グレブナ基底輪読
論文紹介

研究室セミナー2012/11/2

長谷川禎彦

http://metabolomics.jp/mediawiki/index.php?title=User:Aritalab/Internal/PaperReading

論文

- **A computational algebra approach to the reverse engineering of gene regulatory networks**
  - R. Laubenbacher & B. Stigler
  - J. Theor. Biol. 229 (2004) pp.523-537
- **引用回数182回**（Google scholar），96回（Scopus）

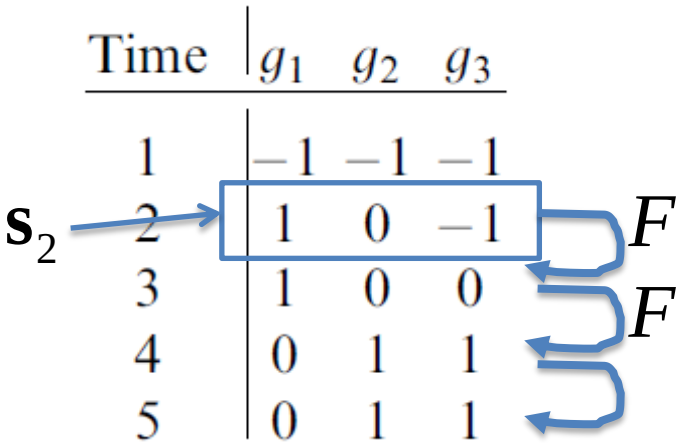
Reverse engineering of GRN

- **時系列発現量データから遺伝子間の制御・被制御関係を推定する問題**
  - e.g. Bayesian network inference

GRNのモデル化

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

- 離散モデル（濃度＋時間）
- Sを有限状態の集合とする
  - 例：S={0, 1}
- n次元なS上の力学系F（写像）　$F : S^n \to S^n$



$$F = (f_1, f_2, \ldots, f_n)$$

$$s_{j+1,i} = f_i(\mathbf{s}_j)$$

GRNのモデル化

- 目的：写像F（多項式）をデータから同定する

- 説明可能な写像Fは多く（無数に）ある
  - オッカムの剃刀
    - データを説明可能なモデルが複数ある場合，単純なモデルを選択せよ
      - BNでは，ベイズモデル選択，MDL，AIC，L1正則化
    - Over fittingを回避
  - 単純なものを一意に選ぶ　⇔　グレブナ基底の割り算の余りは一意に決まる

剰余環

$$\mathbb{Z}/p = \{0, 1, \cdots, p-1\}$$

- **Z modulo p ($\mathbb{Z}/p$) は剰余環**
  - **pが素数ならば体になる（p元体）**

$$a+b \equiv a+b \bmod p$$
$$a \times b \equiv a \times b \bmod p$$

$$-a = p - a \qquad a - a = 0$$
$$aa^{-1} = pc + 1 \bmod p = 1$$

単位元の存在，逆元の存在，乗法の逆元の存在を満たす

$Z/3$

$1^{-1} = 1, 2^{-1} = 2$

- **例 : Z modulo 2**

$$\begin{aligned} x \wedge y &= xy \\ x \vee y &= x + y + xy \\ \neg x &= x + 1 \end{aligned} \qquad \mathbb{Z}/2$$

剰余環

- **Z/4**

  $1^{-1} = 1 \in Z/4$

  2^{-1} does not exist

  - 2 * X mod 4 = 1 となるXがない

- **これからZ/4は環であっても体ではない**

- **この論文ではZ/2を数値実験で用いている**

多項式の構成アルゴリズム

- 与えられたデータを満たす方程式を生成する

$$f_i^0(\mathbf{s}_j) = s_{j+1,i}$$

となる必要があるが，なんでも良く，「最小」である必要はない（後で最小化する）

時間　インデックス

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

$$f_i^0(\mathbf{x}) = \sum_{j=1}^{m-1} s_{j+1,i} r_j(\mathbf{x})$$

$$r_j(\mathbf{x}) = \prod_{k=1}^{m-1} b_{jk}(\mathbf{x}) \quad b_{jk}(\mathbf{x}) = (s_{j,\ell} - s_{k,\ell})^{p-2}(x_\ell - s_{k,\ell})$$

l is the first coordinate in which they differ

$$r_j(\mathbf{s}_j) = 1, \quad r_j(\mathbf{x}) = 0$$

otherwise

多項式の構成アルゴリズム

$$r_j(\mathbf{x}) = \prod_{k=1}^{m-1} b_{jk}(\mathbf{x}) \qquad b_{jk}(\mathbf{x}) = (s_{j,\ell} - s_{k,\ell})^{p-2}(x_\ell - s_{k,\ell})$$

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

フェルマーの小定理になっている

$a^{p-1} = 1 \pmod{p}$

$$r_2 = b_{21} \cdot b_{23} \cdot b_{24}$$

$$b_{21} = (1-(-1))^{p-2}(x_1+1), \quad \ell = 1$$

$$b_{23} = (-1-0)^{p-2}(x_3-0), \quad \ell = 3$$

$$b_{24} = (1-0)^{p-2}(x_1-0), \quad \ell = 1$$

$$f_i^0(\mathbf{x}) = \sum_{j=1}^{m-1} s_{j+1,i} r_j(\mathbf{x})$$

$$r_2(\mathbf{s}_2) = 1 \qquad r_2(\mathbf{x}) = 0 \quad \text{otherwise}$$

$$f_i^0(\mathbf{s}_2) = s_{3,i}, \quad i = 1,2,3$$

写像Fを一意に決める

- 先ほど求めた関数Fはとても冗長
- 仮に説明可能な関数が二つあるとする

$$f_i(\mathbf{s}_j) = s_{j+1,i} = h_i(\mathbf{s}_j)$$

$$f_i(\mathbf{s}_j) - h_i(\mathbf{s}_j) = g(\mathbf{s}_j) = 0$$

- 二つの関数の差gは，データ上で常に0となる関数分だけしか違わない
  - 常に0となる関数部分を除けば一意に決まる

$\mathbf{s}_3$

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

写像Fを一意に決める

- 集合Iをデータ上で常に0である(消失する)関数の集合とする
- 可能な全ての関数(写像)の集合は

$$f_i + I := \{ f_i + g \mid g \in I \}$$

- 求めるべき最小の関数は，集合Iでこれ以上割ることの出来ない関数 → Iで割った余り

$$f_i = \overline{f_i^0}^I$$

一般の多項式集合で割ると余りは一意でない

→ グレブナ基底を使う

データ上で消失する，非0な多項式

- **データ上で常に0である，非0多項式**
  - **時間 i で消失するイデアル**　　$h(\mathbf{s}_t) = 0 \Leftrightarrow h \in I_t$

$$I_i = \langle x_1 - s_{i1}, x_2 - s_{i2}, \cdots, x_n - s_{in} \rangle$$

  - **データ全体で消失するイデアル**

$$I = \bigcap_{i=1}^{m} I_i$$

  - **イデアルIのグレブナ基底を計算**

注：イデアルの積イデアルは本の4章§3で出てくる．自明な方法では出ない．

$I_2 = \langle x_1 - 1, x_2, x_3 + 1 \rangle$

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

例の場合 $\mathbb{Z}/3$

$$I_1 = \langle x_1 + 1, x_2 + 1, x_3 + 1 \rangle$$

$$I_2 = \langle x_1 - 1, x_2, x_3 + 1 \rangle$$

$$I_3 = \langle x_1 - 1, x_2, x_3 \rangle$$

$$I_4 = I_5 = \langle x_1, x_2 - 1, x_3 - 1 \rangle$$

Time	g_1	g_2	g_3
1	-1	-1	-1
2	1	0	-1
3	1	0	0
4	0	1	1
5	0	1	1

積イデアル

$$\bigcap_{t=1}^{5} I_t = \langle 2\,x_2 + x_2^3, 2\,x_2^2 + x_3\,x_2, 2 + x_1 + x_2, x_3 + 2\,x_2 + 2\,x_2^2 + x_3^2 \rangle$$

グレブナ基底

$$\langle 2x_3 + x_3^3, x_2 + 2x_3 + 2x_3^2 + x_3x_2, x_2 + 2x_3 + 2x_3^2 + x_2^2, 2 + x_1 + x_2 \rangle$$

$$x_n^3 = x_n, 3x_n^m = 0$$

$$\langle x_2 + 2x_3 + 2x_3^2 + x_3x_2, x_2 + 2x_3 + 2x_3^2 + x_2^2, 2 + x_1 + x_2 \rangle$$

離散時系列データを用意

データを満たす多項式写像Fを生成

時系列データ上で常にゼロとなる
ノンゼロ多項式のイデアルを生成

イデアルのグレブナ基底
を計算

多項式Fをグレブナ基底で割り，
余りを出力する

Time	g_1	g_2	g_3
1	−1	−1	−1
2	1	0	−1
3	1	0	0
4	0	1	1
5	0	1	1

$$f_1^0 = x_1^2x_3 - x_1^2 + x_1x_3 + x_1,$$
$$f_2^0 = -x_1^2x_3 + x_1^2 - x_1x_3 - x_1 + 1,$$
$$f_3^0 = -x_1^2x_3 - x_1^2 - x_1x_3 + x_1 + 1.$$

時間jのデータから
j+1のデータを出す
写像

データ上で常に
消失する多項式イデアル

$$\langle 2\,x2 + x2^3,\ 2\,x2^2 + x2\,x3,\ 2 + x1 + x2,\ x3 + 2\,x2 + 2\,x2^2 + x3^2 \rangle$$

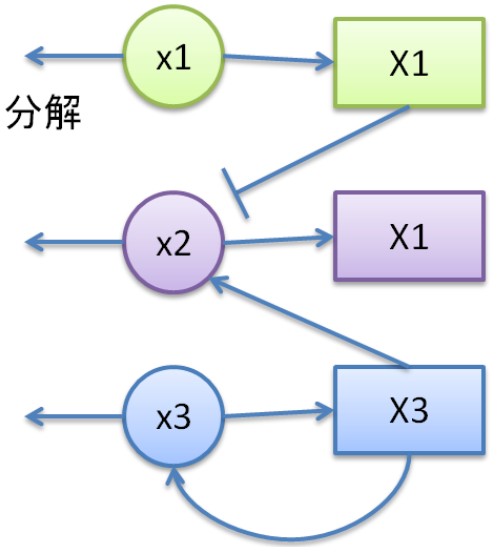
$$I = \langle x_1 + x_2 - 1, x_2x_3 - x_3^2 + x_2 - x_3,$$
$$x_2^2 - x_3^2 + x_2 - x_3 \rangle.$$

データ上で常に
消失する多項式イデアル
のグレブナ基底

$$f_1 = -x_3^2 + x_3,$$
$$f_2 = x_3^2 - x_3 + 1,$$
$$f_3 = -x_3^2 + x_2 + 1.$$

アルゴリズムの直感的要約

時系列を説明する多項式の写像Fの中で，常に0となる多項式を含まない，最小の多項式写像Fを求めている



t=1 t=2

$s_{1,1}$ $s_{2,1}$

$s_{1,2}$ $\xrightarrow{F}$ $s_{2,2}$

$s_{1,3}$ $s_{2,3}$

$$F = (f_1, f_2, \ldots, f_n)$$

$$s_{j+1,i} = f_i(\mathbf{s}_j)$$

計算量

- m : 時系列のデータ数
- n : ノード数
- p : Sの状態数

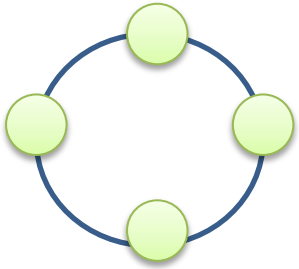
- nに対してquadratic
- mに対してexponential

評価

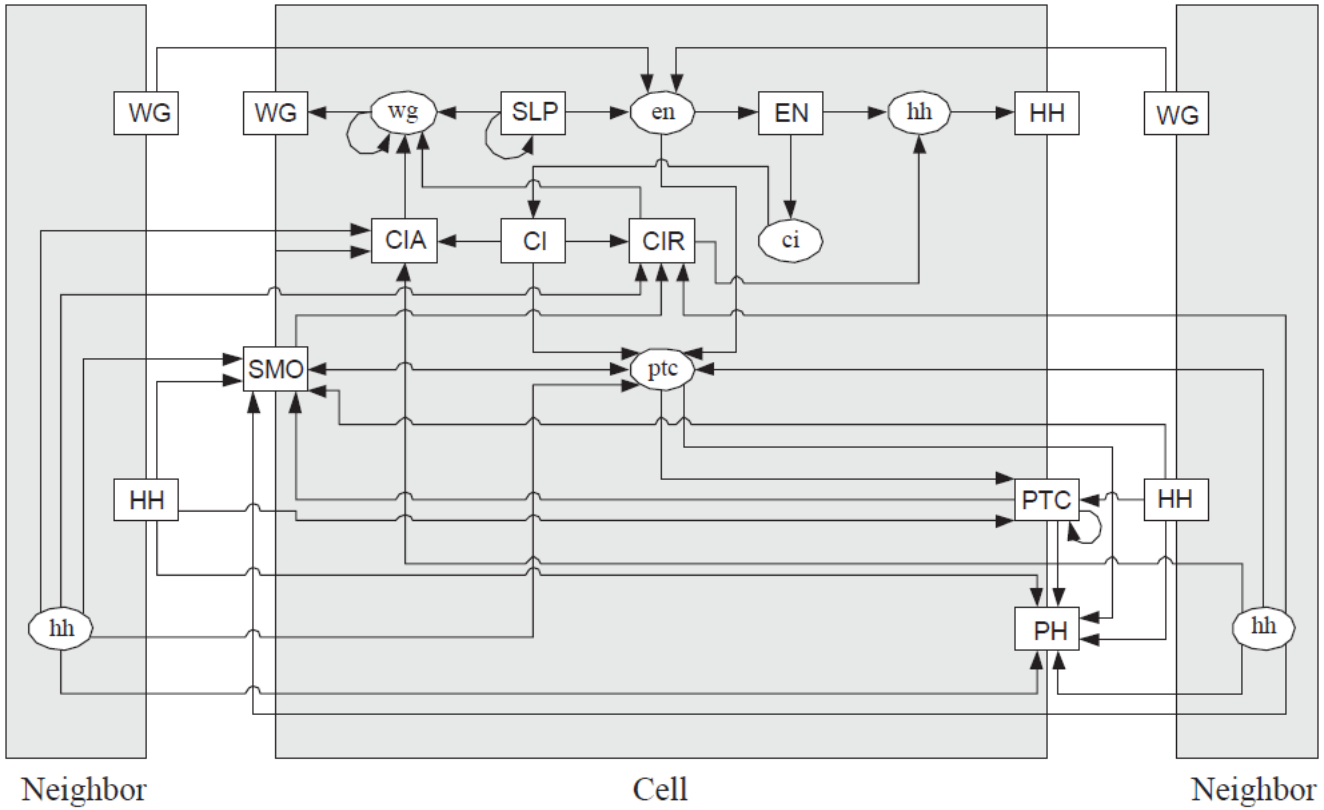
- 実際の発現量データでは行っていない
- キイロショウジョウバエ胚発生GRNの既知のBoolean networkを利用
  - Albert (2003)らのネットワーク
- 既知のBoolean networkより学習データを生成

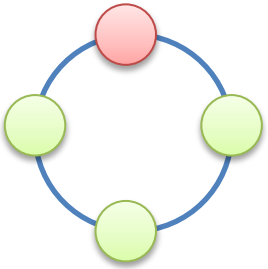
データの詳細

- 4つの細胞集団
- 5遺伝子＋それらのタンパク質
  - 15 molecular species
- 5遺伝子の初期条件はAlbertに従う
  - タンパク質は一個前のmRNAに従う
- 人工ノックアウトデータも生成
  - WT＋ノックアウト＝6種類のデータ
- 時間方向に8ステップ
  - 計6*(8-1)=42ポイントのBooleanデータ



学習元のネットワーク


WG
WG
wg
SLP
en
EN
hh
HH
WG
CIA
CI
CIR
ci
SMO
ptc
HH
PTC
HH
hh
PH
hh
Neighbor
Cell
Neighbor



実験方法

- 4つの細胞で同じなので，一つのみのネットワーク構造に注目する
- 4種類の単項式順序を用いる
  - x1>…>x21, x21>…>x1, その他二つ
- WT＋ノックアウトデータにおいて，違う単項式順序で得られたネットワークのIntersectionを計算する
  - Intersection = ANDのネットワーク

結果

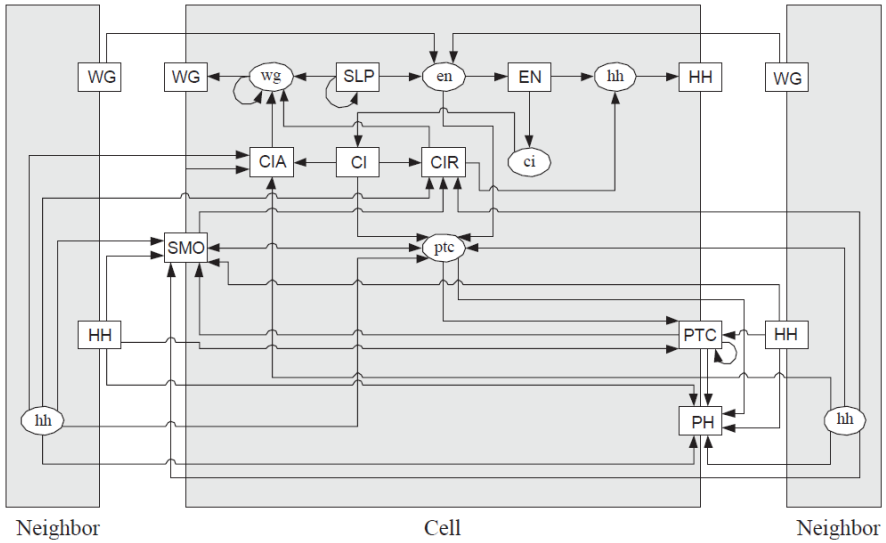
- WTデータのみ
  - 14 intracellular links (44 links in data)
- KO+WTデータ
  - 27 intracellular links (44 links in data)
- *wg*, *ci*, SLP, WG, EN, HH, PTCの制御を正しく推定

結果

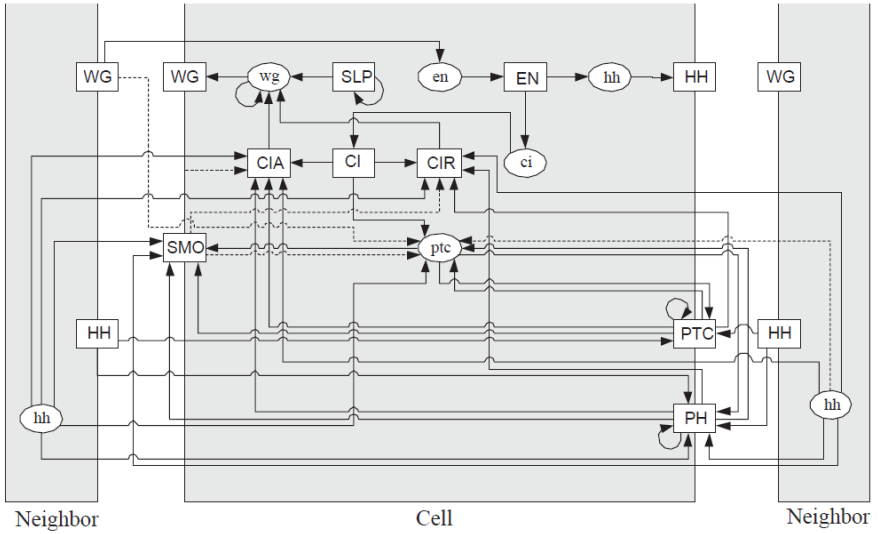
- 84%のトポロジーを推定

4順序で出現

3順序で出現



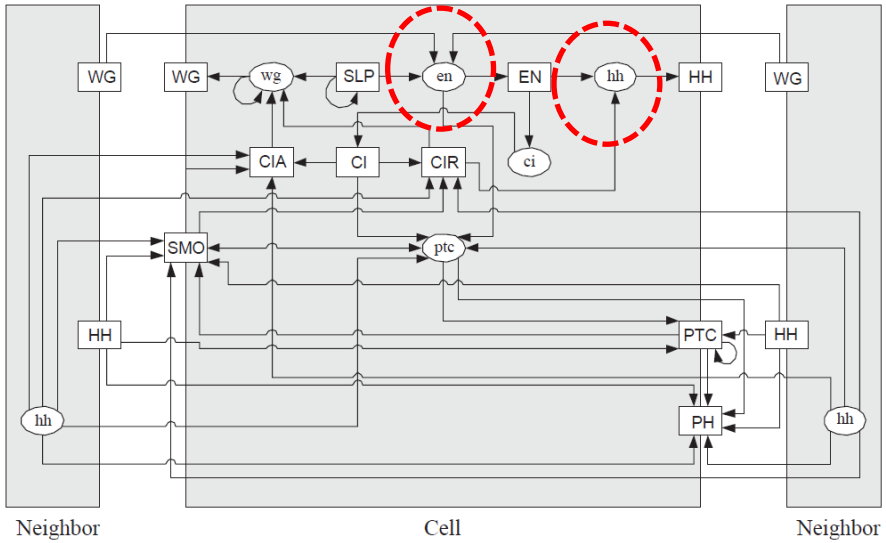
学習データに使った
ネットワーク



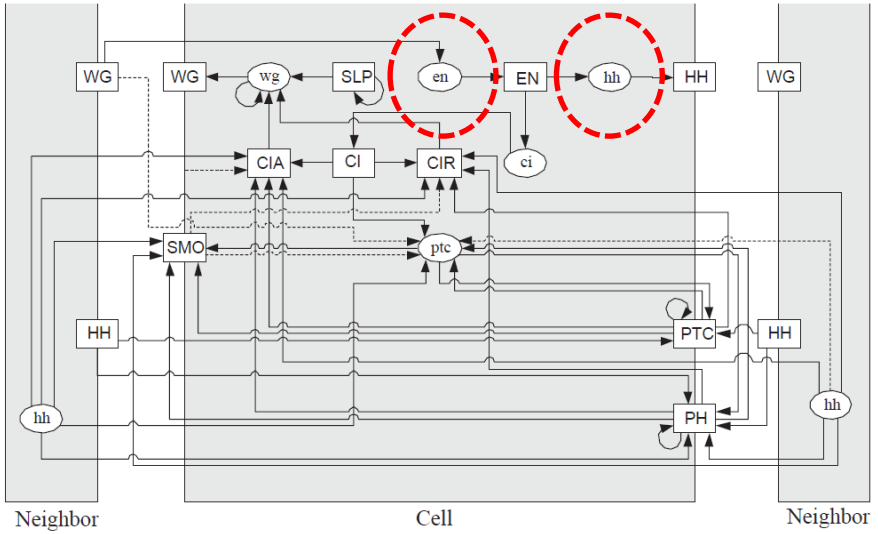
学習されたデータ

結果

- 余分なものは学習していない



学習データに使った
ネットワーク



学習されたデータ

Boolean dynamics

- もとのBoolean関数から，代数写像を計算
  - データ上で常に0となるものはどのようなアルゴリズムでも学習できない
- 計算した写像をデータ上で0になるイデアルで割った関数と，ブールデータから推定した写像を比較

Boolean dynamics

Table 5
Boolean functions for the network $\mathcal{N}$ in one cell of the ring

$$f_1 = SLP_i^{t+1} = \begin{cases} 0, & \text{if } i \bmod 4 = 1 \text{ or } i \bmod 4 = 2 \\ 1, & \text{if } i \bmod 4 = 3 \text{ or } i \bmod 4 = 0 \end{cases}$$
$$f_2 = wg_i^{t+1} = (CIA_i^t \wedge SLP_i^t \wedge \neg CIR_i^t) \vee (wg_i^t \wedge (CIA_i^t \vee SLP_i^t) \wedge \neg CIR_i^t)$$
$$f_3 = WG_i^{t+1} = wg_i^t$$
$$f_4 = en_i^{t+1} = (WG_{i-1}^t \vee WG_{i+1}^t) \wedge \neg SLP_i^t$$
$$f_5 = EN_i^{t+1} = en_i^t$$
$$f_6 = hh_i^{t+1} = EN_i^t \wedge \neg CIR_i^t$$
$$f_7 = HH_i^{t+1} = hh_i^t$$
$$f_8 = ptc_i^{t+1} = CIA_i^{t+1} \wedge \neg EN_i^{t+1} \wedge \neg CIR_i^{t+1}$$
$$f_9 = PTC_i^{t+1} = ptc_i^t \vee (PTC_i^t \wedge \neg HH_{i-1}^t \wedge \neg HH_{i+1}^t)$$
$$f_{10} = PH_i^{t+1} = PTC_i^{t+1} \wedge (HH_{i-1}^{t+1} \vee HH_{i+1}^{t+1})$$
$$f_{11} = SMO_i^{t+1} = \neg PTC_i^{t+1} \vee HH_{i-1}^{t+1} \vee HH_{i+1}^{t+1}$$
$$f_{12} = ci_i^{t+1} = \neg EN_i^t$$
$$f_{13} = CI_i^{t+1} = ci_i^t$$
$$f_{14} = CIA_i^{t+1} = CI_i^t \wedge (SMO_i^t \vee hh_{i-1}^t \vee hh_{i+1}^t)$$
$$f_{15} = CIR_i^{t+1} = CI_i^t \wedge \neg SMO_i^t \wedge \neg hh_{i-1}^t \wedge \neg hh_{i+1}^t$$

答えのbool func

関数にしてイデアルIで割る

Table 6
Polynomial representations of the Boolean functions in Table 4, together with the legend of variable names

$$f_1 = x_1$$
$$f_2 = (x_{15} + 1)(x_1x_{14} + x_2(x_1 + x_{14} + x_1x_{14}) + x_1x_2x_{14}(x_1 + x_{14} + x_1x_{14}))$$
$$f_3 = x_2$$
$$f_4 = (x_{16} + x_{17} + x_{16}x_{17})(x_1 + 1)$$
$$f_5 = x_4$$
$$f_6 = x_5(x_{15} + 1)$$
$$f_7 = x_6$$
$$f_8 = x_{13}((x_{11} + x_{20} + x_{11}x_{20}) + x_{21} + (x_{11} + x_{20} + x_{11}x_{20})x_{21})(x_4 + 1)$$
$$(x_{13}(x_{11} + 1)(x_{20} + 1)(x_{21} + 1) + 1)$$
$$f_9 = x_8 + x_9(x_{18} + 1)(x_{19} + 1) + x_8x_9(x_{18} + 1)(x_{19} + 1)$$
$$f_{10} = (x_8 + x_9(x_{18} + 1)(x_{19} + 1) + x_8x_9(x_{18} + 1)(x_{19} + 1))(x_{20} + x_{21} + x_{20}x_{21})$$
$$f_{11} = x_8 + x_9Y + x_8x_9Y + 1 + x_{20} + ((x_8 + x_9Y + x_8x_9Y + 1)x_{20}) + x_{21}$$
$$+(x_8 + x_9Y + x_8x_9Y + 1 + x_{20} + (x_8 + x_9Y + x_8x_9Y + 1)x_{20})x_{21}$$
$$f_{12} = x_5 + 1$$
$$f_{13} = x_{12}$$
$$f_{14} = x_{13}((x_{11} + x_{20} + x_{11}x_{20}) + x_{21} + (x_{11} + x_{20} + x_{11}x_{20})x_{21})$$
$$f_{15} = x_{13}(x_{11} + 1)(x_{20} + 1)(x_{21} + 1)$$

Boolean dynamics

- 依存関係（ノード間の繋がり）は完全に一致（全単項式順序）

Table 7
Boolean functions reduced by the ideal of wildtype and knock-out time series

$$
\begin{aligned}
f_1 &= x_1\\
f_2 &= x_1x_{14} + x_2x_{14} + x_2x_{15} + x_2\\
f_3 &= x_2\\
f_4 &= x_{16}\\
f_5 &= x_4\\
f_6 &= x_5\\
f_7 &= x_6\\
f_8 &= x_{12}x_{13} + x_{13}x_{16} + x_{13}x_{20} + x_{18}x_{20} + x_{13}x_{21} + x_{19}x_{21} + x_{13} + x_{18} + x_{19}\\
f_9 &= x_{10}x_{14} + x_{14}x_{18} + x_{14}x_{19} + x_9x_{20} + x_{18}x_{20} + x_9x_{21} + x_{19}x_{21} + x_8 + x_9 + x_{10}\\
f_{10} &= x_{10}x_{14} + x_{14}x_{18} + x_{14}x_{19} + x_8x_{20} + x_8x_{21} + x_{10} + x_{18} + x_{19}\\
f_{11} &= x_8x_{20} + x_9x_{20} + x_{18}x_{20} + x_8x_{21} + x_9x_{21} + x_{19}x_{21} + x_8 + x_9 + x_{18} + x_{19} + 1\\
f_{12} &= x_5 + 1\\
f_{13} &= x_{12}\\
f_{14} &= x_{11}x_{13} + x_9x_{20} + x_{18}x_{20} + x_9x_{21} + x_{19}x_{21} + x_{10} + x_{18} + x_{19}\\
f_{15} &= x_{11}x_{13} + x_9x_{20} + x_{18}x_{20} + x_9x_{21} + x_{19}x_{21} + x_{10} + x_{13} + x_{18} + x_{19}
\end{aligned}
$$

学習データから
推定された写像

Boolean dynamics

Table 4
Performance of dynamics detection for one cell of $\mathcal{N}$

Total single interactions in $\mathcal{N}$		13
Total cooperative interactions in $\mathcal{N}$		30
Single interactions	4 TO	3 TO
Total predicted	18	21
True positives	12	12
False positives	6	9
Cooperative interactions	4 TO	3 TO
Total predicted	3	11
True positives	3	8
False positives	0	3

Single interactions = degree-one terms; cooperative interactions = degree-two terms. 4 TO denotes results for all 4 term orders used, whereas 3 TO denotes results for any 3 of the 4 term orders used.

Discussion

- データの離散処理
  - pを大きくすると影響は少なくなる
- ノイズの影響
  - 1%のノイズを離散データに追加