計算機の中の生命科学

– ゲノム･ポストゲノムの解析 –

計算科学研究センター発足シンポジウム

平成16年6月11日

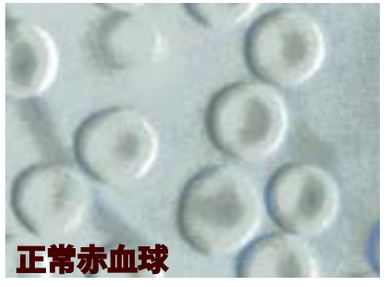
筑波大学生命環境科学研究科　漆原秀子

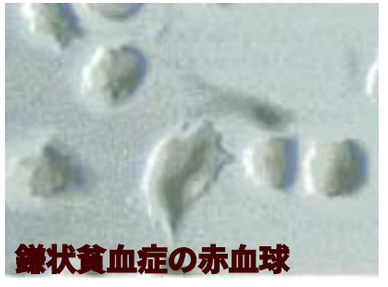
本日の内容

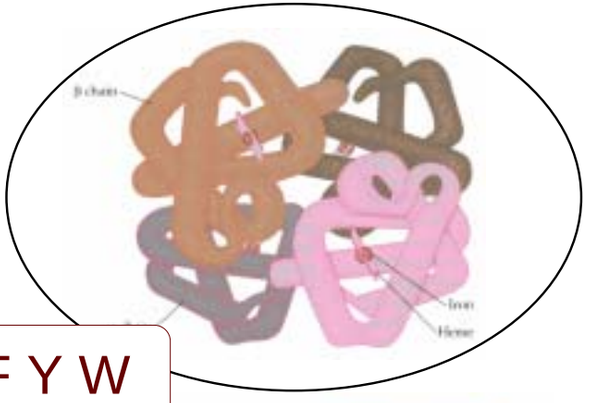
- 生物遺伝情報体系の概略
- ゲノム解析において大規模な計算が実行されている例
- ポストゲノム解析で計算に苦闘している例
- 生命現象のシミュレーションへの期待と模索

生物の遺伝情報システム

形 質







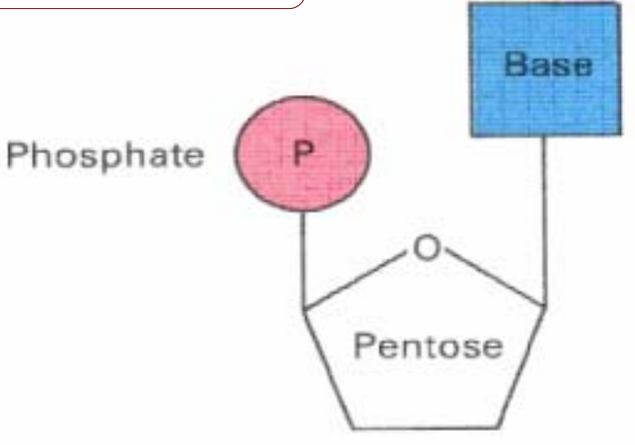
タンパク質 C S T P A G N D E Q H R K M I L V F Y W

翻訳（Translation）

RNA A G U C

転写（Transcription）

DNA A G T C

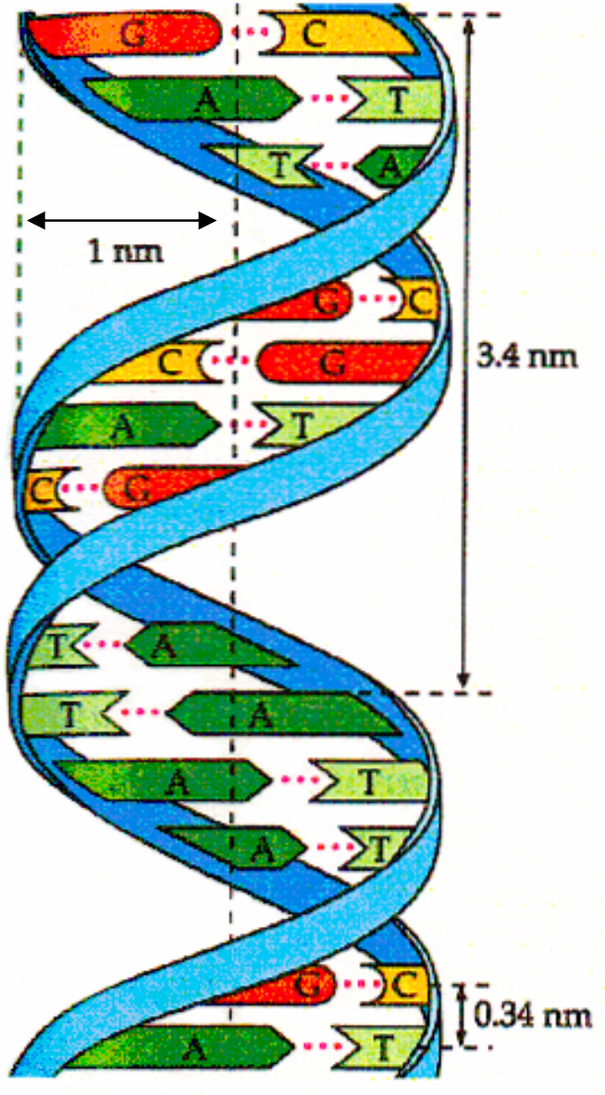


ヌクレオチド・塩基・base (b)・base pair (bp)

ゲノム：その生物に固有な遺伝情報の１セット；DNA**または**RNA

生物のゲノムサイズ (例)

生物種	ゲノムサイズ (Mb)
トウモロコシ	5000
ヒト	3000
フグ	400
ショウジョウバエ	140
線虫	100
細胞性粘菌	34.0
酵母	12.1
大腸菌	4.64



GGTAATAATAATAATGGTATAAGTGATATTCAACTTAGAACTGATTTTGA
ATTAATGGGTAAAAAATTAGGTGAACTTGGTACAGAAAATTATAAATTAG
ATGAAAGAAATAGAACATTAGAATTAAAGAATAGGCAATTAACAGAAGAT
TTGGAAAAGAAATCAAATGCAGTTAGATTATTAGTTTCAAAAACTCAACT
TGGTAAAGCAACCTCTGAAGAAGAAAAAGCTAAAAAAATTAAAAAGTGGT
GGATTTATGGGAAGTTTTTGGAGAAATAATGATCCAAAGATTGCTGCTGA
AATGGTCGAAAAAATGGAAGTTATGTTACAAGAGAATGTTCTTAAAAATT
TCCAACTTTCAAATGATTTAGAATTATTAGGTACTGAAACTGGAAAATTA
AAATCTCAATTATCAATTTATGAATCAATTTTAAAAGAAAATAATATCGA
AAAACCAACTTTTAAAAATTTACATGAAAATTTAAAACCAAATAATTTTG
CACTACTTCAATTCAAATTGGAAGTGAAGAAAATTATAAAATTTAATTAT
TATTATAATATTTTAAAATAAATTAAAAAAAATAAATAAAAAAAACTATA
TATATTATATATTAAATTAAAAACAAATAAATAAATAAAAAAGTAAAAAA
TTCAATTGAAAATAATAATAAAAATAATAATAAAAATAATAATAAGATTT
GGTTAATCACAGGAACATCAAGTGGCATTGGTTTAGAATTAGTAAAGAAG
TTATTAACATATGGTTATAAAGTTTCAGCATTAACTCGTAGACCAGAAGA
AATTGAAAAAGAGATTAAAGAAATTCAATTTGAAAAAGATAATTTATTAA
TTGTTAAAACTGATATTACAAATAATGAATCAGTAAAGAGTGCAGTAGAA
GAGACTATTAAACAGTTTGGTAGAATTGATGTGTTAGTAAATAACGCTGG
AAAAAAAAAAAAAAAAAAABGACTAGTTCTAGATCGCGAGCGGCCGACC

3 Gb = 1000 bases x 10 slides/sec x 4 days

ゲノム解析過程における計算（1）

- DNA断片の配列をアセンブルする
  - 配列決定は技術的制約により、たかだか0.5〜1 Kb程度の断片としてしか行えない
  - この断片を正しく並べて10〜100Mbのゲノム配列を得る

ゲノム解析過程における計算（1）

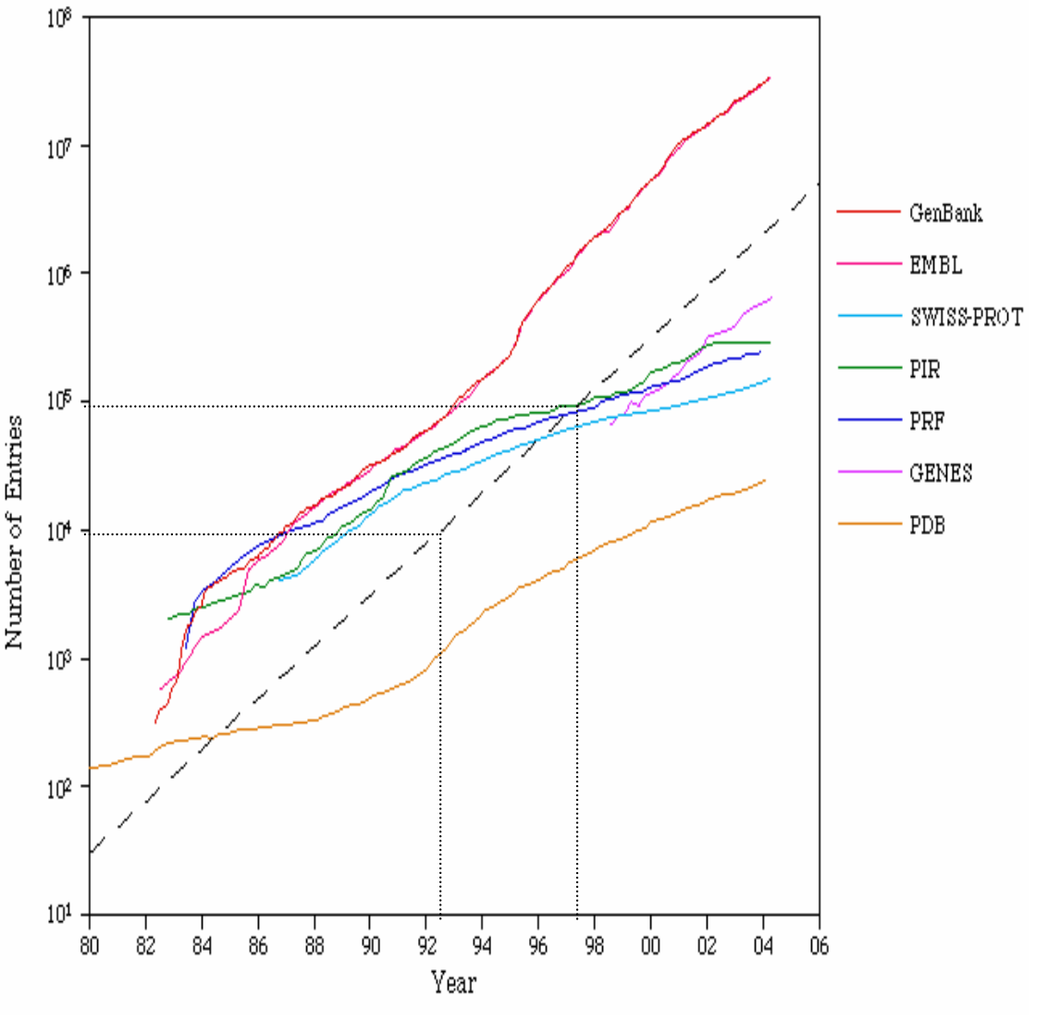
- DNA断片の配列をアセンブルする
  - 配列決定は技術的制約により、たかだか0.5〜1 Kb程度の断片としてしか行えない
  - この断片を正しく並べて10〜100Mbのゲノム配列を得る

ゲノム解析過程における計算（1）

- DNA断片の配列をアセンブルする
  - 配列決定は技術的制約により、たかだか0.5〜1 Kb程度の断片としてしか行えない
  - この断片を正しく並べて10〜100Mbのゲノム配列を得る
  - 通常10倍程度に重複させるので、10^6近い配列を同時に処理することとなり、大メモリの計算機が必要となる

ゲノム解析過程における計算(2)

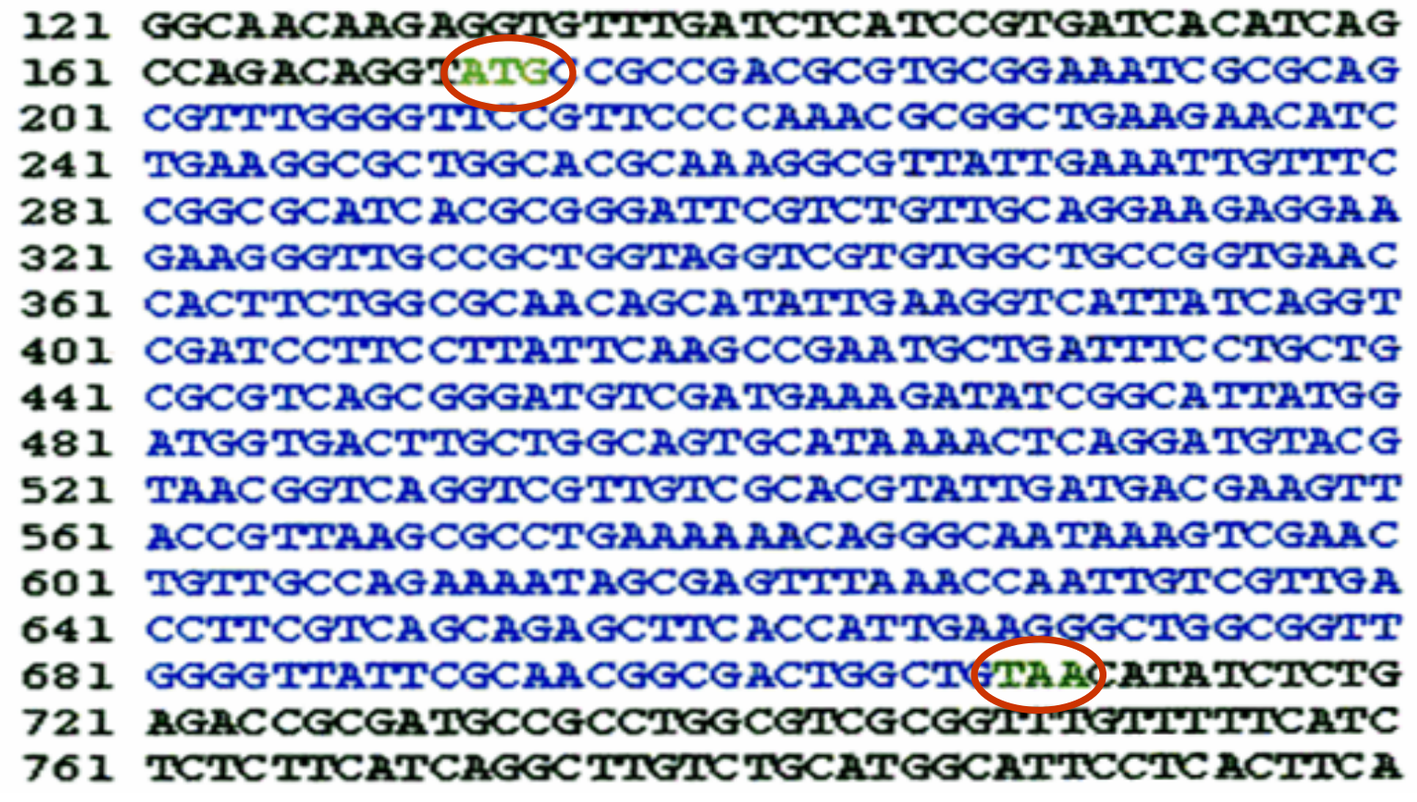
- アノテーション(説明づけ)
  - 既知配列のデータベースに対して相同性検索を行い、スコアの最も高い配列の情報を得る
  - ダイナミックプログラミングによるS-searchは膨大な計算を必要とする
  - 通常はFASTA, BLASTなどの速さを精度に優先させたプログラムを用いている
  - データベース登録配列は5年間で10倍に増加している
    (計算機能力は10年間で10倍)



配列データベース登録件数の推移

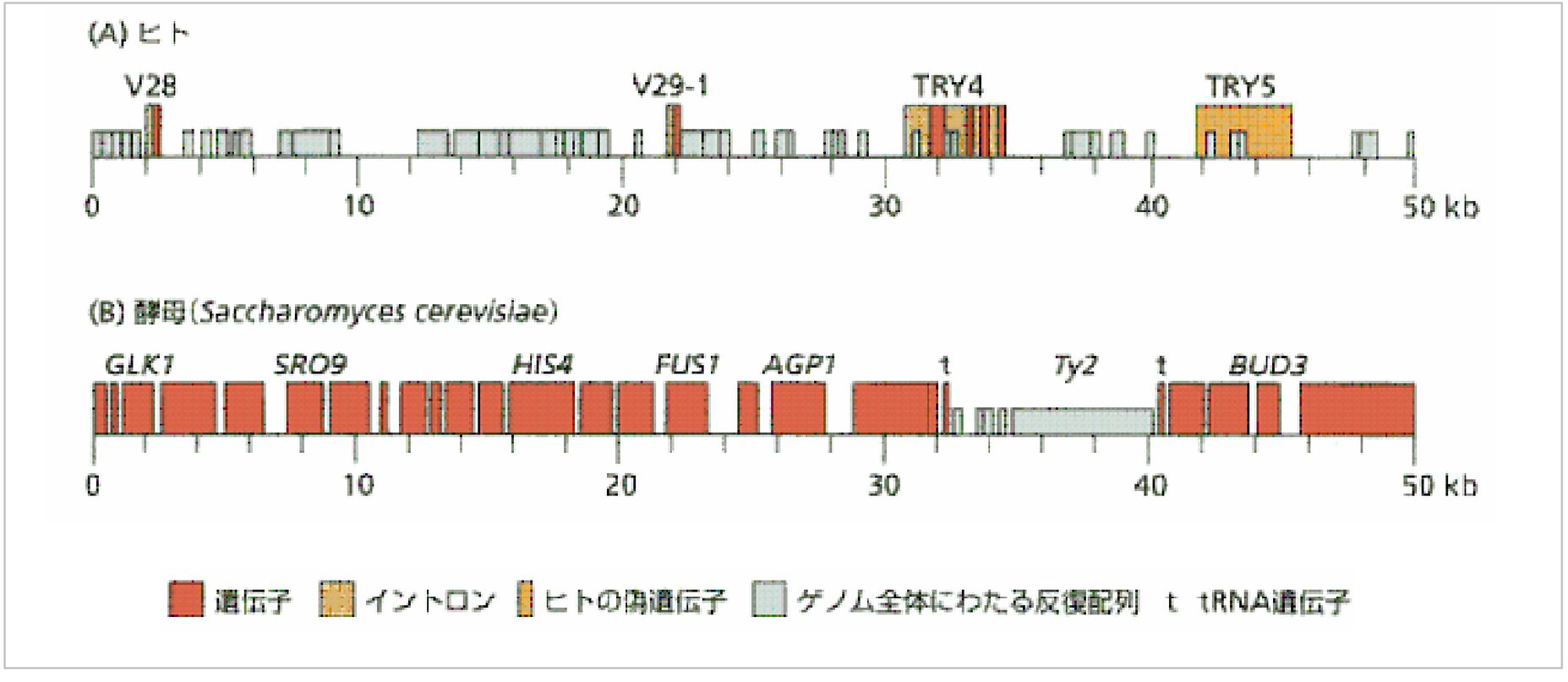
ポストゲノム解析における計算課題 (1)

- 遺伝子領域を予測する
  - ORF (開始コドンから終始コドンの前まで) を検出する



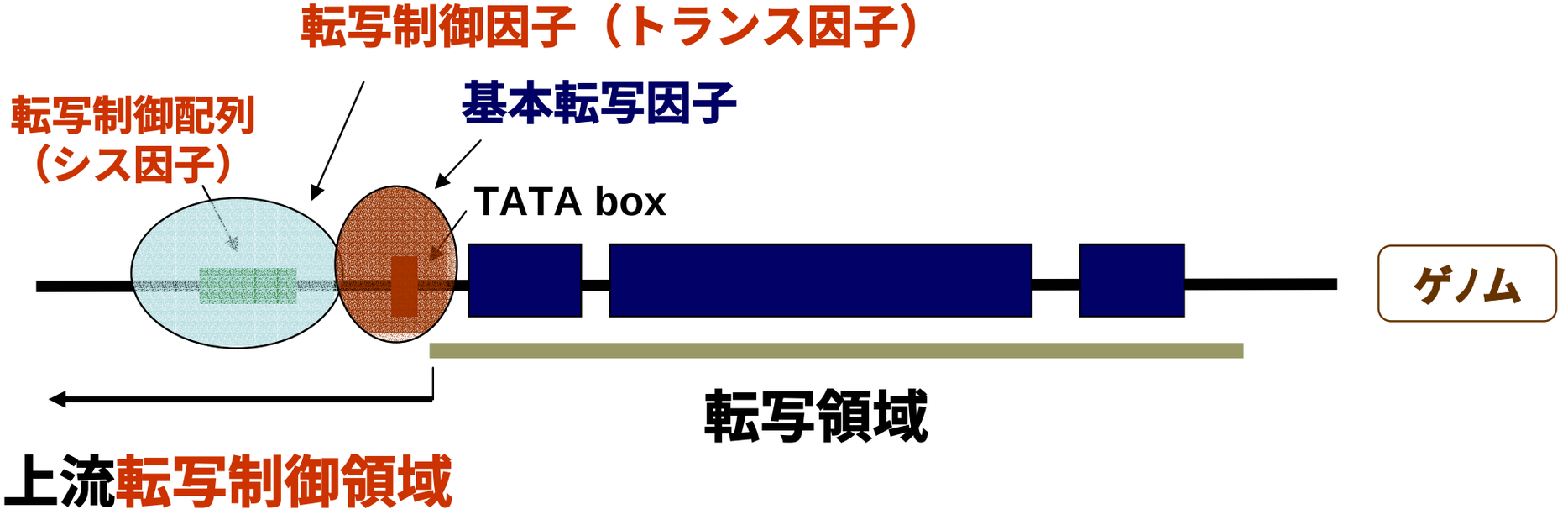
ポストゲノム解析における計算課題 (1)

- 遺伝子領域を予測する
  - イントロンの存在が遺伝子予測を困難にする



ポストゲノム解析における計算課題（2）

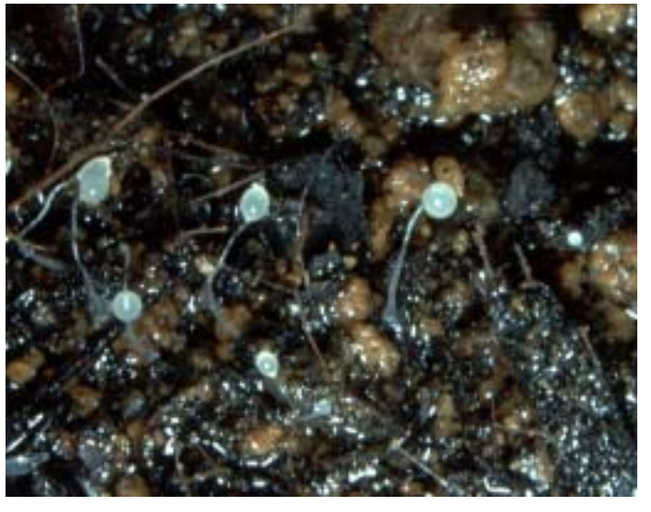
- 制御システムを解明する
  - 遺伝子の発現は時間的･空間的に制御されている
  - 遺伝子発現制御はトランス因子とシス因子の相互作用で行われる



細胞性粘菌cDNAの解析

- **細胞性粘菌ゲノムの性質**
  - **ゲノムサイズが小さい**（34 Mbp）
  - **遺伝子数は約10,000から12,000と予想されている**
  - **遺伝子密度が高く、イントロンも少ない**
    - **平均3,000塩基対に1遺伝子（ヒトの30倍）**
    - **遺伝子あたり平均1.2個のイントロン**
  - **A,Tの割合が高い**
- cDNA（complementary DNA）
  - **RNAを逆転写したDNAで、発現遺伝子の情報が得られる**

ヒトの 1/100
酵母の 3倍



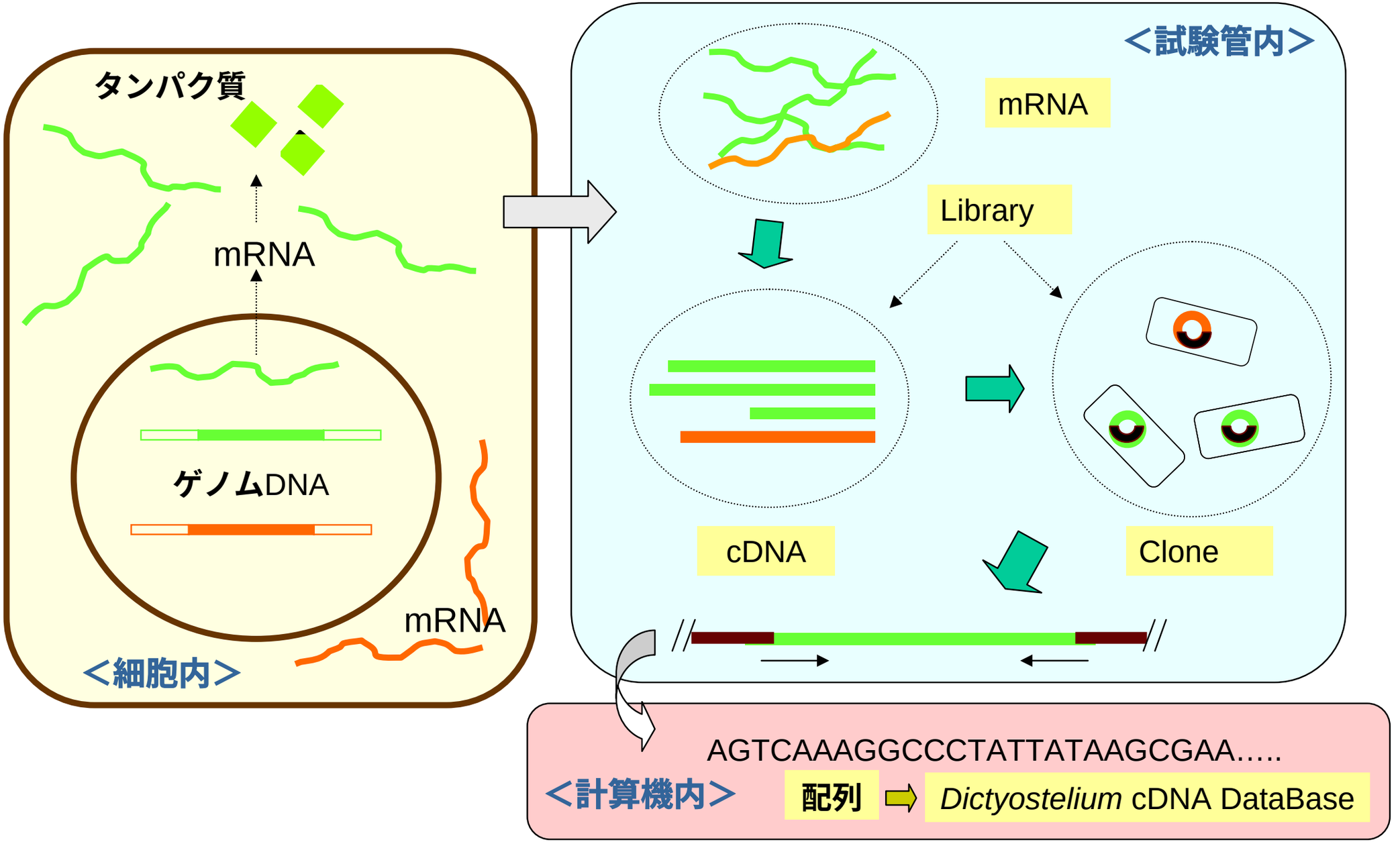
細胞性粘菌は発生･分化の優れたモデル系

多細胞体 ⇄ 単細胞アメーバ

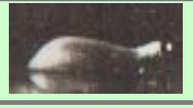
cDNA解析の課題

- 遺伝子レパートリーを決定する
- 遺伝子発現パターンを大規模に解析する
- 発現制御の遺伝子ネットワークを解析する
  - シス制御因子の解析
  - トランス制御因子の解析
  - 遺伝子ネットワークの記述

細胞性粘菌cDNAの大規模構造解析



約16万の配列から独立の遺伝子6,718個を取得

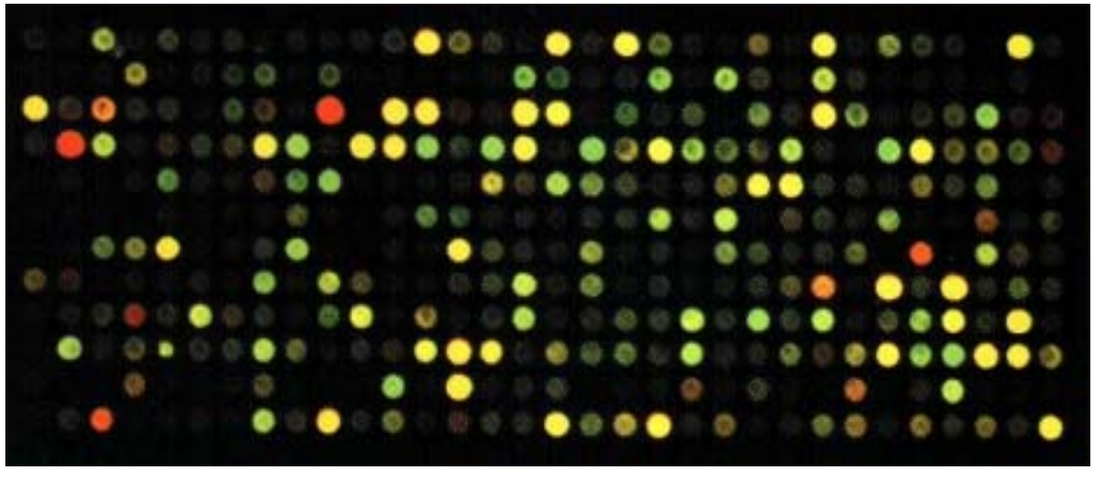
発生ステージ	配列合計	ライブラリー	オリゴ(dT)	完全長	差分化
 増殖期	49,568	Vegetative	10,850	38,718	*
 集合期	33,022	Aggregation	*	33,022	*
 移動体期	42,565	Slug	19,759	22,806	*
 形態形成期	28,783	Culmination	*	28,783	*
 有性生殖期	3,346	Fusion-competent	1,450	*	1,896
全体	157,284		32,059	123,329	1,896

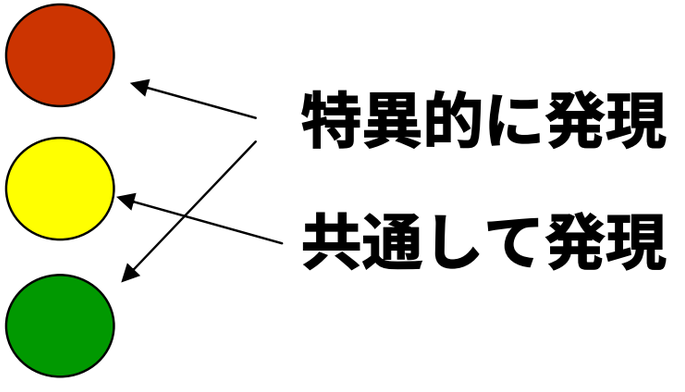
独立遺伝子	6,718
ギャップなし	3,735
ギャップあり	2,983
Orphan contig	1,643
内部配列決定	822

＊ゲノム配列からの予測： 約12,000遺伝子

DNAマイクロアレイを用いた遺伝子発現解析

- スライドグラスにDNAを固定
- 2種類のmRNA集団を別々の蛍光色素で標識して添加
- スポットの「色」を検出
- 統計処理して遺伝子の発現パターンを解析





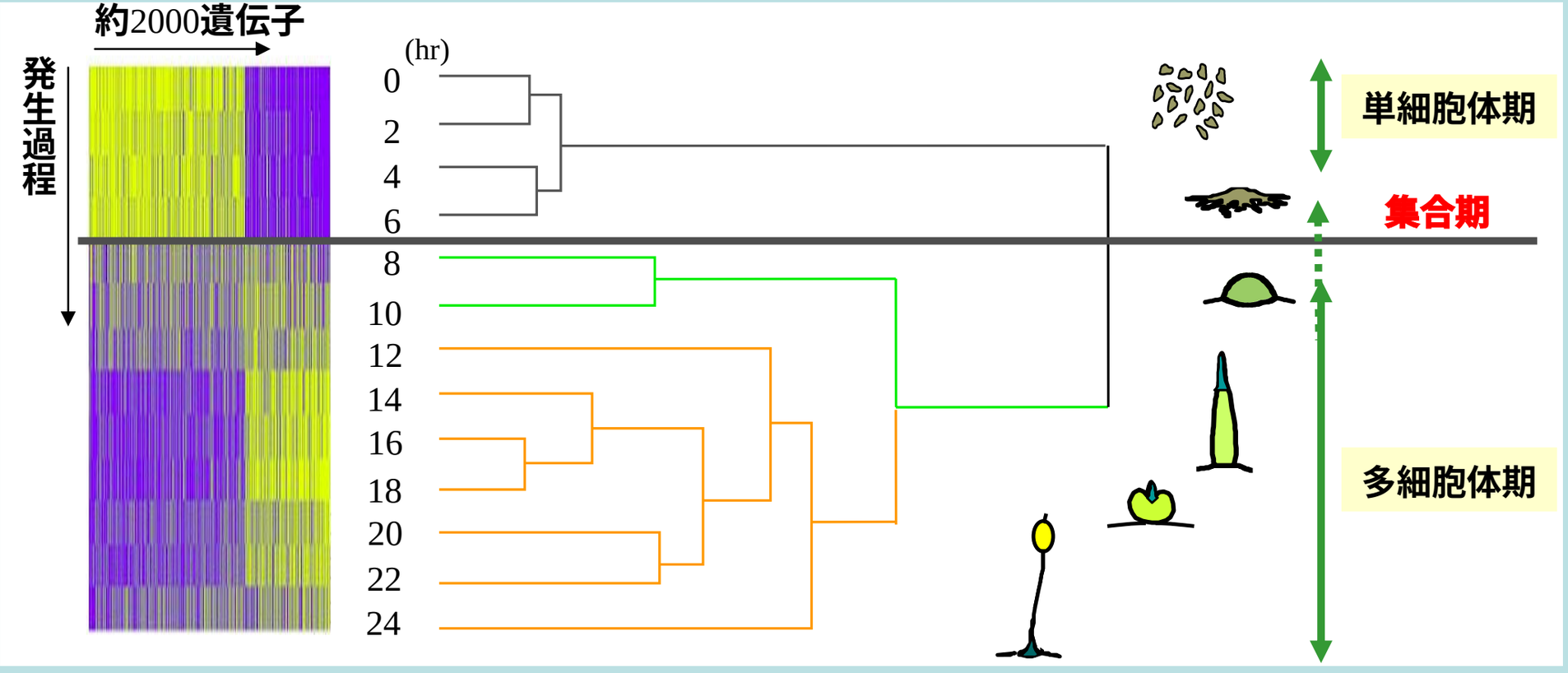
DNAマイクロアレイの生データ

Clone	0 hr	2 hr	4 hr	6 hr	8 hr	10 hr	12 hr	14 hr	16 hr	18 hr	20 hr	22 hr	24 hr
SLB855	3.9355	4.8221	19.0074	14.1344	-2.7017	-3.9273	-4.1082	-4.6417	-10.1923	-8.5741	-6.2508	-1.4991	-0.0042
SLE485	-7.7066	-7.858	-7.2362	-5.9978	-3.3286	-5.7987	6.1344	-3.6182	13.4575	16.8135	0.4504	1.267	3.4212
SSA322	-10.9649	-6.4283	-10.6752	-4.178	1.4092	3.32	2.9937	2.3047	3.553	9.0822	3.6494	1.4655	4.4688
SLK887	-3.3436	-4.5103	-6.3033	-4.0391	-4.657	-4.6845	3.495	-2.7829	9.732	10.8212	1.0199	0.2534	4.999
SLJ768	-1.0696	-0.4841	8.8838	12.7732	-3.9083	-3.3456	1.5682	3.5282	-3.7411	-4.8465	-4.1729	-3.0514	-2.1339
SSA136	-5.3441	-3.9835	-5.1164	-4.2782	-0.5016	-1.7543	1.2139	-1.0473	11.5656	9.3608	0.1382	1.16	-1.4132
SSH889	-7.9347	-5.3846	-8.4705	-5.9409	1.3669	1.3786	3.5525	2.9791	4.4909	5.6365	2.0925	3.2209	3.0127
SSH241	-9.5843	-8.0308	-4.5254	-2.3735	1.5352	2.1471	2.7205	1.7607	4.0032	6.3262	2.957	-0.79	3.8541
SSB695	-4.021	-0.8391	-2.6789	-1.7363	-3.9821	-4.9579	4.0333	11.5797	2.1721	5.1379	-4.4281	-1.6594	1.3797
SSL316	-2.9372	-2.547											
SLD219	-1.6841	-1.676											
SLA767	-3.0317	-1.26											
SLK851	-3.8257	-5.866											
SSG694	1.4655	-4.3743	-0.5548	2.0329	8.7179	7.1831	-5.3379	-4.5333	-1.7082	-1.8266	0.5521	1.7289	-3.3453
SLJ129	-2.0061	-0.8093	-1.3341	-1.3813	-1.7433	-9.0286	1.0745	0.419	3.0163	9.5232	-1.3796	1.0952	2.5541
SSJ231	-5.3166	-5.0627	-6.3445	-2.9014	3.2047	2.3881	-1.6468	-0.4165	4.5549	3.2116	2.6222	6.0777	-0.3706
SSA137	-5.9269	-3.2812	-4.9375	-4.6988	-0.1334	1.2623	1.9565	0.9167	3.0619	9.0337	2.0774	0.016	0.6533
SSF596	-5.3468	-4.5544	-5.8027	-4.5497	0.9167	1.5641	0.7907	1.5379	2.5837	7.1139	2.3699	-0.5035	3.8802
SSM591	-4.5908	-6.6144	-6.397	-4.0565	2.4779	2.1545	2.2163	1.5993	3.3941	4.1563	1.7393	1.514	2.4069

遺伝子数×測定ポイント数×４（再現性のため）の数値

有意なデータを得るために統計処理が必要である

細胞の集合によって遺伝子発現が大きく変化する



遺伝子を発現量の時間的変化でグループ化し、変化する2000個を示した。

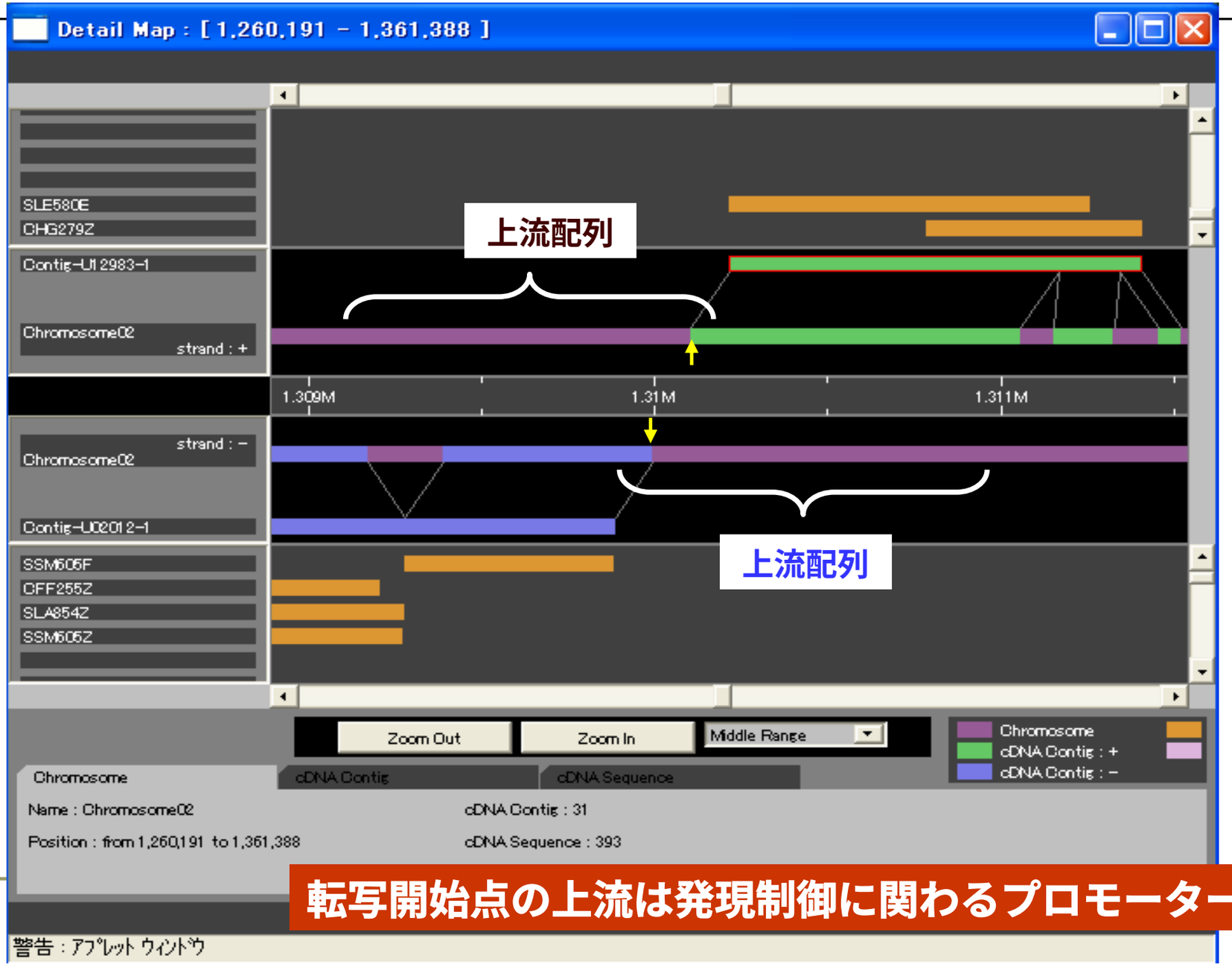
青いバーは発現量が平均の1/2以下であること、黄色いバーは２倍以上であることを示す。

シス因子（上流調節配列）の解析

- 遺伝子を発現パターンでグループ化
- 遺伝上流配列の取得
- グループに特徴的な上列配列エレメントを抽出

- エレメントが実際に機能していることの確認
- エレメントに結合する因子の単離・同定

遺伝子上流配列の取得とシス因子の抽出



転写開始点の上流は発現制御に関わるプロモーター領域である

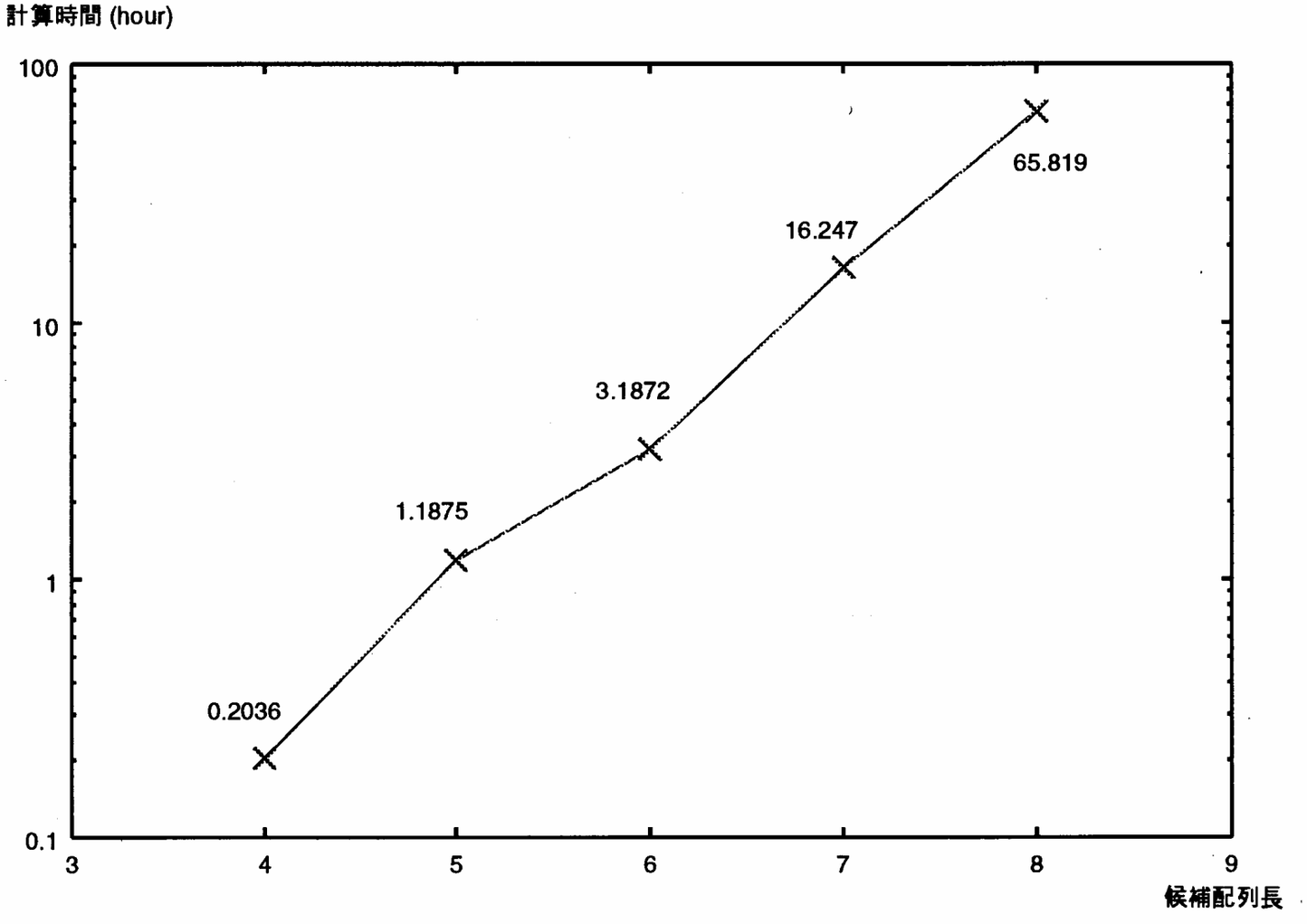
シス因子（上流調節配列）の解析

- 遺伝子を発現パターンでグループ化
- 遺伝上流配列の取得
- グループに特徴的な上列配列エレメントを抽出

- エレメントが実際に機能していることの確認
- エレメントに結合する因子の単離・同定

（シス因子から攻める遺伝子間相互作用）

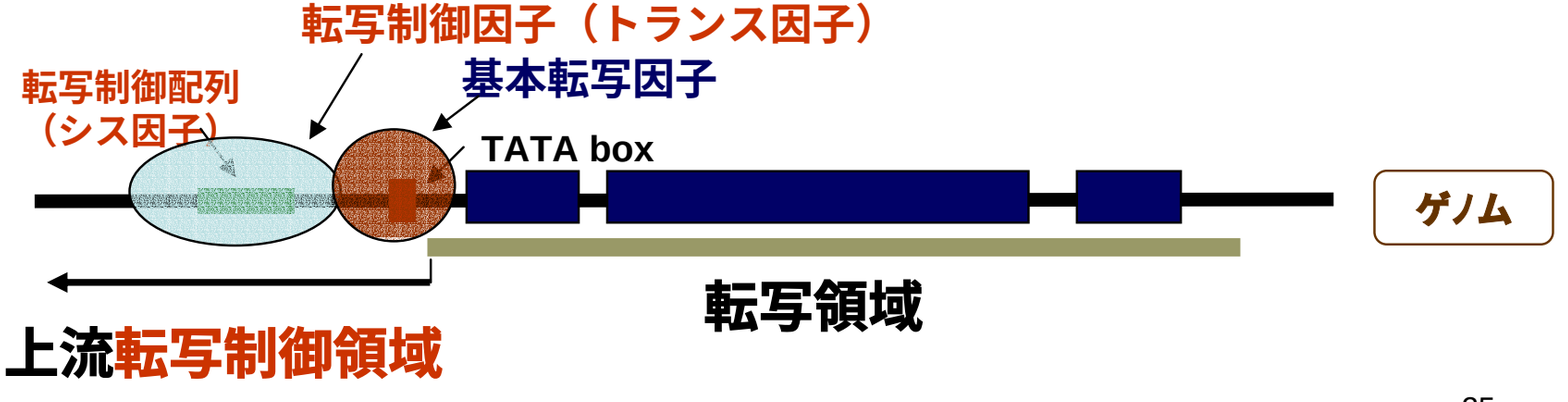
特徴配列の抽出に必要な計算時間



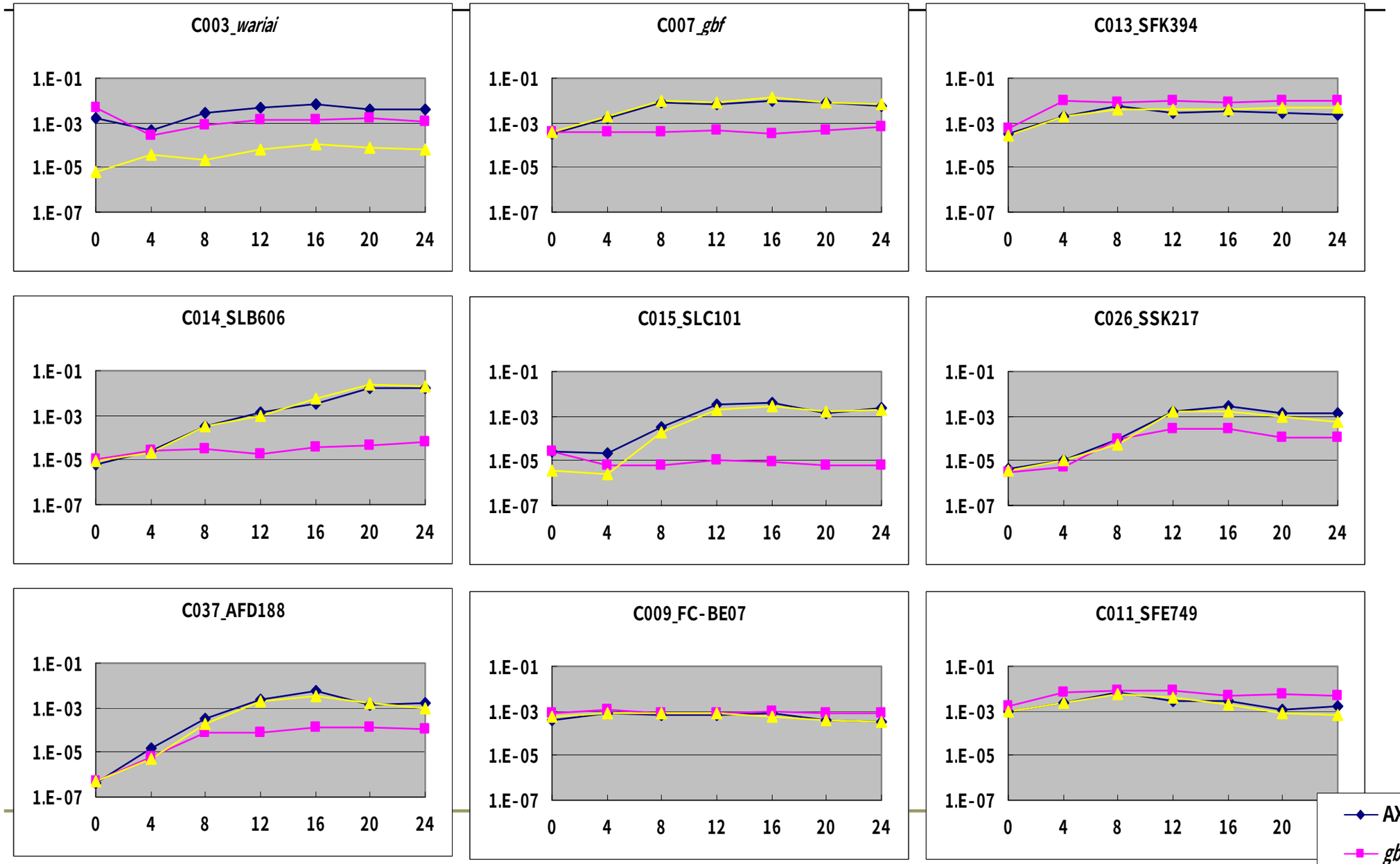
電子情報工学系安永研究室　牛山建太郎氏　修士論文

遺伝子ネットワークの推定

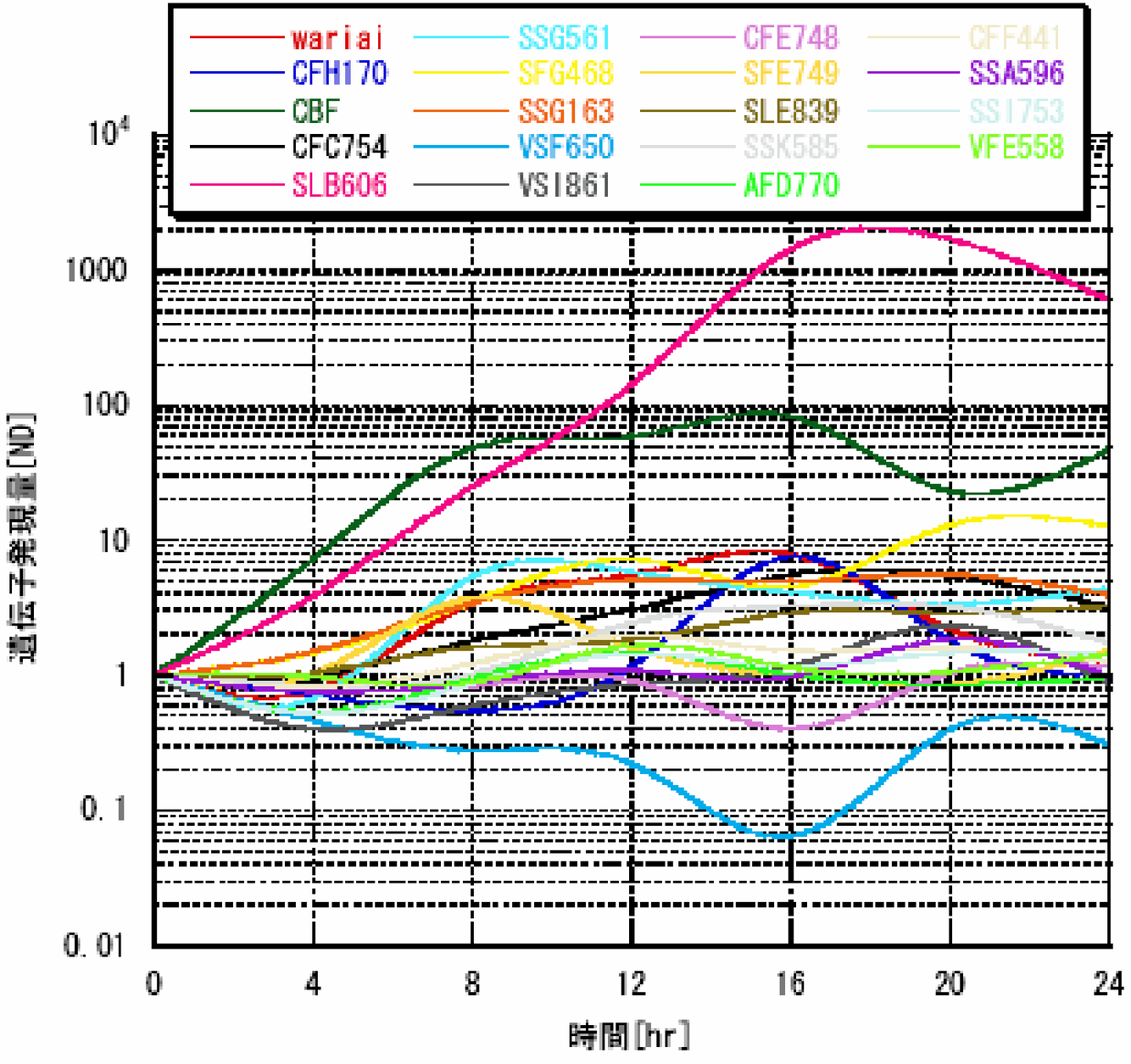
- 遺伝子破壊株での発現変化データを利用
- 遺伝子発現の時系列変化データを利用
  - S-systemモデルによる推定
    - 遺伝子数が少ない場合
  - 線形微分方程式モデルによる推定
    - 遺伝子数が多い場合



定量PCRによる時系列発現データの取得



遺伝子ネットワーク推定用のデータ



S-systemによるネットワーク解析

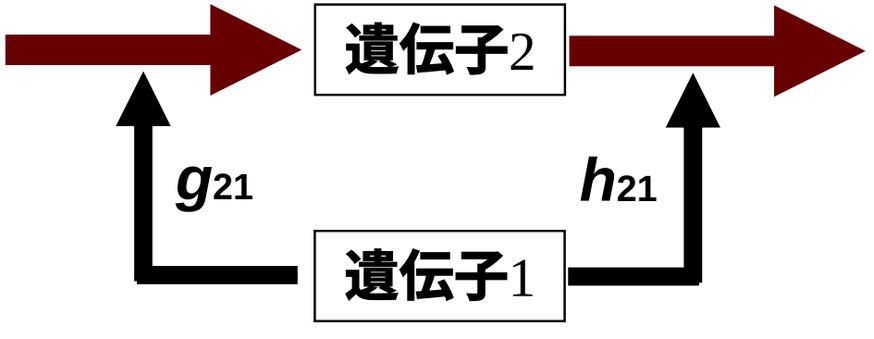
・さまざまな動的挙動を表現できる連立微分方程式

・本研究での方程式の解：
　　定量PCRによる遺伝子発現データ（タイムコース）

$$\frac{dX_i}{dt} = \underbrace{\alpha_i \prod_{j=1}^{n} X_j^{g_{ij}}}_{\text{合成項}} - \underbrace{\beta_i \prod_{j=1}^{n} X_j^{h_{ij}}}_{\text{分解項}} \qquad (i = 1,2,\ldots,n)$$

n	遺伝子の数
X	遺伝子の発現量
α, β	速度係数
g, *h*	相関係数

例えば、g_{21}というパラメータについて考える



遺伝子1が遺伝子2の合成過程を

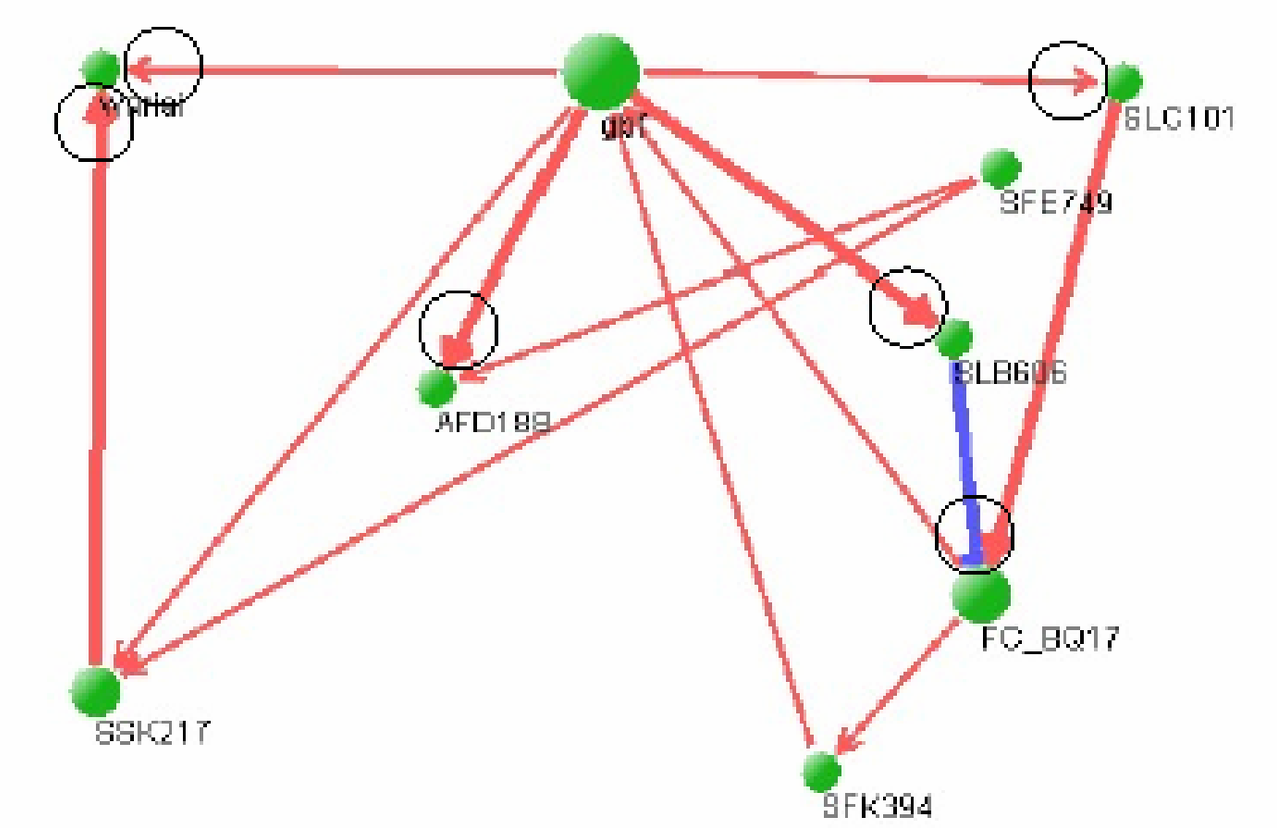
$g_{21} > 0$ ・・・促進

$g_{21} < 0$ ・・・抑制

$g_{21} = 0$ ・・・無関係

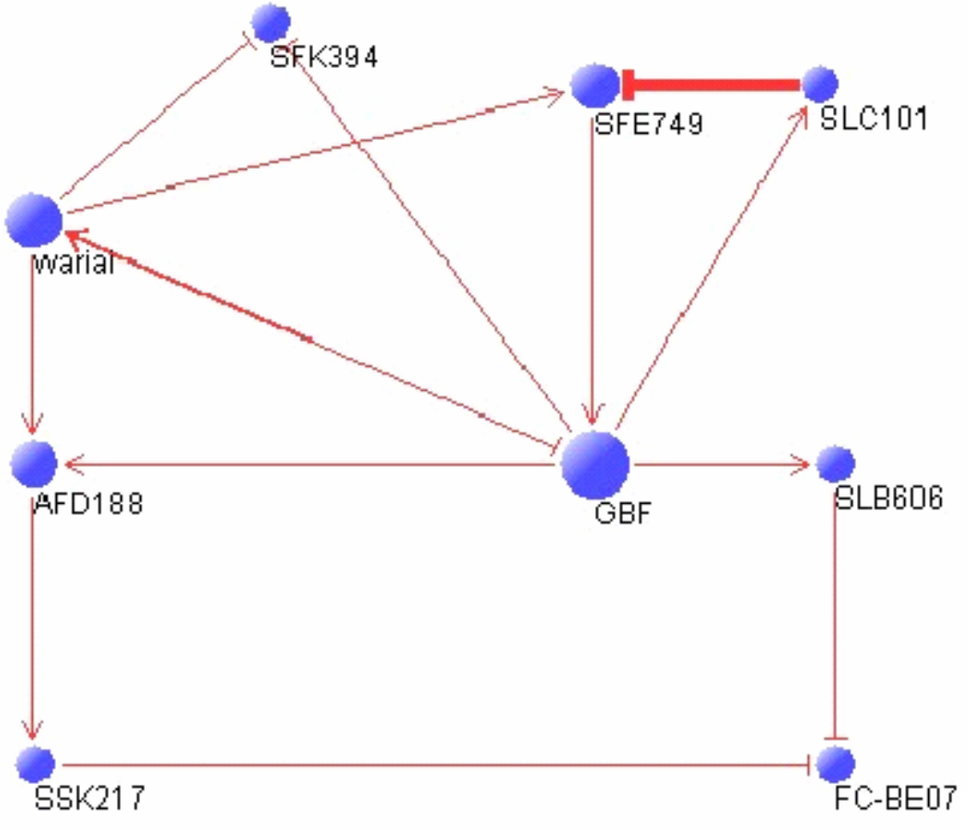
ネットワークの推定

発現パターンデータをS-system に適用し、
9個の転写因子遺伝子のネットワークを推定した結果

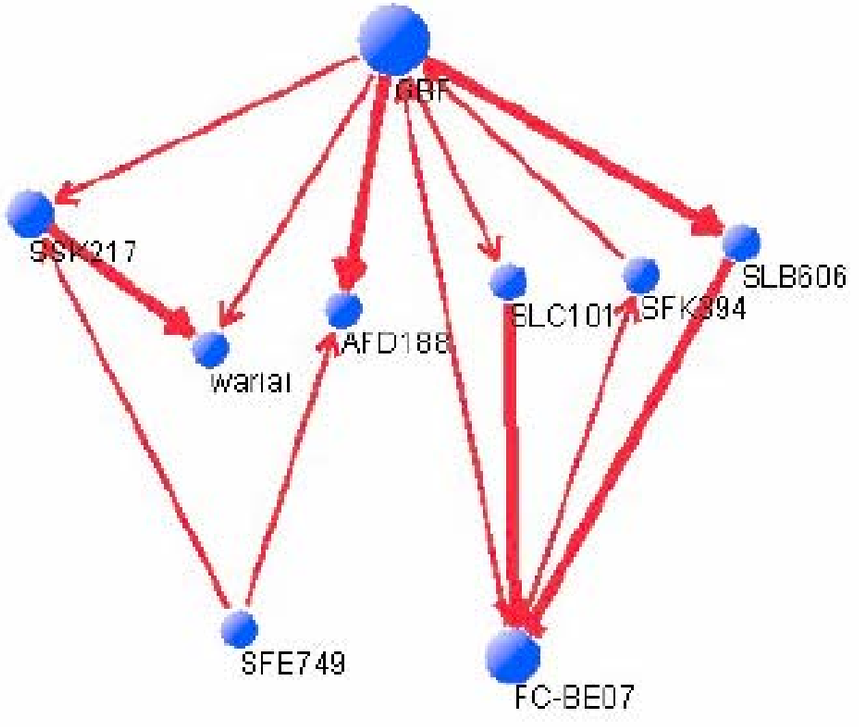


○：2回の実験に共通して推定された相互作用

遺伝子ネットワークの推定



線形微分方程式モデルの解

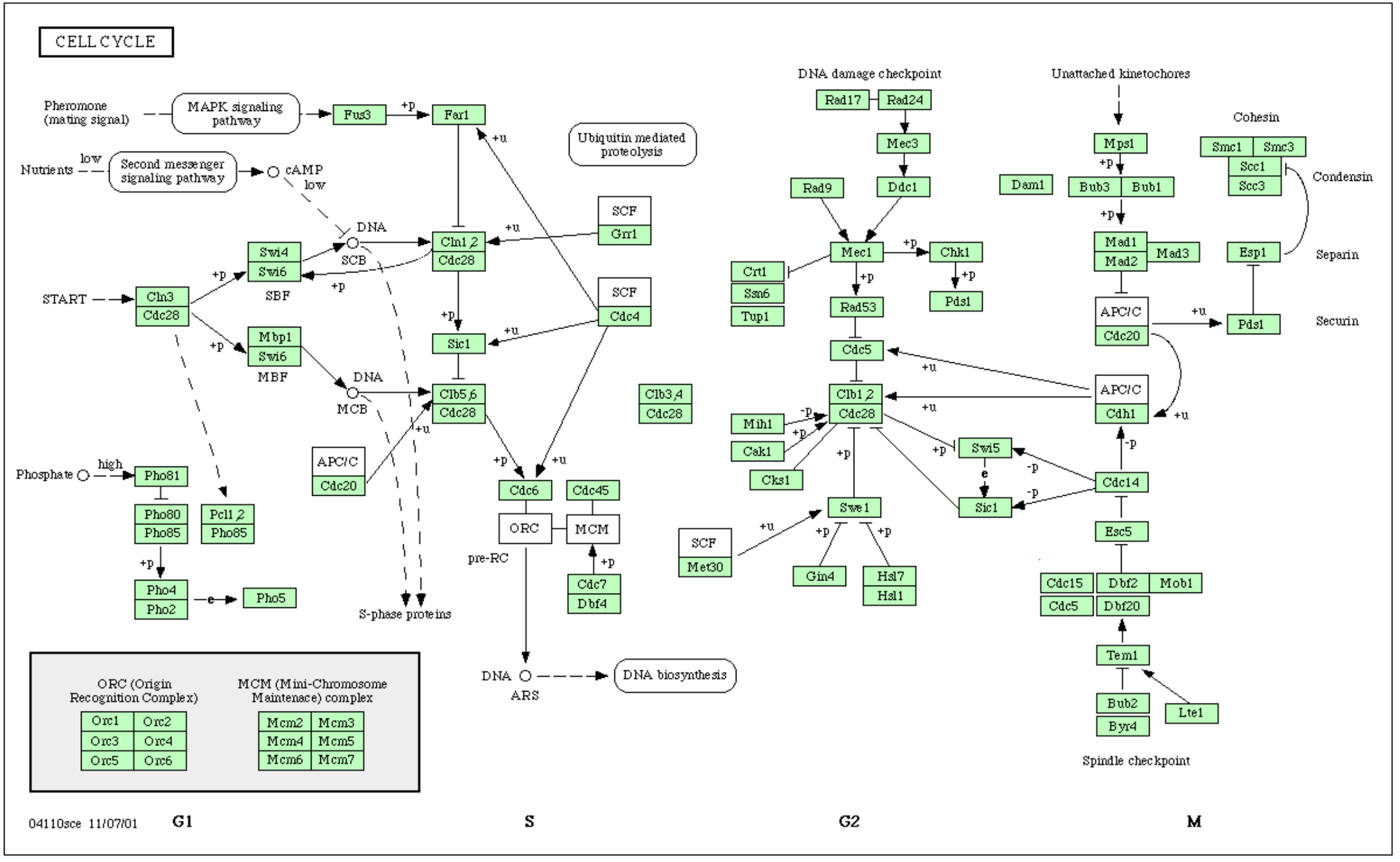


S-systemモデルの解

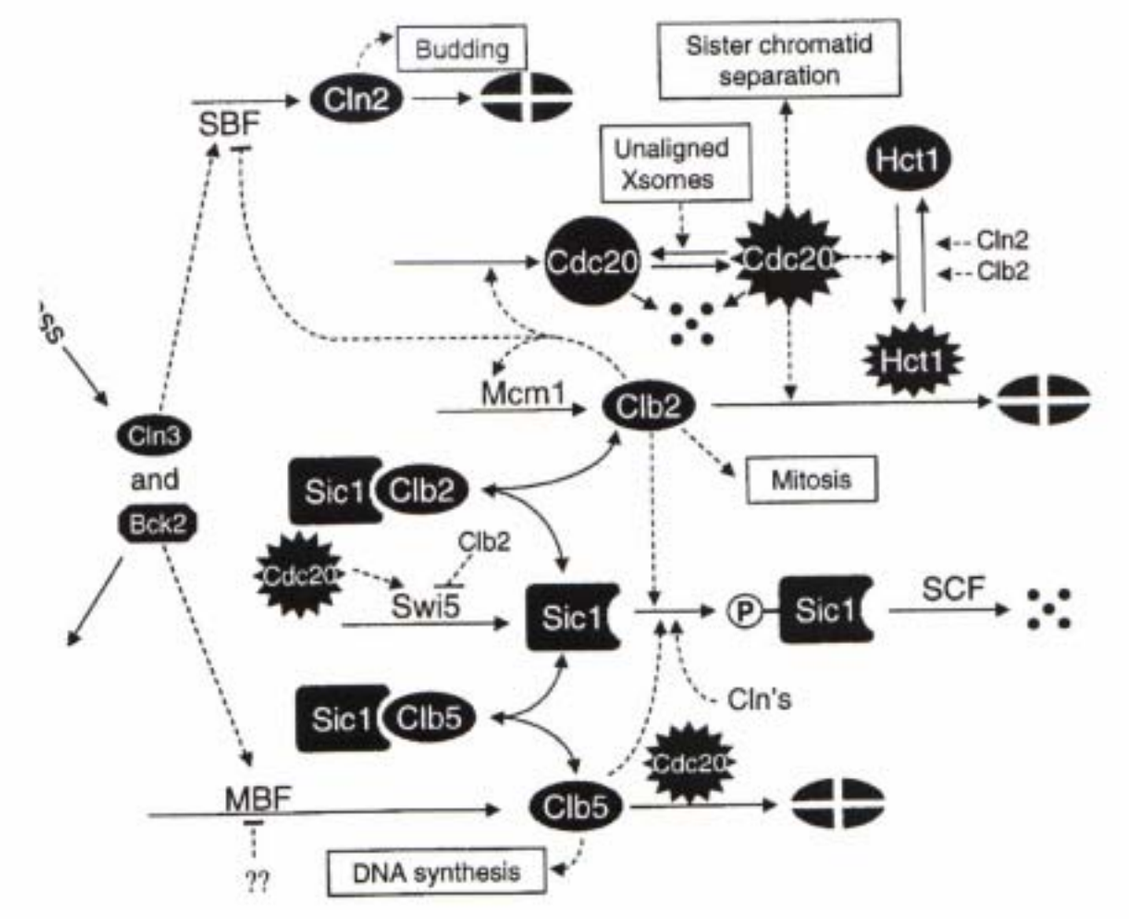
ポスト「ポストゲノム」の課題

- 生命活動のシミュレーション
  - 代謝
  - 分裂 (細胞周期)
  - コミュニケーション
  - 遺伝子制御システム

細胞周期制御の遺伝子ネットワーク



細胞周期コントロールのネットワークを計算機内に模擬することによって解析できないか



Chen *et al., Mol. Biol. Cell* (2000).**11**:369-391

Step 1. 細胞周期関連分子に関する微分方程式を作る

$$\frac{d}{dt}[\mathrm{Cln2}] = (k'_{s,n2} + k''_{s,n2}[\mathrm{SBF}]) \cdot \mathrm{mass} - k_{d,n2}[\mathrm{Cln2}]$$

Equations governing cyclin-dependent kinases

$$\frac{d}{dt}[\mathrm{Cln2}] = (k'_{s,n2} + k''_{s,n2}[\mathrm{SBF}]) \cdot \mathrm{mass} - k_{d,n2}[\mathrm{Cln2}]$$

$$\frac{d}{dt}[\mathrm{Clb2}]_T = (k'_{s,b2} + k''_{s,b2}[\mathrm{Mcm1}]) \cdot \mathrm{mass} - V_{d,b2}[\mathrm{Clb2}]_T,\ V_{d,b2} = k'_{d,b2}([\mathrm{Hct1}]_T - [\mathrm{Hct1}]) + k''_{d,b2}[\mathrm{Hct1}] + k'''_{d,b2}[\mathrm{Cdc2}$$

$$\frac{d}{dt}[\mathrm{Clb5}]_T = (k'_{s,b5} + k''_{s,b5}[\mathrm{MBF}]) \cdot \mathrm{mass} - V_{d,b5} \cdot [\mathrm{Clb5}]_T,\ V_{d,b5} = k'_{d,b5} + k''_{d,b5}[\mathrm{Cdc20}]$$

$$[\mathrm{Bck2}] = [\mathrm{Bck2}]^0 \cdot \mathrm{mass}, \qquad [\mathrm{Cln3}]^* = [\mathrm{Cln3}]_{\max} \frac{D_{n3} \cdot \mathrm{mass}}{J_{n3} + D_{n3} \cdot \mathrm{mass}}$$

$$[\mathrm{Clb2}]_T = [\mathrm{Clb2}] + [\mathrm{Clb2/Sic1}], \quad [\mathrm{Clb5}]_T = [\mathrm{Clb5}] + [\mathrm{Clb5/Sic1}]$$

$$[\mathrm{Sic1}]_T = [\mathrm{Sic1}] + [\mathrm{Clb2/Sic1}] + [\mathrm{Clb5/Sic1}]$$

Equations governing the inhibitor of Clb-dependent kinases

$$\frac{d}{dt}[\mathrm{Sic1}]_T = k'_{s,c1} + k''_{s,c1}[\mathrm{Swi5}] - \left(k_{d1,c1} + \frac{V_{d2,c1}}{J_{d2,c1} + [\mathrm{Sic1}]_T}\right) \cdot [\mathrm{Sic1}]_T$$

$$\frac{d}{dt}[\mathrm{Clb2/Sic1}] = k_{as,b2}[\mathrm{Clb2}] \cdot [\mathrm{Sic1}] - \left(k_{di,b2} + V_{d,b2} + k_{d1,c1} + \frac{V_{d2,c1}}{J_{d2,c1} + [\mathrm{Sic1}]_T}\right) \cdot [\mathrm{Clb2/Sic1}]$$

$$\frac{d}{dt}[\mathrm{Clb5/Sic1}] = k_{as,b5}[\mathrm{Clb5}] \cdot [\mathrm{Sic1}] - \left(k_{di,b5} + V_{d,b5} + k_{d1,c1} + \frac{V_{d2,c1}}{J_{d2,c1} + [\mathrm{Sic1}]_T}\right) \cdot [\mathrm{Clb5/Sic1}]$$

Step 2. **反応定数を求める**

$k'_{s,n2} = 0$ $k''_{s,n2} = 0.05$ $k_{d,n2} = 0.1$

Table 2. Kinetic constants for the budding yeast model

Rate constants (min^{-1})

$k'_{s,n2} = 0$
$k'_{s,b2} = 0.002$
$k'_{d,b2} = 0.01$
$k'_{s,b5} = 0.006$
$k'_{s,c1} = 0.02$
$k_{as,b2} = k_{as,b5} = 50$
$k'_{s,20} = 0.005$
$k_{a,20} = 1$
$k'_{a,t1} = 0.04$
$k_{s,ori} = 2$
$k_{d,ori} = k_{d,bud} = k_{d,spn} = 0.06$
$k'_{i,sbf} = 0.5$
$k_{i,mcm} = 0.15$

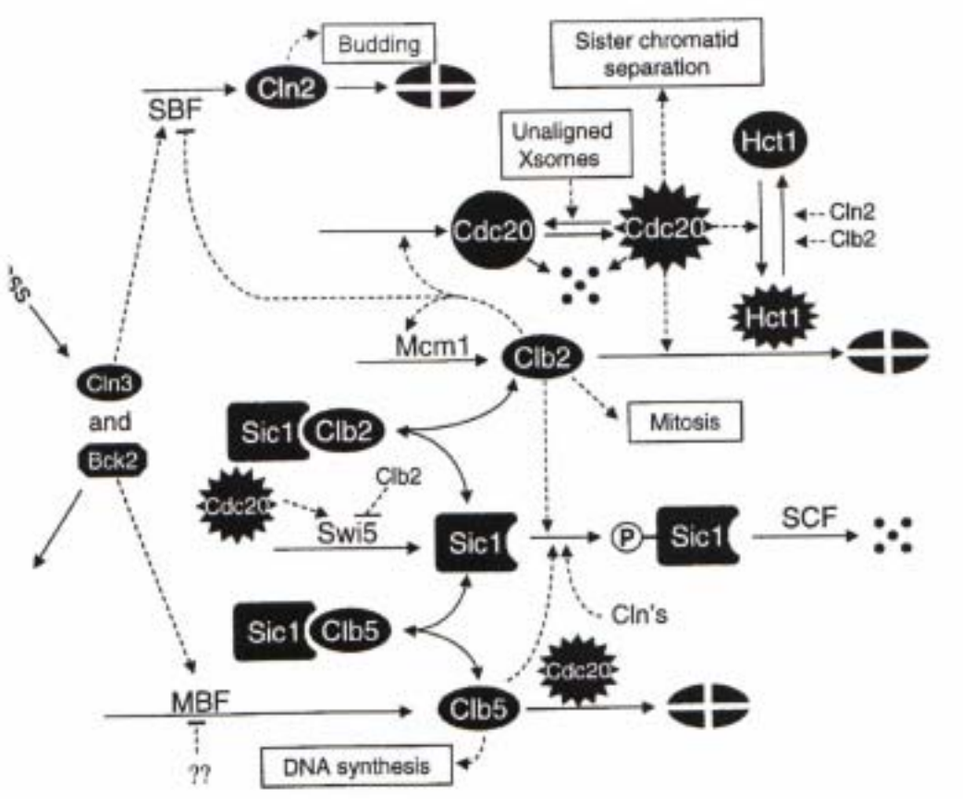
$k''_{s,n2} = 0.05$
$k''_{s,b2} = 0.05$
$k''_{d,b2} = 2$
$k''_{s,b5} = 0.02$
$k''_{s,c1} = 0.1$
$k_{di,b2} = k_{di,b5} = 0.05$
$k''_{s,20} = 0.06$
$k'_{i,20} = 0.1$
$k''_{a,t1} = 2$
$k_{s,bud} = 0.3$

$k''_{i,sbf} = 6$
$\mu = 0.005776$

Characteristic concentrations (dimensionless)

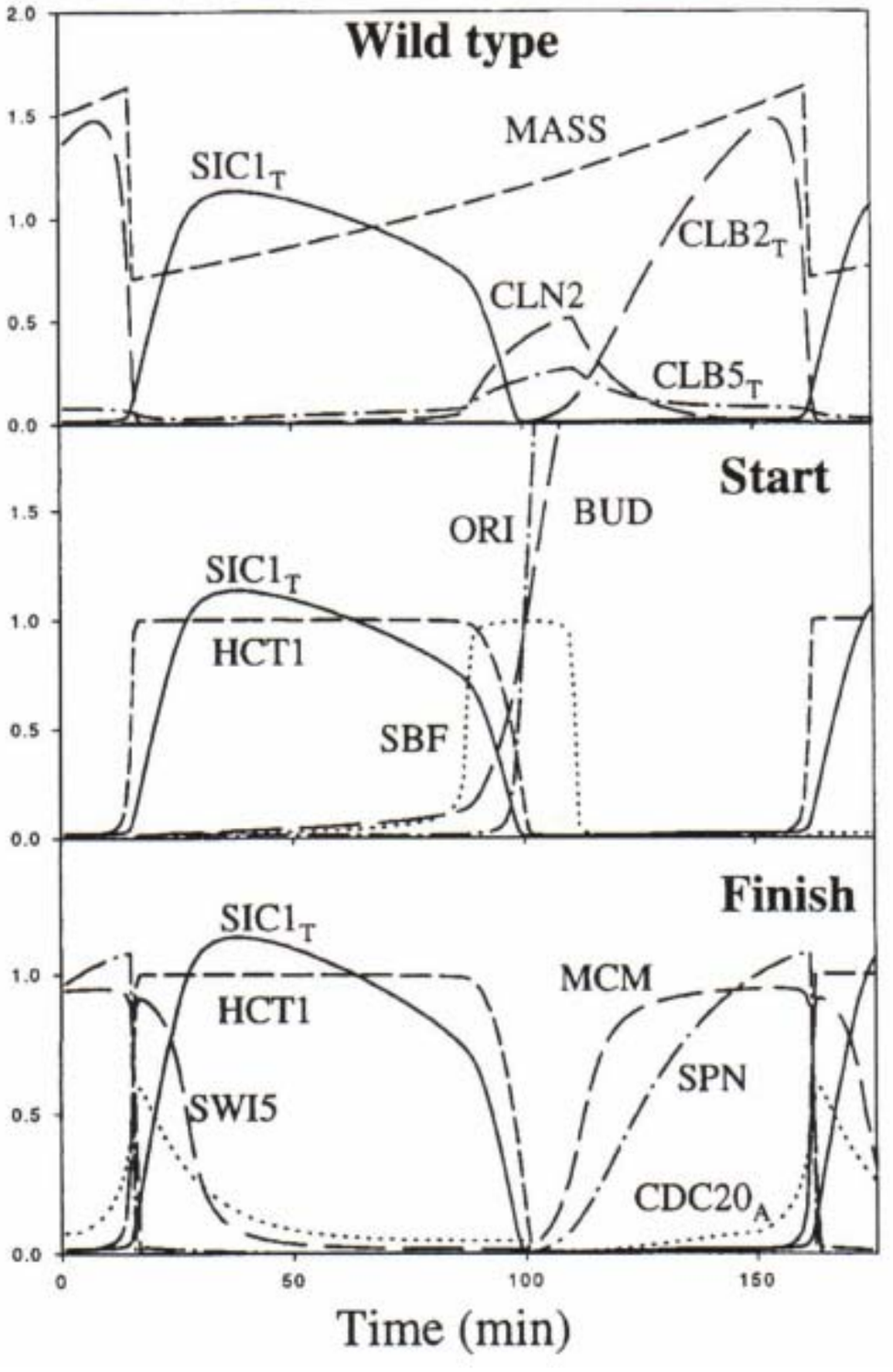
$[Cln3]_{max} = 0.02$

$[Bck2]^0 = 0.0027$



細胞周期のシミュレーション

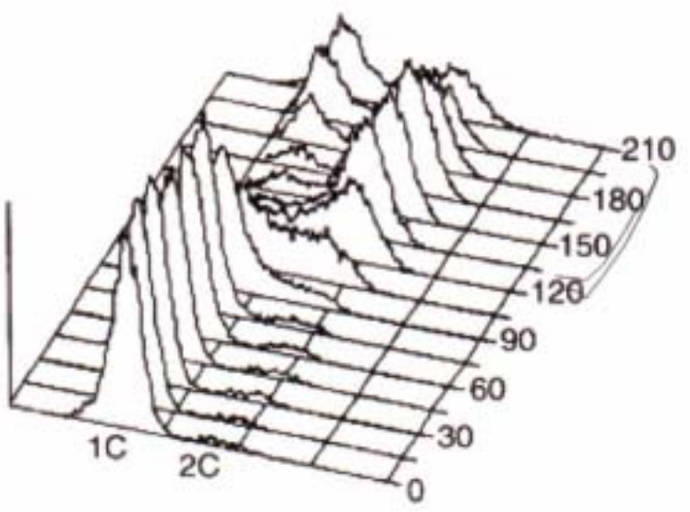
(計算機内の細胞増殖)



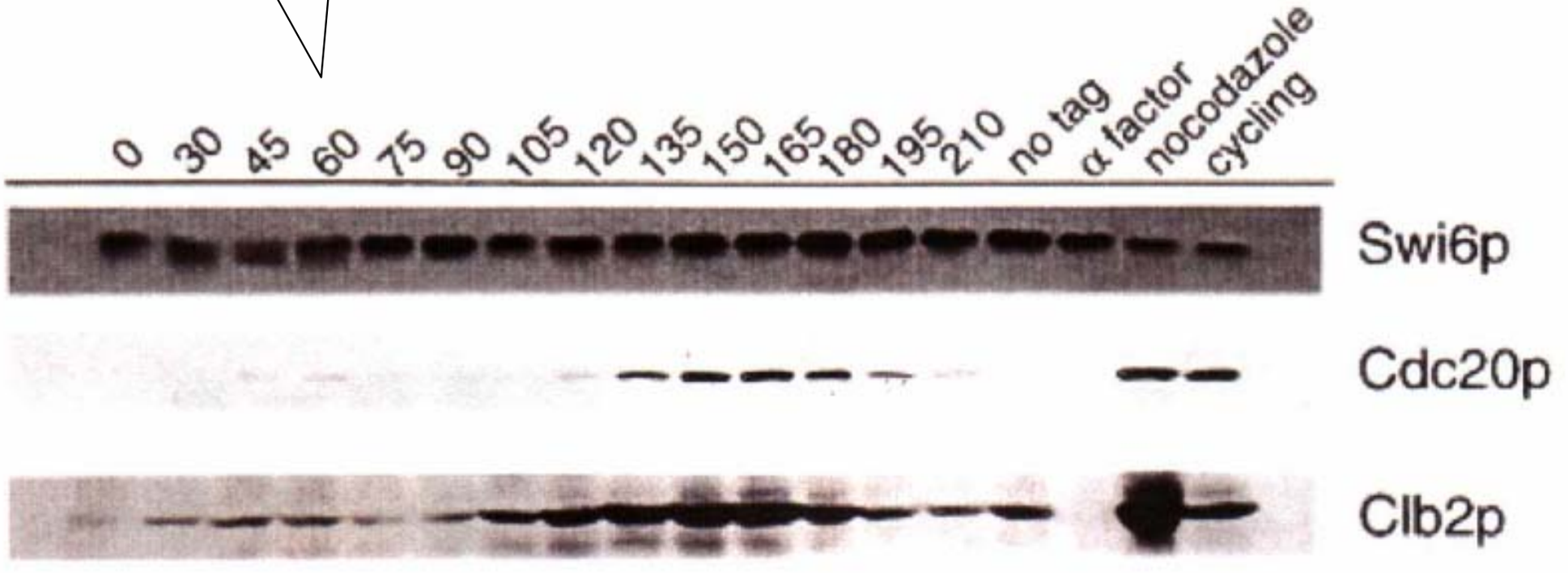
Our simulation

Step 0. 反応定数を決めるための実験データ

細胞周期の進行状況チェック



タンパク質濃度変化の測定

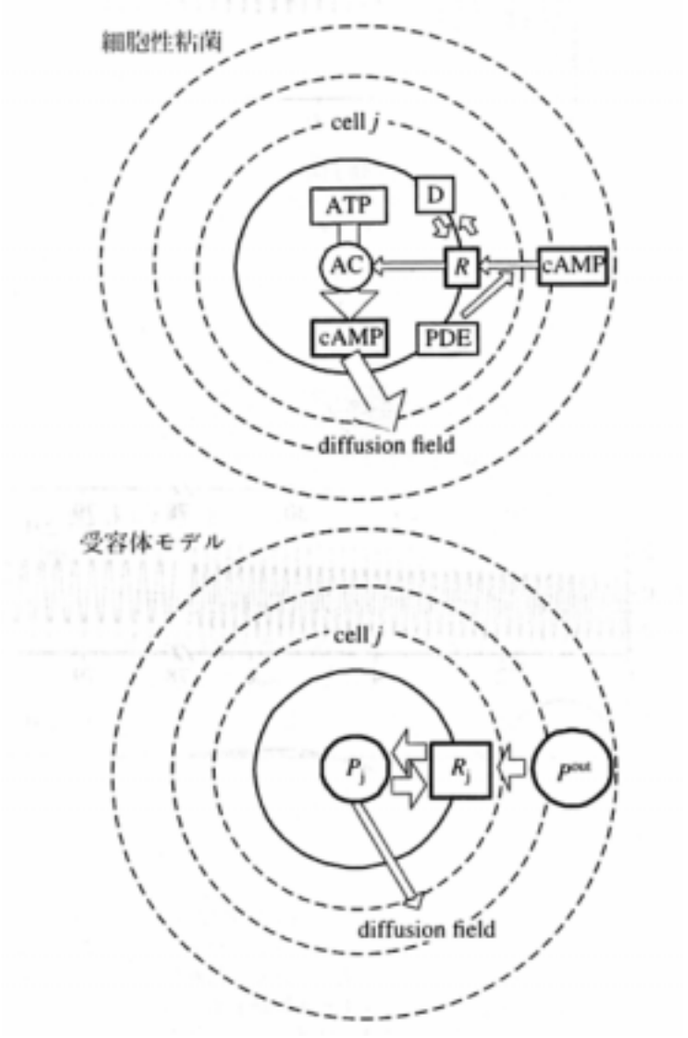


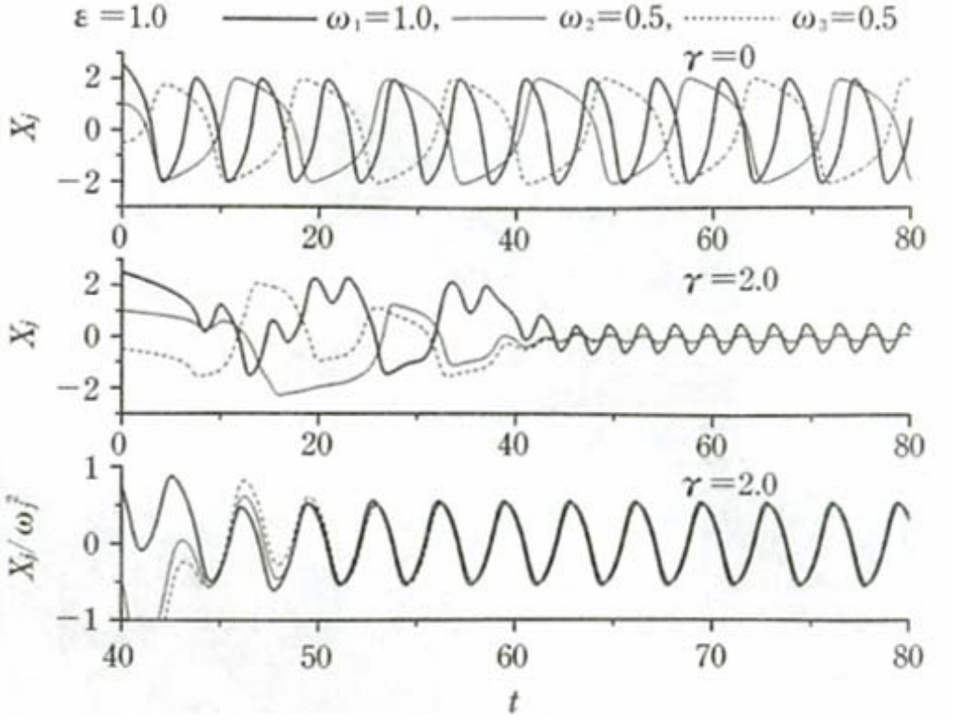
simulation

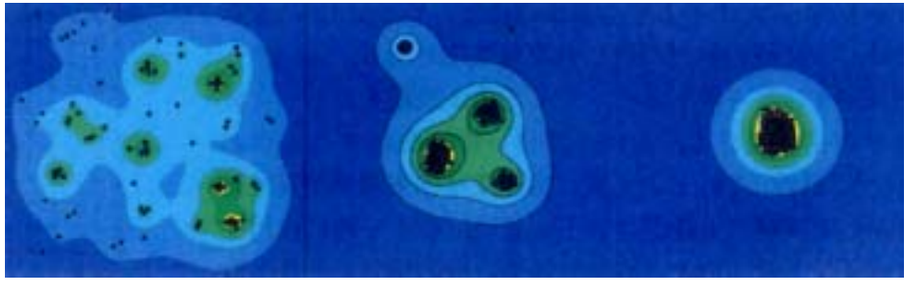
ポスト「ポストゲノム」の課題

- 生命活動のシミュレーション
  - 代謝
  - 分裂 (細胞周期)
  - コミュニケーション
  - 遺伝子制御システム

細胞性粘菌集合のシミュレーションから システム同期の新たな知見を得る







長野正道　応用物理70, 2001

本日の内容

- 生物遺伝情報体系の概略
- ゲノム解析過程では大規模な計算が実行されている (軌道に乗っている)
- ポストゲノム解析でも大規模な計算が必要である (悪戦苦闘している)
- 生命現象のシミュレーションから自然現象を説明できる新たな知見が得られる (と期待している)

…ゲノム科学は計算科学の助けを必要としている