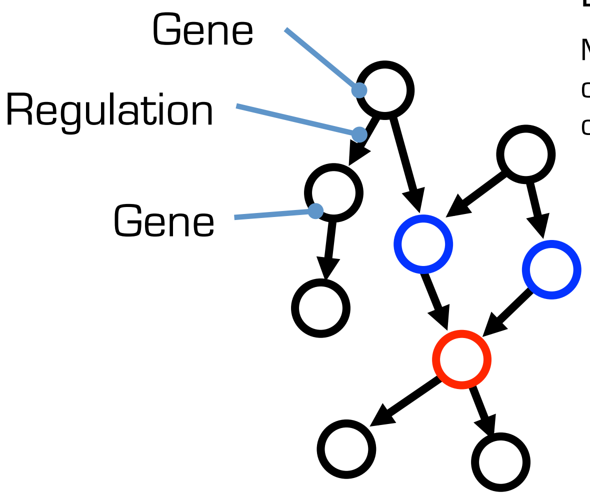
本日の内容

- 遺伝子ネットワーク推定ソフトウェアSiGN（サイン）の紹介
  - SiGN-SSM, SiGN-BN, SiGN-L1

- ベイジアンネットワークを用いたSiGN-BNの機能紹介

Gene Network Estimation

- **Gene network**: model for **transcriptome level gene–gene regulation** using directed graphs.
- Gene network estimation is to estimate gene networks from **high-throughput biological data** e.g. **gene expression data**.



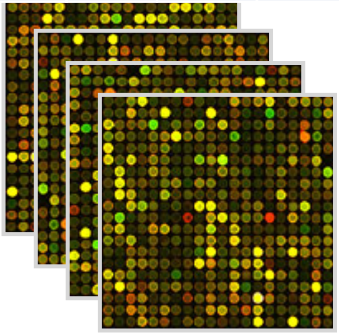
Gene Network

Directed Graph:

Mathematical model consisting of nodes and directed edges connecting them.

Estimation

Gene	KD1	KD2	KD3	…
Gene 1	1.45	-1.54	1.23	…
Gene 2	3.21	-2.1	1.44	…
…	…	…	…	…



Gene Expression Data

SiGN （サイン）

- SiGN: A collection of large-scale gene network estimation software designed for utilizing super computers.
  - SiGN-BN: Bayesian networks （ベイジアンネットワーク） 本日ベータ版公開
  - SiGN-SSM: State space models （状態空間モデル） オープンソース公開中
  - SiGN-L1: L1-regularization based models （L1正則化） 準備中

SiGN Web Site: http://sign.hgc.jp/

SiGN　（サイン）

- HGCスパコンで動くスパコン用専用ソフトウェア
- 実行にはHGCのアカウントが必要
- 実行はターミナルアプリから手でコマンド入力
  - 多少の Unix 操作の知識が必要

SiGNでできること（まとめ）

- 遺伝子発現データから遺伝子間の発現の依存関係を予測・推定する
  - 遺伝子発現データ（マイクロアレイデータ）
    - 患者サンプルから得られる細胞
    - ノックダウン実験
    - 薬剤投与などの時系列に観測したデータ
- ある程度のサンプル数が必要
  - 必要な量はモデルや実験デザインにより様々

SiGN （サイン）

- SiGN: A collection of large-scale gene network estimation software designed for utilizing super computers.
  - SiGN-BN: Bayesian networks 本日ベータ版公開
    （ベイジアンネットワーク）
  - SiGN-SSM: State space models オープンソース
    （状態空間モデル） 公開中
  - SiGN-L1: L1-regularization based models 準備中
    （L1正則化）

SiGN Web Site: http://sign.hgc.jp/

SiGN-SSM

Dynamic gene network estimation software using a State Space Model (SSM).

（状態空間モデル）

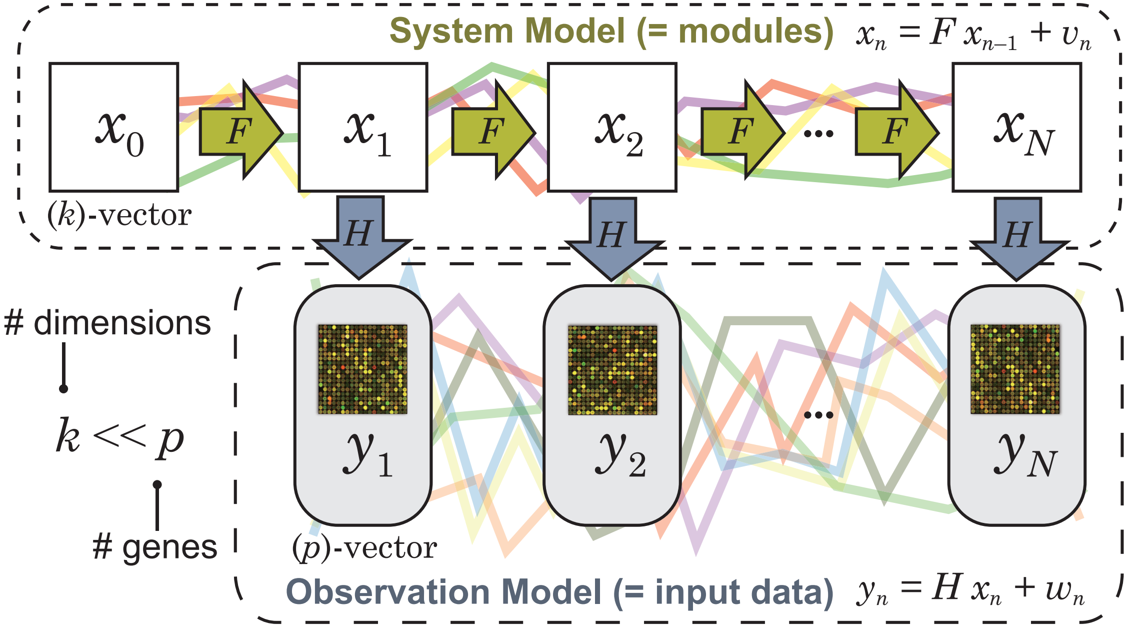
Open source software distributed at http://sign.hgc.jp/signssm/

Suitable for modeling time series gene expression data.

Definition:

$$x_n = Fx_{n-1} + v_n, \quad v_n \sim N(0,Q) \qquad \text{[System model]}$$

$$y_n = Hx_n + w_n, \quad w_n \sim N(0,R) \qquad \text{[Observation model]}$$



Parameter estimation = EM algorithm (x_0, F, Q, H, R and k)

Needs to repeat estimation with various initial values.

SiGN-SSM: 特徴

- 時系列マイクロアレイデータからの遺伝子ネットワーク推定

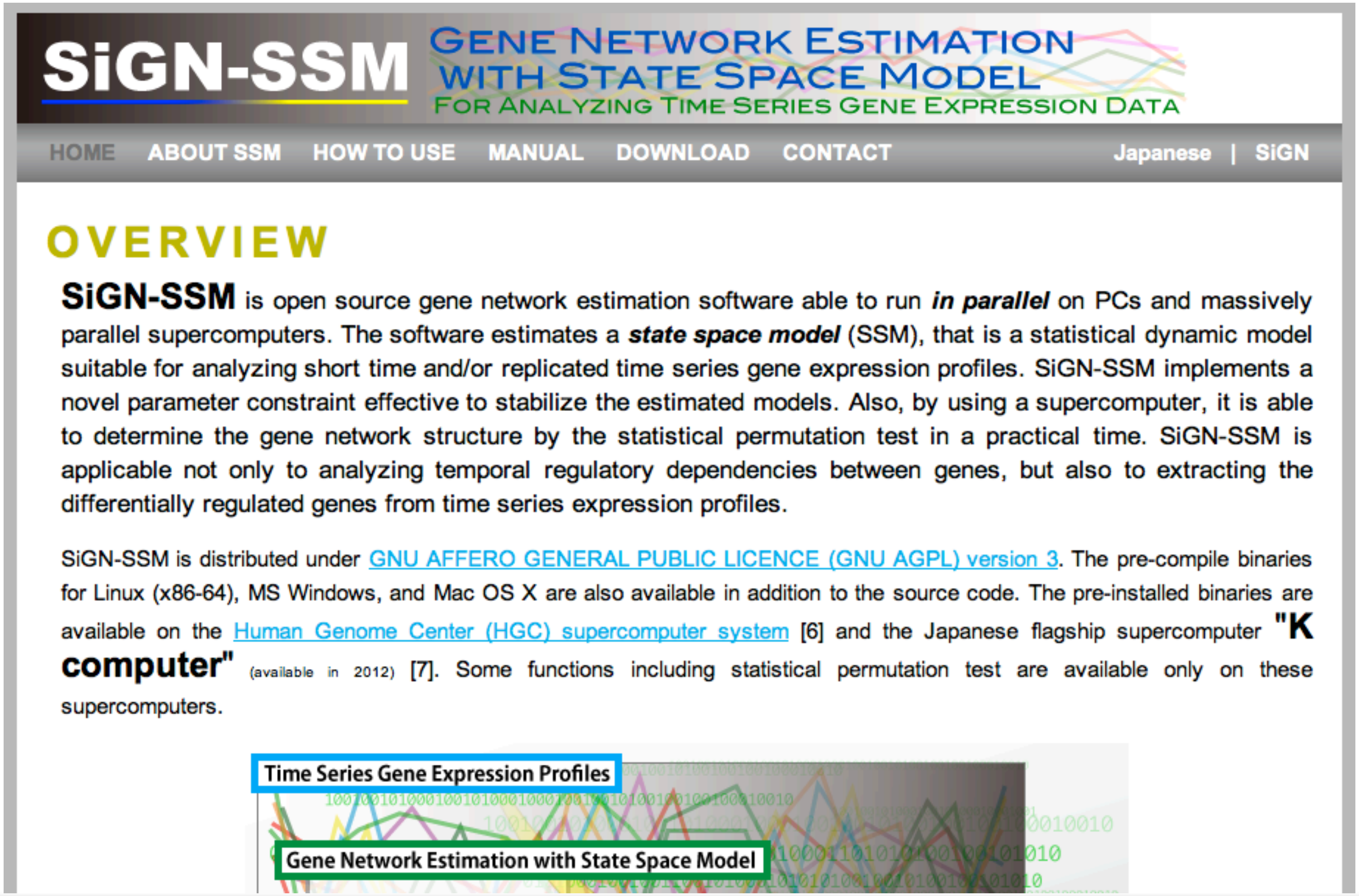
- 不等間隔・繰り返し計測データに対応

- SSMのパラメータ推定.
- 推定結果に基づきネットワーク構造の決定.

- オープンソース
  - 一部機能はHGCスパコン限定

SiGN-SSM: 詳細はウェブサイトへ

http://sign.hgc.jp/signssm/


SiGN-SSM
GENE NETWORK ESTIMATION WITH STATE SPACE MODEL
FOR ANALYZING TIME SERIES GENE EXPRESSION DATA
HOME
ABOUT SSM
HOW TO USE
MANUAL
DOWNLOAD
CONTACT
Japanese | SiGN
Time Series Gene Expression Profiles
Gene Network Estimation with State Space Model

OVERVIEW

SiGN-SSM is open source gene network estimation software able to run ***in parallel*** on PCs and massively parallel supercomputers. The software estimates a ***state space model*** (SSM), that is a statistical dynamic model suitable for analyzing short time and/or replicated time series gene expression profiles. SiGN-SSM implements a novel parameter constraint effective to stabilize the estimated models. Also, by using a supercomputer, it is able to determine the gene network structure by the statistical permutation test in a practical time. SiGN-SSM is applicable not only to analyzing temporal regulatory dependencies between genes, but also to extracting the differentially regulated genes from time series expression profiles.

SiGN-SSM is distributed under GNU AFFERO GENERAL PUBLIC LICENCE (GNU AGPL) version 3. The pre-compile binaries for Linux (x86-64), MS Windows, and Mac OS X are also available in addition to the source code. The pre-installed binaries are available on the Human Genome Center (HGC) supercomputer system [6] and the Japanese flagship supercomputer **"K computer"** (available in 2012) [7]. Some functions including statistical permutation test are available only on these supercomputers.

SiGN （サイン）

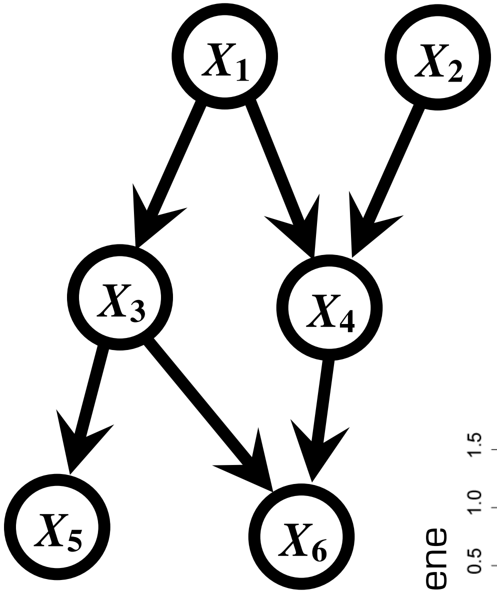
- SiGN: A collection of large-scale gene network estimation software designed for utilizing super computers.
  - SiGN-BN: Bayesian networks （ベイジアンネットワーク） 本日ベータ版公開
  - SiGN-SSM: State space models （状態空間モデル） オープンソース 公開中
  - SiGN-L1: L1-regularization based models （L1正則化） 準備中

SiGN Web Site: http://sign.hgc.jp/

Nonparametric Bayesian Network Model

We use the nonparametric Bayesian network as models for gene networks

Node = Gene



Directed Edge = Regulatory Relationships

Joint Probability by a DAG (Directed Acyclic Graph)

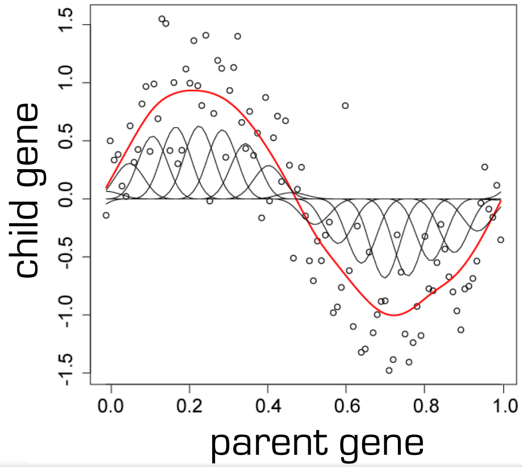
$$f(X_1,X_2,\ldots,X_6)$$

$$= f_1(X_1)f_2(X_2)f_3(X_3\mid X_1)\cdots f_6(X_6\mid X_3,X_4)$$

Network score = Posterior Probability

$$\pi(G\mid X)\propto \pi(G)\int\prod_{i=1}^{n} f(x_{i1},\ldots,x_{ip}\mid \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G\mid\lambda)d\boldsymbol{\theta}_G$$

G : gene network $\quad$ X : expression data



Nonparametric regression by B-spline

$$x_{ij}=m_{j1}(p_{i1}^{(j)})+\ldots+m_{jq_j}(p_{iq_j}^{(j)})+\varepsilon_{ij}$$

$$\varepsilon_{ij}\sim N(0,\sigma_j^2)$$

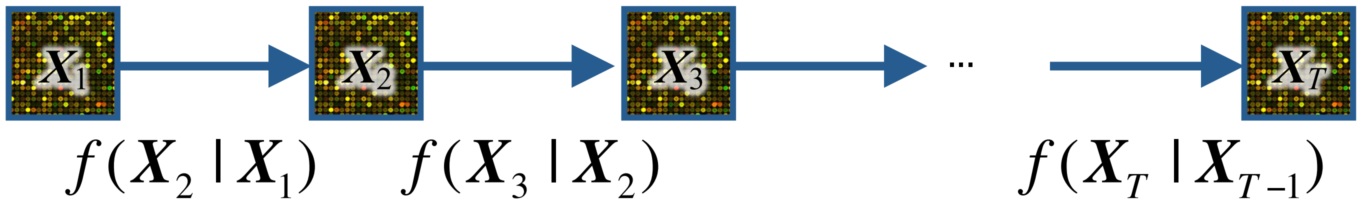
$$m_{jk}(p_{ik}^{(j)})=\sum_{l=1}^{M_{jk}}\gamma_{lk}b_{lk}^{(j)}(p_{ik}^{(j)})$$

Dynamic Bayesian Networks

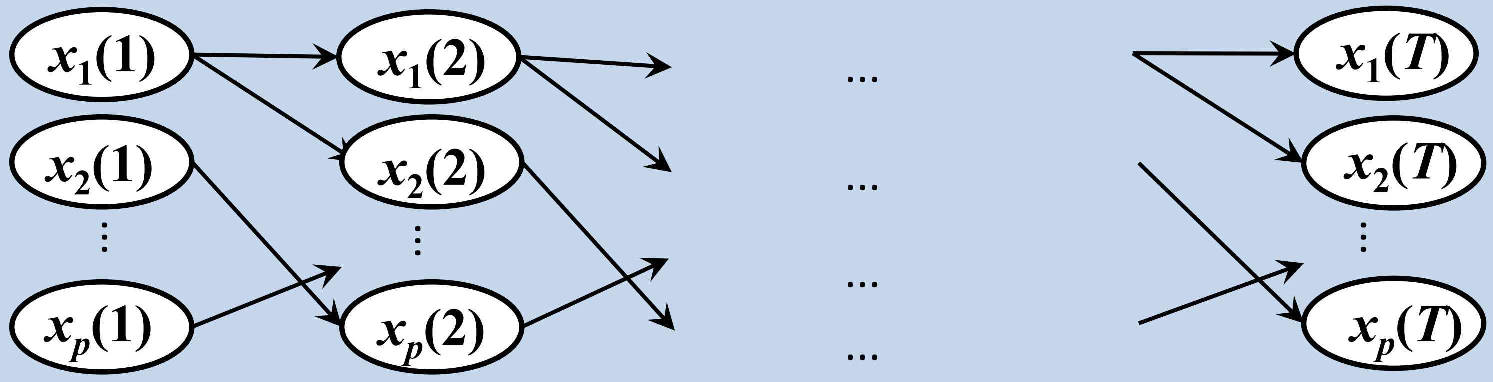
DBN (Dynamic Bayesian Network): model for time course data.

DBN assumes the dependency between consecutive time points.

Time course gene expression data: $X_1, \ldots, X_T$



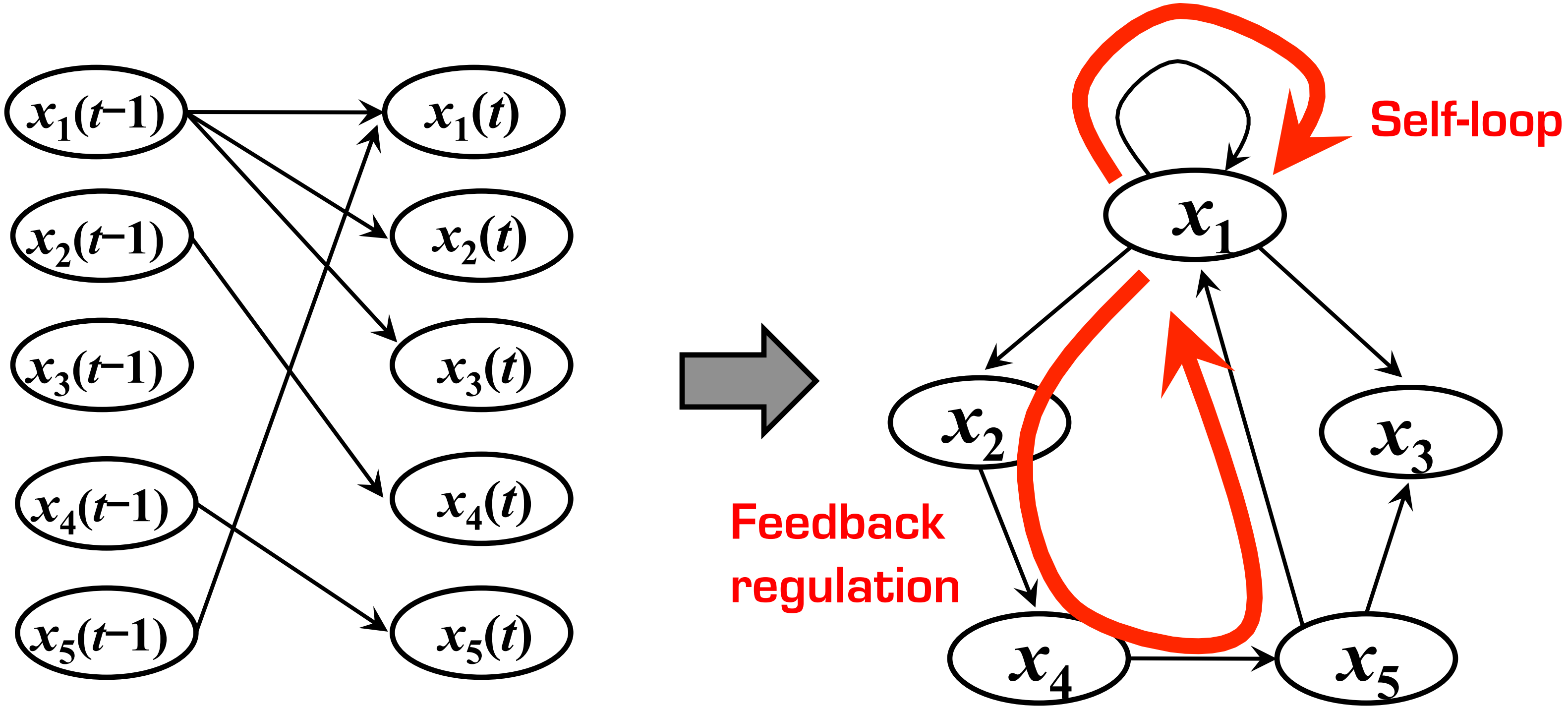
Gene Network: Bipartite Graph



Gene Network Estimation by Dynamic Bayesian Networks

Estimated Bipartite Graph

Corresponding Gene Network



Self-loop

Feedback regulation

Difficulty in Bayesian Network Estimation

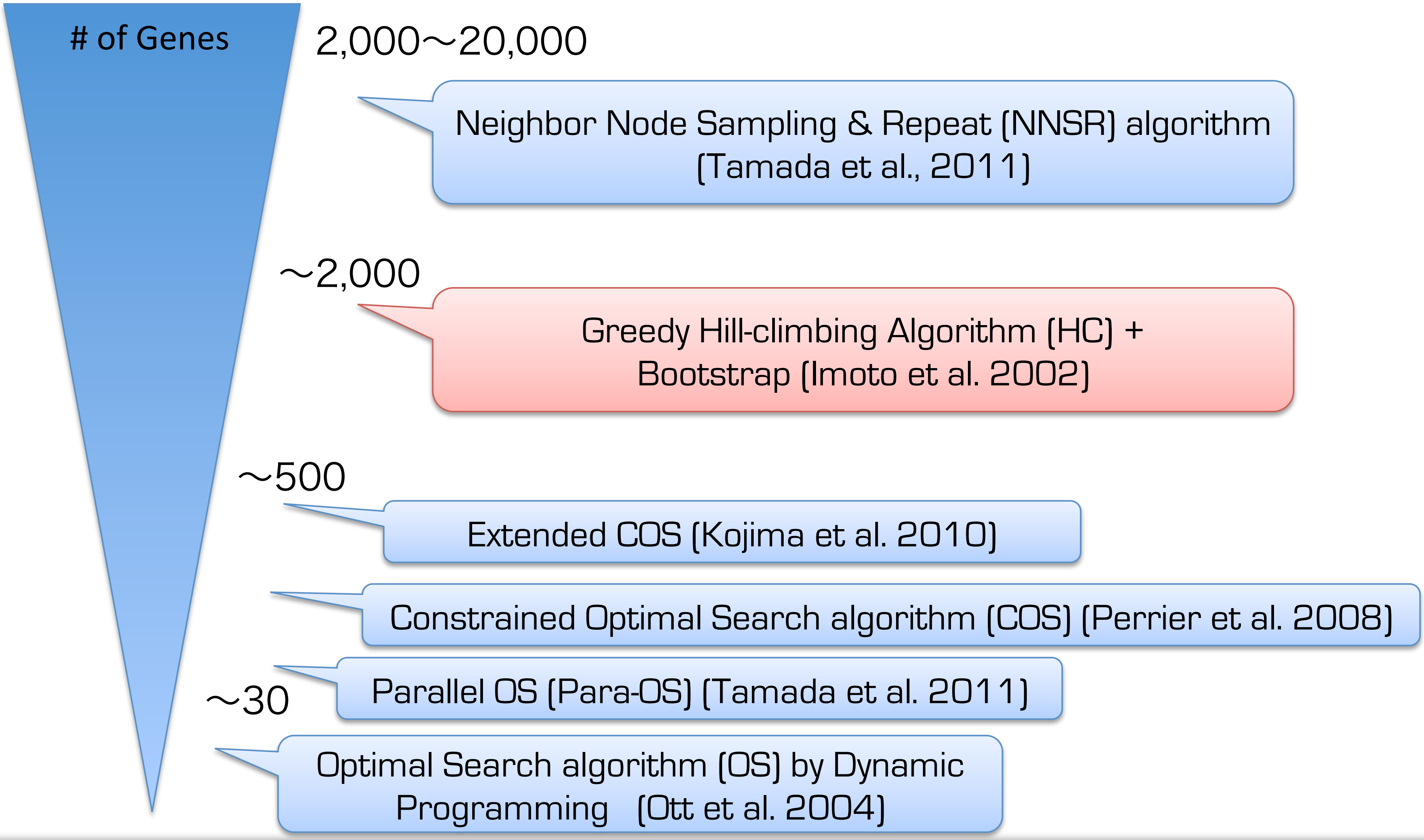
A huge number of possible DAGs : impossible to search the optimal one

# of nodes	# of DAGs	# of nodes	# of DAGs
1	1	16	$\approx 8.37 \times 10^{46}$
2	3	17	$\approx 6.26 \times 10^{52}$
3	25	18	$\approx 9.93 \times 10^{58}$
4	543	19	$\approx 3.32 \times 10^{65}$
5	29,281	20	$\approx 2.34 \times 10^{72}$
6	3,781,503	21	$\approx 3.46 \times 10^{79}$
7	1,138,779,265	22	$\approx 1.07 \times 10^{87}$
8	783,702,329,343	23	$\approx 6.97 \times 10^{94}$
9	1,213,442,454,842,881	24	$\approx 9.43 \times 10^{102}$
10	$\approx 4.17 \times 10^{18}$	25	$\approx 1.86 \times 10^{111}$
11	$\approx 3.15 \times 10^{22}$	⋯	⋯
12	$\approx 5.21 \times 10^{26}$	30	$\approx 2.71 \times 10^{158}$
13	$\approx 1.86 \times 10^{31}$	⋯	⋯
14	$\approx 1.43 \times 10^{36}$	⋯	⋯
15	$\approx 2.37 \times 10^{41}$	40	$\approx 1.12 \times 10^{276}$

Exceeds the number of atoms in the universe

Network Size and Algorithms

Different search algorithms are developed depending on the size of networks



# of Genes

2,000〜20,000

Neighbor Node Sampling & Repeat (NNSR) algorithm (Tamada et al., 2011)

〜2,000

Greedy Hill-climbing Algorithm (HC) + Bootstrap (Imoto et al. 2002)

〜500

Extended COS (Kojima et al. 2010)

Constrained Optimal Search algorithm (COS) (Perrier et al. 2008)

Parallel OS (Para-OS) (Tamada et al. 2011)

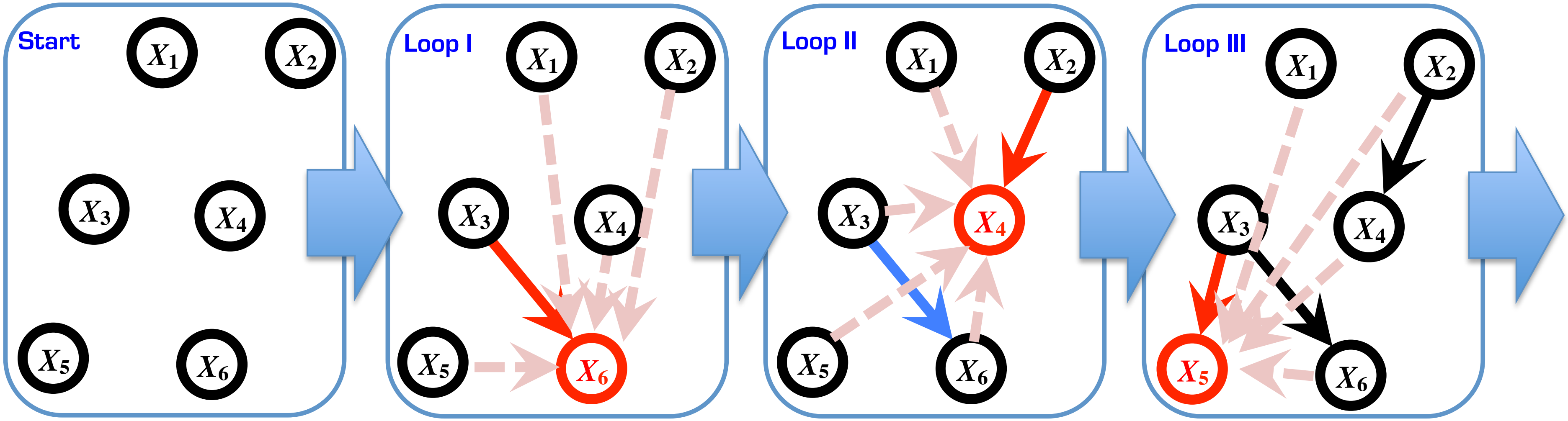
〜30

Optimal Search algorithm (OS) by Dynamic Programming (Ott et al. 2004)

Greedy Hill-Climbing Algorithm (HC)

Algorithm for searching the local optimal DAG structure

Heuristics algorithm applicable to estimate gene networks for ～ **100 genes**.



1. Begins with an empty graph.

2. Visits nodes in a random order.

3. Calculates local scores for all possible candidate parents.

4. Employs the best operation that improves the score.

Add/Delete/Reverse

5. Repeats until any operation can improve the score.

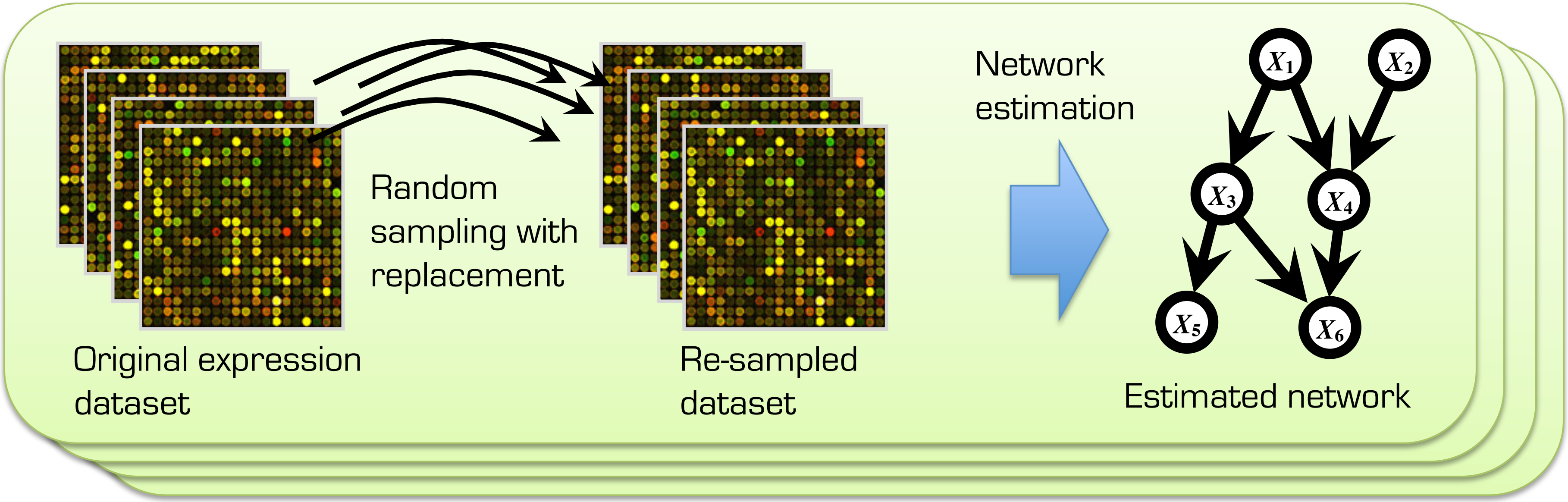
※ Need to check every time whether a cyclic path is made or not.

※ Repeats this many times, then employs the best structure because they are local optimal.

HC + Bootstrap

〜 1000 genes

Bootstrapping is required for calculating the reliability of edges.



Original expression dataset

Random sampling with replacement

Re-sampled dataset

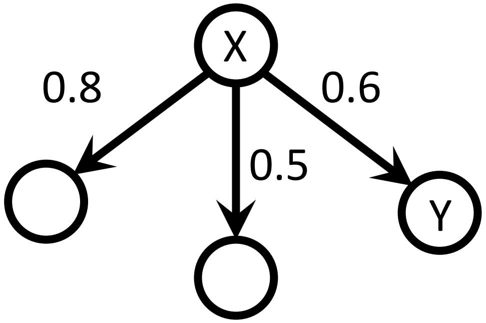
Network estimation

Estimated network

- Estimate networks **many times for re-sampled dataset.** (1,000 times 〜)
- The final structure is determined by the frequencies of edges during the repeated estimation.
- We can perform each network estimation **independently for the re-sampled datasets in parallel.**
  - **Parallelization is easy for Bootstrap HC.**

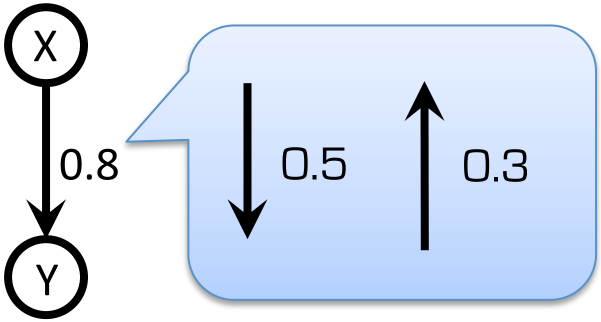
Estimated Networks

Bootstrap probabilities



- Edges in estimated networks have "Bootstrap Probability."

Direction reliabilities (Static model only)



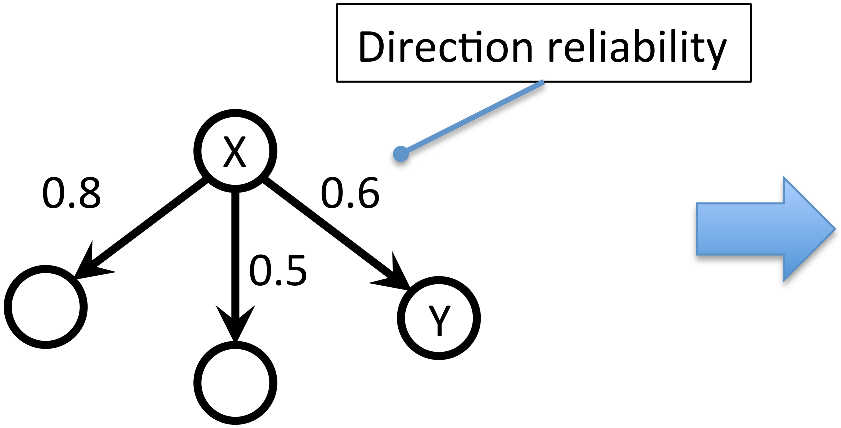
Direction reliability = 0.625

- Edges also have reliability of the direction.
  - Dynamic model can have edges in both directions.

Hubness & Hub Index

Hubness: simply counts the number of children regardless of the confidence of the edge direction.

Hub genes may have important roles as *master regulators* in the networks.
Useful for analyzing the estimated networks.



Hubness of X = 3

Hub Index of X = 1.9

Hubness of Y = 0

Hub Index of Y = 0.4

Hub Index: takes the confidence of edge direction into account. Simply the sum of “BS.Direction” for all edges connected to the target node.

Note: Hub Index does not take bootstrap probabilities into account.

実行・解析の流れ

1. 入力となるアレイデータの準備
   - テキストファイルの入力ファイル
     - EDF, タブ区切り独自形式
2. HGCスパコンへのジョブの投入
   - SGE のアレイジョブによりBootstrapの繰り返し計算を並列実行
3. 出力結果のまとめ処理
   - 結果がファイルに出力されるので，それを１つのネットワークファイルにまとめる
     - CSML, タブ区切りテキストファイル
4. Cell Illustrator Online などによる解析
   - ハブ遺伝子の分析やターゲット遺伝子関連サブネットワークの抽出・分析

SiGN-BN の公開状況

- ベータ版 (rel. 0.9.0) が `~tamada/sign` 以下にインストールされている

- 近日中に正式版およびドキュメントを公開予定
  - HC+Bootstrap 法以外の SiGN-BN のアルゴリズムや SiGN-L1 などども準備が整い次第順次HGCスパコンユーザ向けに公開予定

入力ファイル

- EDF形式によるアレイデータ
- 独自タブ区切りテキストファイル
  - 詳細は後日公開するドキュメントを参照

タブ区切りテキストファイルによる独自フォーマット

Probe_1	Gene_1	0.5	0.3	0.52	…
Probe_2	Gene_2	1.5	1.44	1.55	….
Probe_3	Gene_3	…	…	…	
….	…	…			
Probe_p	Gene_p				

実行

- HGCスパコンにログインし，SGEのqsubコマンドによりジョブを投入・実行

ブートストラップの実行

```
qsub -t 1-X [SGE options] /usr/local/bin/signbn-hc.sh
   --bs -o file_prefix [SiGN options] input_file
```

※ X はブートストラップ回数．1000以上を推奨．

結果のまとめ処理

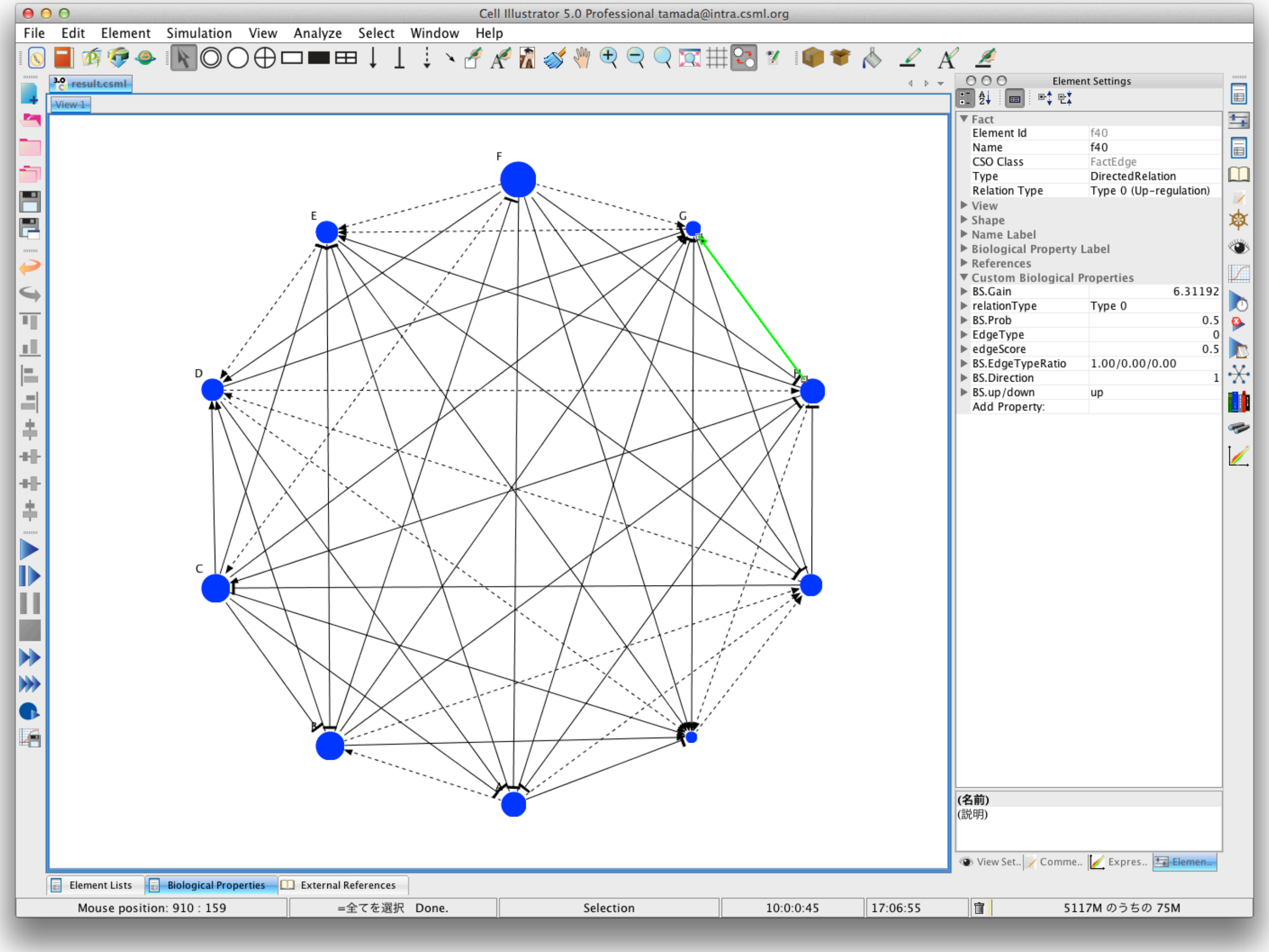
```
qsub [SGE options] /usr/local/bin/signbn-hc.sh --bin
   signproc --bs prefix=file_prefix [,other options] --output
   file=output_file,type=file_type
```

※ SGE の -t オプションは付けない

file_type は CSML か TXT

Cell Illustrator Online による解析

推定したネットワーク (CSML) は CIO で閲覧・編集・解析が可能



▼ Fact	
Element Id	f40
Name	f40
CSO Class	FactEdge
Type	DirectedRelation
Relation Type	Type 0 (Up-regulation)
▶ View	
▶ Shape	
▶ Name Label	
▶ Biological Property Label	
▶ References	
▼ Custom Biological Properties	
▶ BS.Gain	6.31192
▶ relationType	Type 0
▶ BS.Prob	0.5
▶ EdgeType	0
▶ edgeScore	0.5
▶ BS.EdgeTypeRatio	1.00/0.00/0.00
▶ BS.Direction	1
▶ BS.up/down	up
Add Property:	

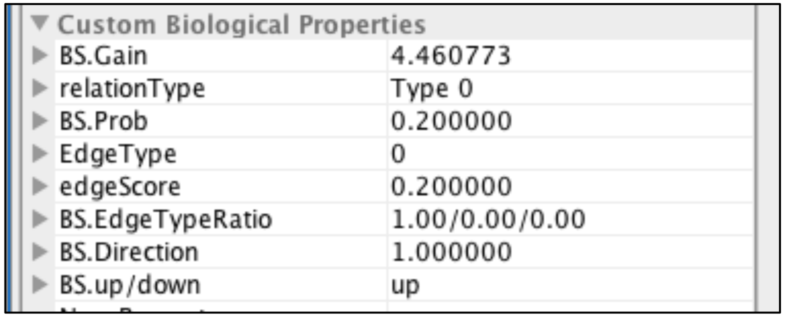
(名前)
(説明)

Mouse position: 910 : 159 | =全てを選択 Done. | Selection | 10:0:0:45 | 17:06:55 | 5117M のうちの 75M

Edge Properties

The following information is generated and assigned to each estimated edge. You can see them on CI Online.

`BS.Prob`	Bootstrap probability
`edgeScore`	Same as `BS.Prob`.
`BS.Gain`	Average of the edge gain
`BS.up/down`	One of “up”, “down”, or “unknown” that represents the estimated type of regulation.
`BS.edgeTypeRatio`	The ratio of up, down, and unknown regulation.
`BS.Direction`	Frequency (confidence) of the edge direction.



▼ Custom Biological Properties

▶ BS.Gain	4.460773
▶ relationType	Type 0
▶ BS.Prob	0.200000
▶ EdgeType	0
▶ edgeScore	0.200000
▶ BS.EdgeTypeRatio	1.00/0.00/0.00
▶ BS.Direction	1.000000
▶ BS.up/down	up

Edge properties shown in CI Online.

サンプルデータによる実習

（ベータ版）

サンプルファイル sample003.txt が ~tamada/sign/samples にあります.

(1) 出力用ディレクトリの準備と移動

```
mkdir ~/test
cd test
```

(2) SGEへのジョブの投入

赤字はベータ版でのみの指定

```
qsub -t 1-10 ~tamada/sign/signbn-hc.sh --dir ~tamada/sign --bs -o
   result ~tamada/sign/samples/sample003.txt
```

(3) SGEのジョブの確認

```
qstat
```

(4) 出力ファイルのまとめ

```
qsub ~tamada/sign/signbn-hc.sh --dir ~tamada/sign --bin signproc --bs
   prefix=result --output file=result.csml,type=csml
```

~/test に result.csml が作られます. ローカルPCに転送しCIOで開いてください.

Acknowledgments

Computational time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo, and RIKEN Integrated Cluster of Clusters (RICC) system.

Members

Teppei Shimamura, Rui Yamaguchi, Masao Nagasaki, Seiya Imoto, Satoru Miyano

Collaborators

Cristin Print, Hiromitsu Araki	The University of Auckland, NZ
D. Steve Charnock-Jones	The University of Cambridge, UK
Osamu Hirose	The University of Tokyo
Ryo Yoshida, Tomoyuki Higuchi	The Institute of Statistical Mathematics