トランスクリプトーム解析・プロテオーム解析入門

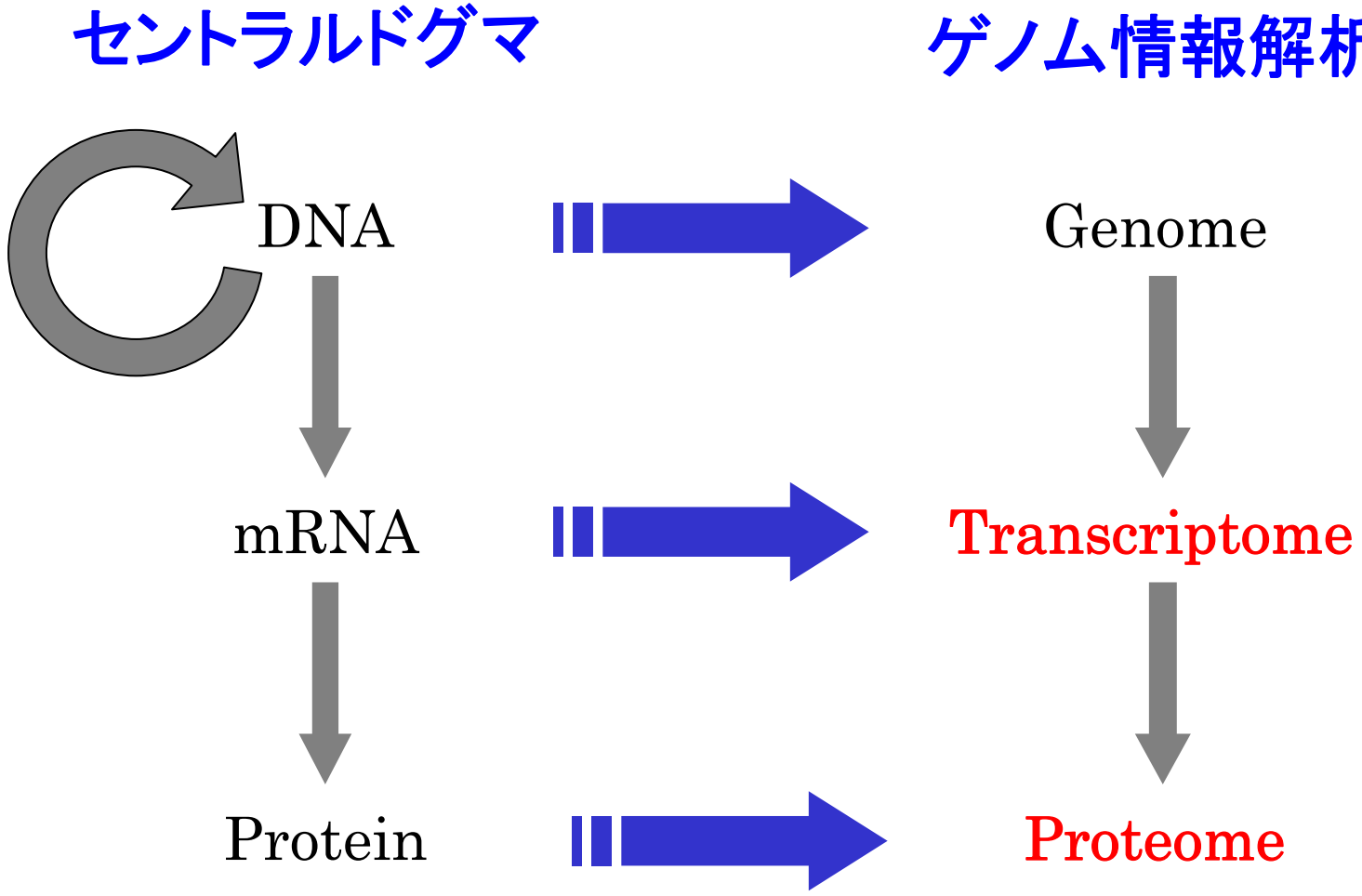
産業技術総合研究所
生命情報工学研究センター

油谷　幸代

内容

- 背景
- トランスクリプトーム解析
- プロテオーム解析

背景(1/6)
-セントラルドグマとゲノム情報解析-



背景(2/6)
-ゲノムとは？-

Genomeとは？

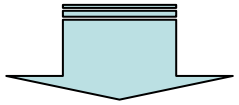
定義:
ある生物をその生物たらしめるのに必須な遺伝情報。

由来:
遺伝情報を持つ単位である**gene**(遺伝子)と
chromo**some**(染色体)を組み合わせた造語であり、
細胞における遺伝子全体を対象とする。
1920年にドイツのハンブルク大学の植物学者Hans Winklerにより造られた。

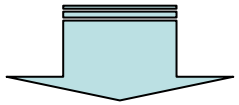
背景(3/6)
-ゲノム配列解析-

微生物ゲノム

1995 *H.influenzae*
Fleischmann, R.D., *et al.*



2003 Complete 95
On going 344



2010 **Complete**
Bacteria: 1014
Archaea: 79

真核生物

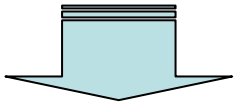
1996 *S.cerevisiae*
Goffeau, A. *et al.*

1998 *C.elegans*
The *C. elegans* Sequencing Consortium.

2000 *D. melanogaster*
Adams, M.D. *et al.*

2001 *H.sapiens (draft)*
Lander, E.S. *et al.*

2002 *S.pombe*
Wood, V. *et al*.



2010 Eukaryotes: 129
+ 13(draft)

背景(4/6)
-配列決定された主な生物-

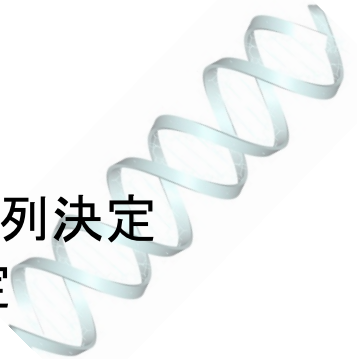
生物種	ゲノムサイズ(bp)	遺伝子推定領域
ヒトミトコンドリア	1.7×10^{4}	13
λファージ	4.8×10^{4}	50
H. salinarium (高度好塩菌)	2.6×10^{6}	2749
E. coli (大腸菌)	4.6×10^{6}	4149
S. cerevisiae (出芽酵母)	1.2×10^{7}	5880
C. elegans (線虫)	9.7×10^{7}	約20000
A. thaliana (シロイヌナズナ)	1.3×10^{8}	約27000
D. melanogaster (ショウジョウバエ)	1.8×10^{8}	約14000
H. sapiens (ヒト)	3.0×10^{9}	約26000
M. musculus (マウス)	3.3×10^{9}	約29000
O. sativa (イネ)	3.9×10^{9}	約37000
Amoeba dubia (アメーバ)	6.7×10^{11}	

背景(5/6)
-ポストシークエンス解析とは？-

シークエンス解析

対象：DNA

解析内容：ゲノムワイドでのDNA配列決定
遺伝子コード領域の決定

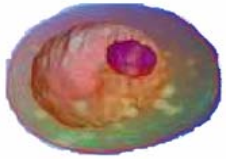


ポストシークエンス解析

対象：mRNA・Protein

解析内容：遺伝子発現・細胞内タンパク質の網羅的解析
未知遺伝子・タンパク質の機能同定
遺伝子間・タンパク質間の相互作用の解明

高次生命システムの解明



背景(6/6)
-トランスクリプトーム解析・プロテオーム解析-

トランスクリプトーム・プロテオーム解析の特徴

▶▶ **網羅性・包括性を目指した大規模解析**

トランスクリプトーム・プロテオーム解析でわかること

▶▶ **未知の遺伝子・たんぱく質の機能**
遺伝子間・たんぱく質間の相互作用

トランスクリプトーム・プロテオーム解析の有用性

▶▶ **生体細胞内における遺伝子やたんぱく質の**
働きを解析できる。 ⇒ ゲノム創薬への応用

CBRC

National Institute of Advanced Industrial Science and Technology AIST

トランスクリプトーム解析

- トランスクリプトーム解析とは？
- トランスクリプトーム解析の実験的手法
  - GeneChip技術
  - スポット型アレイ法
  - タイリングアレイ法
- アレイインフォマティクス
  - アレイインフォマティクスの目的と必要性
  - データ正規化
  - クラスター解析
  - ネットワーク解析

トランスクリプトーム解析とは？(1/3)
-トランスクリプトーム解析の基本-

定義　細胞内における**遺伝子転写産物(mRNA)全て**を要素とする集合。

由来　転写を意味する**Transcription**とGen**ome**を組み合わせて作られた造語。

目的　シークエンス解析によってDNA配列上で遺伝子と推定された部分について細胞レベルでmRNA量を測定・解析し・・・
生体細胞内における遺伝子の発現状況を網羅的に把握することを目的としている。

トランスクリプトーム解析とは？(2/3)
-トランスクリプトーム解析の流れ-

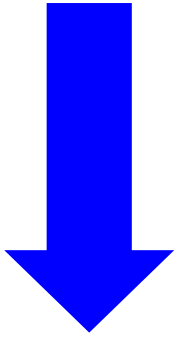
実験的アプローチ

試料の調整

ハイブリダイゼーション

蛍光強度の測定

理論的（情報学的）アプローチ



アレイインフォマティクス

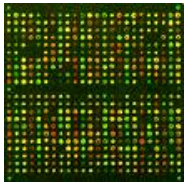
データの正規化
クラスタリング
ネットワーク解析

未知遺伝子の機能発見・遺伝子発現の制御関係の解明

トランスクリプトーム解析とは？(3/3)
-ゲノム創薬への期待-

各種疾患動物モデルや疾患細胞内における・・・

体系的かつ網羅的な遺伝子発現解析

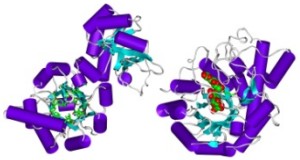


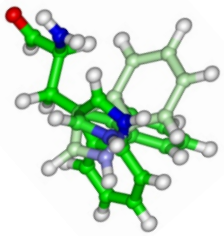
病態に特異的な遺伝子発現パターンから・・・

医薬品開発のターゲット候補遺伝子群の同定

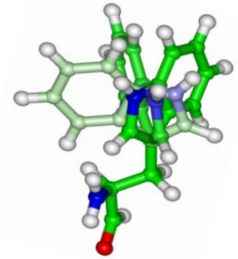
候補遺伝子群の細胞生物学的機能解析によって・・・

標的とする低分子化合物の選択





創薬ターゲットとする分子の決定



トランスクリプトーム解析の実験的手法(1/6)
-マイクロアレイとSAGEの比較-

手法　マイクロアレイ

SAGE（Serial Analysis of Gene Expression）

	マイクロアレイ	SAGE
原理	ハイブリダイゼーション	DNAシークエンス
データの性質	定性的	定量的
mRNA配列情報	事前に必要	不要
解析規模	大規模（全mRNA対象）	少～中規模（選択されたmRNA対象）

トランスクリプトーム解析の実験的手法(2/6)
-マイクロアレイとは？-

ガラスやシリコン製の小基盤上にDNA分子を高密度に配置（アレイ、array）したもの

⇒同時に、数千から数万規模の遺伝子発現を観察が可能

作成原理　　**ハイブリダイゼーション**

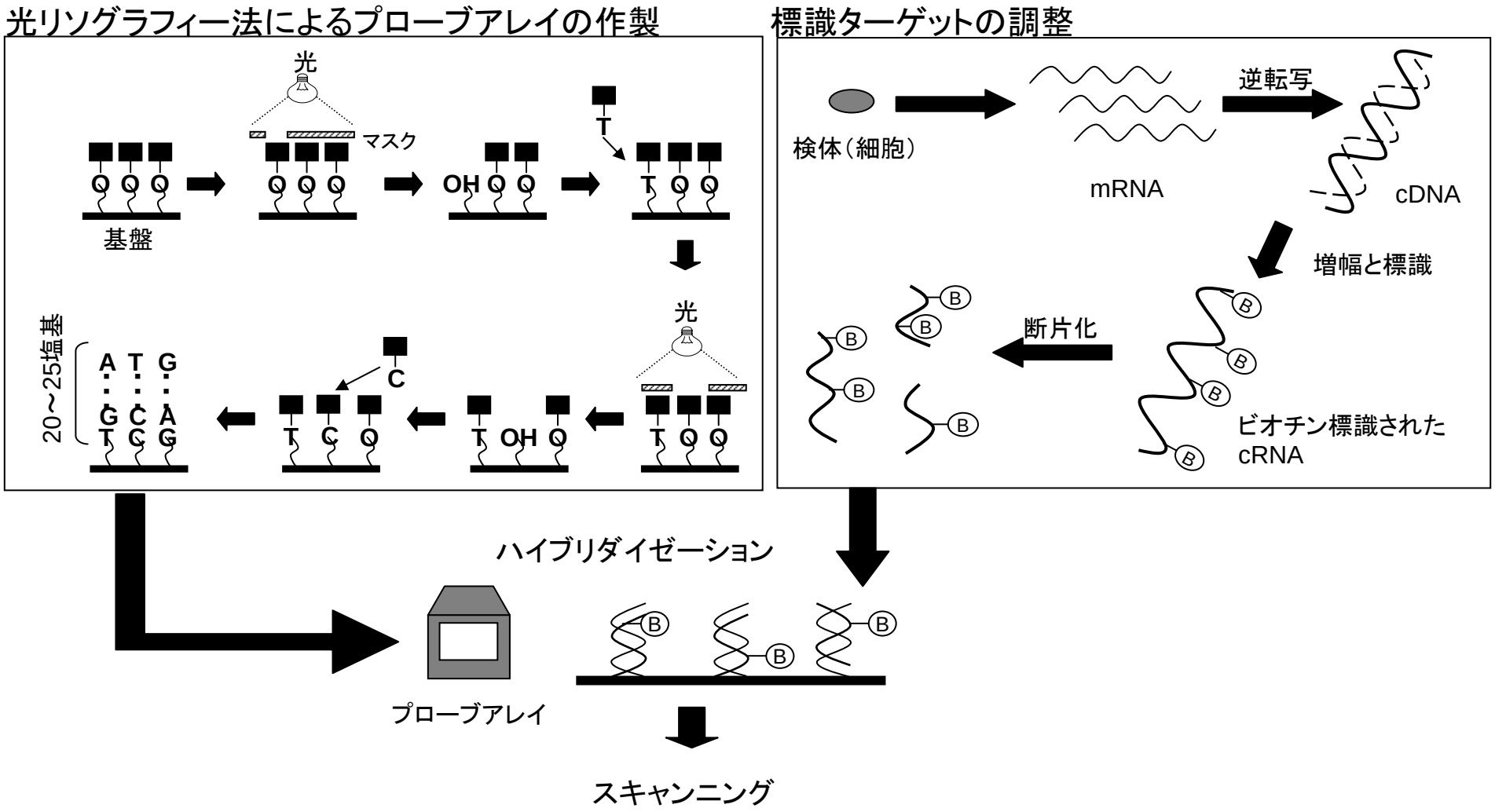
作成方法　　GeneChip技術
スポット型アレイ法　（スタンフォード方式）
タイリングアレイ法

これまでのハイブリダイゼーションを原理とした研究との違い

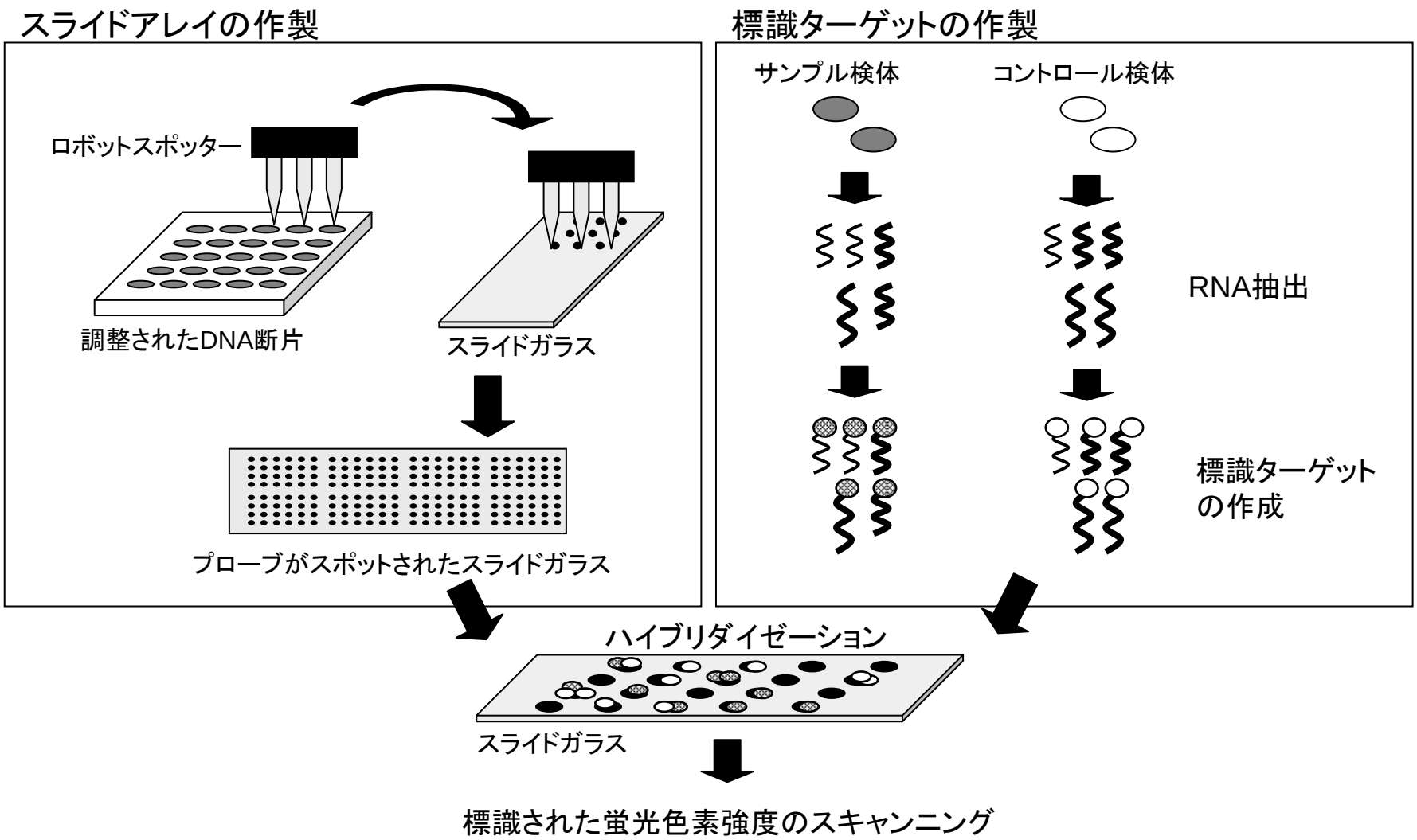
ノーザンブロット・サザンブロット⇒ 1日にせいぜい2，3回の実験

マイクロアレイ⇒1回で数千から数万のハイブリダイゼーション

トランスクリプトーム解析の実験的手法(3/6)
-GeneChip技術-

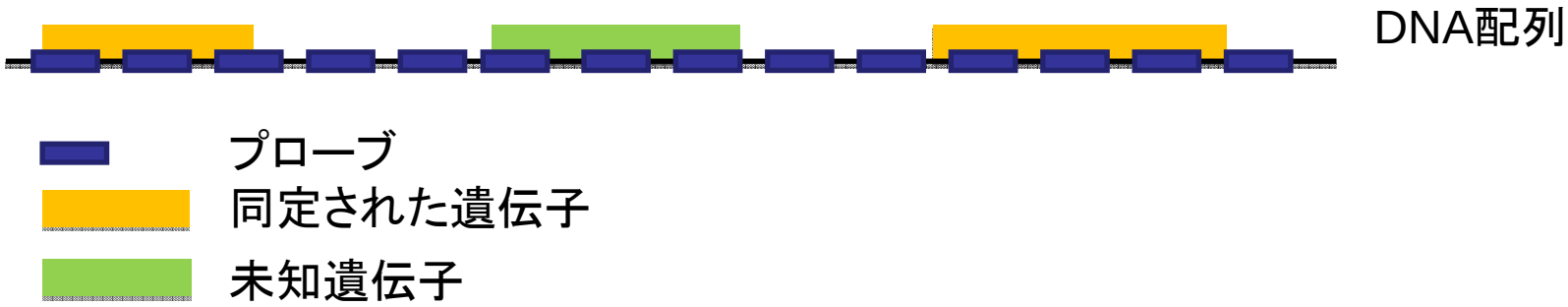


トランスクリプトーム解析の実験的手法(4/6)
-スポット型アレイ法-

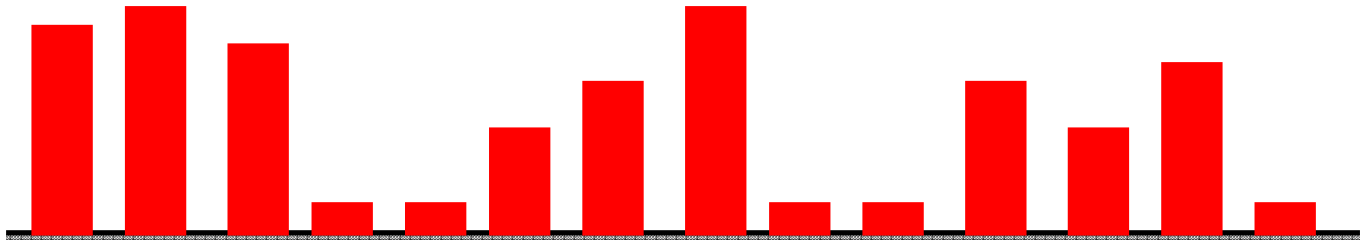


トランスクリプトーム解析の実験的手法(5/6)
-タイリングアレイ法-

解読済みのゲノムデータから**等間隔に抜き出した塩基配列**を検出用プローブとして
タイル状に並べたDNAチップ



発現プロファイル （測定結果）



生体内で転写されたRNAを鋳型 として作った標識cDNAとハイブリダイズさせることで、
RNAに相補的なプローブからシグナルを検出
⇒　未知のRNAについても塩基配列の一部を知ることが可能

トランスクリプトーム解析の実験的手法(6/6)
-発現プロファイルデータとは？-

スポット型アレイで得られるデータとは・・・

Gene name	Intensity	
	Cy3	Cy5
g_a	837	1975
g_b	186	414
g_c	1022	2736
g_d	120	306
⋮	⋮	⋮

数値化された蛍光強度

$$\text{発現量} = \frac{\text{サンプル検体に使用した蛍光色素の強度}}{\text{コントロール検体に使用した蛍光色素の強度}}$$

Gene name	Intensity		ratio
	Cy3	Cy5	
g_a	837	1975	0.423
g_b	186	414	0.449
			0.374
g_d	120	306	0.392
⋮	⋮	⋮	⋮

各遺伝子の発現量

アレイインフォマティクス(1/2)
-目的と必要性-

アレイインフォマティクスとは？

DNAチップやDNAマイクロアレイ等で得られる大量の発現プロファイル情報（個々の遺伝子の発現量)を統計学的手法等により解析し、**遺伝子の機能解析や遺伝子ネットワークの解析を行うための情報処理技術。**

なぜ必要か？

1つのマイクロアレイ実験⇒数千から数万の遺伝子発現プロファイル

実際には**複数、多い場合には数百の実験結果を統合して解析する必要**があり、そのためにはインフォマティクス技術が必要である。

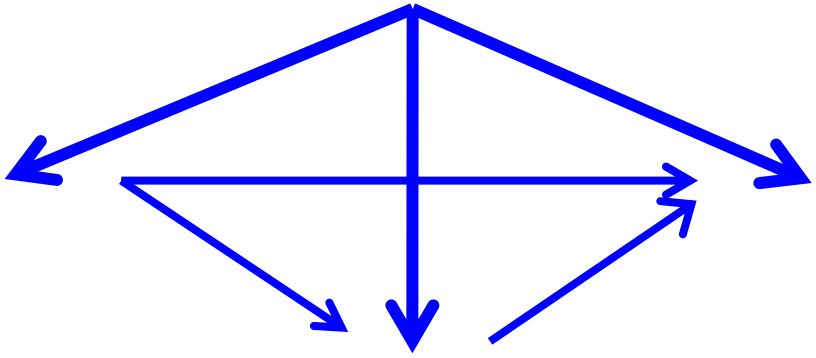
アレイインフォマティクスの主流として・・・

データの正規化
クラスター解析
ネットワーク解析

アレイインフォマティクス(2/2)
-アレイインフォマティクスの流れ-

実験的手法による遺伝子発現情報の測定

測定データの正規化 （標準化）



有意差解析

t-検定

Mann-Whitney U検定

ANOVA

SVM

クラスタリング

階層的クラスタリング

SOM

K-mean

ネットワーク解析

Boolean Model

Bayesian Model

微分方程式 Model

Graphical Gaussian Model

データ正規化(1/7)
-データ正規化の必要性-

国外主要5社におけるマイクロアレイの特徴

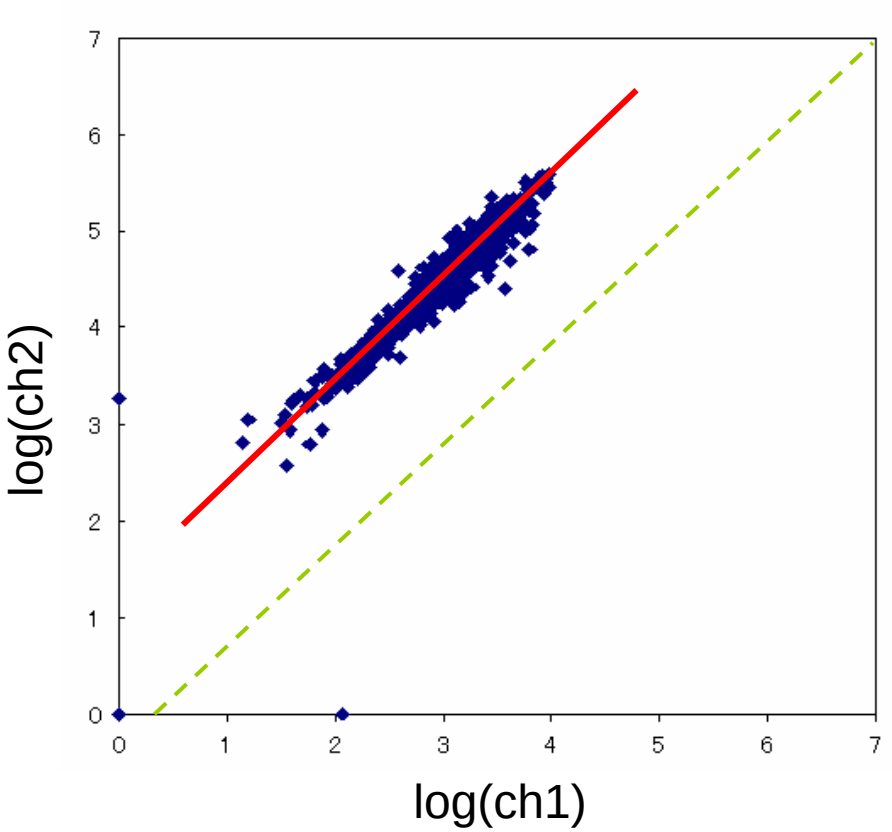
会社名	プローブ数	特徴
アプライドバイオシステムズ	32,878	1色法 1500bp以内
アフィメトリクス	54,120	1色法 25mer, 22probes/gene
アジレントテクノロジーズ	41,000	1色法、2色法 66mer
アプライドマイクロアレイズ	54,841	1色法、30mer, ave: 424bp
イルミナ	48,701	1色法、50mer

＋国内： DNAチップ研究所・東レ・三菱レイヨン・タカラバイオ・・・

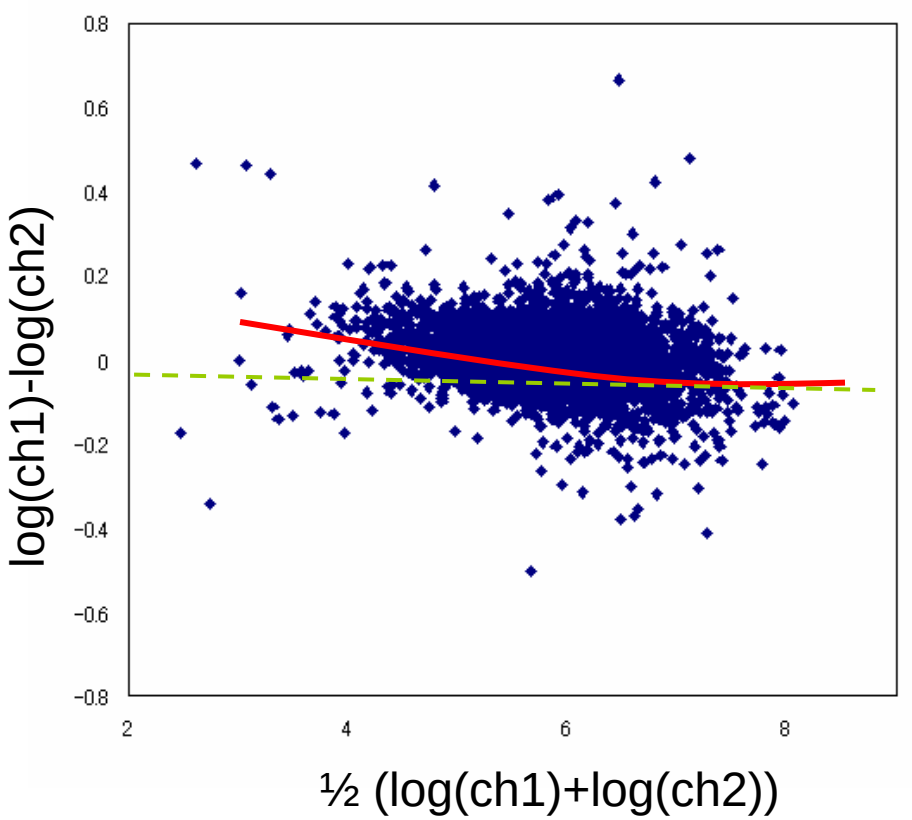
各会社・製品によってデータのプラットフォームが異なる

データ間の比較・解析のためには正規化(標準化)が必要

データ正規化(2/7)
-raw dataの傾向-



細胞間（サンプルvsコントロール）のmRNA量の違い
Cy3,Cy5 の蛍光色素が本来持っている色の特性の違い



MAプロット

M=log(ch1)-log(ch2)=log(ch1/ch2)

A=1/2 (log(ch1)+log(ch2))

データ正規化(3/7)
-線形回帰による補正-

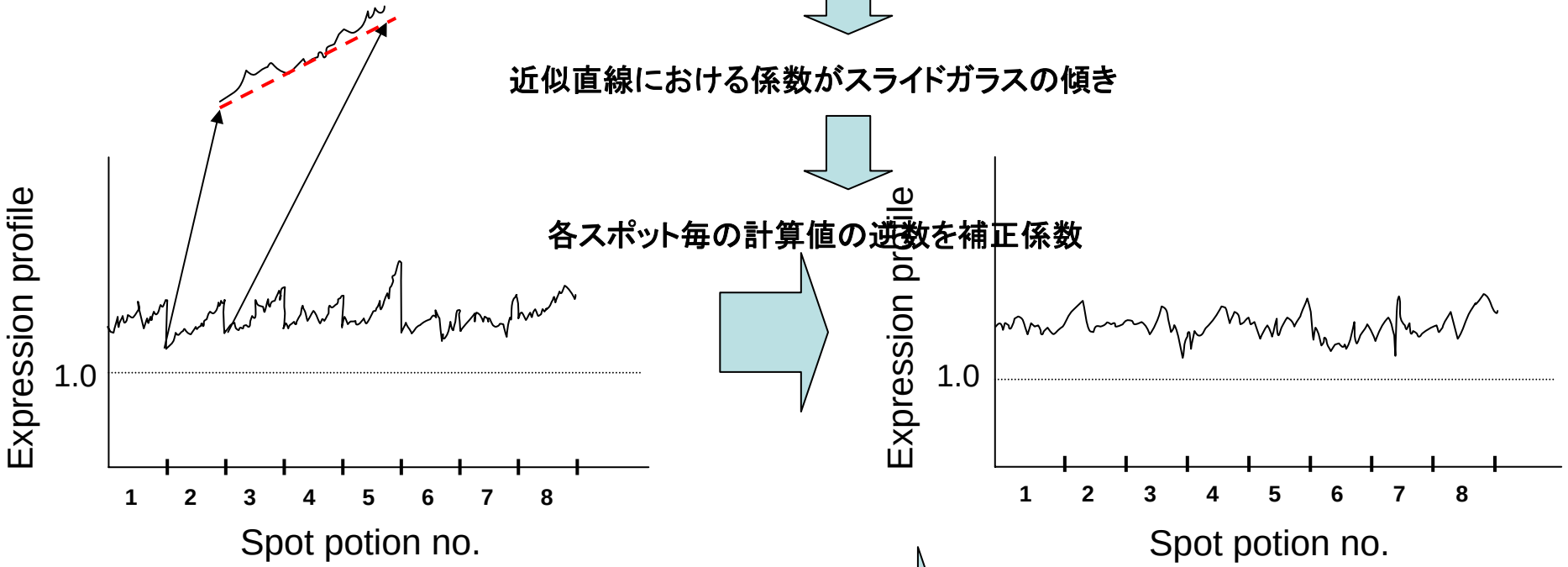
Raw data

一次の近似直線を算出

Pre-normalized data

近似直線における係数がスライドガラスの傾き

各スポット毎の計算値の逆数を補正係数

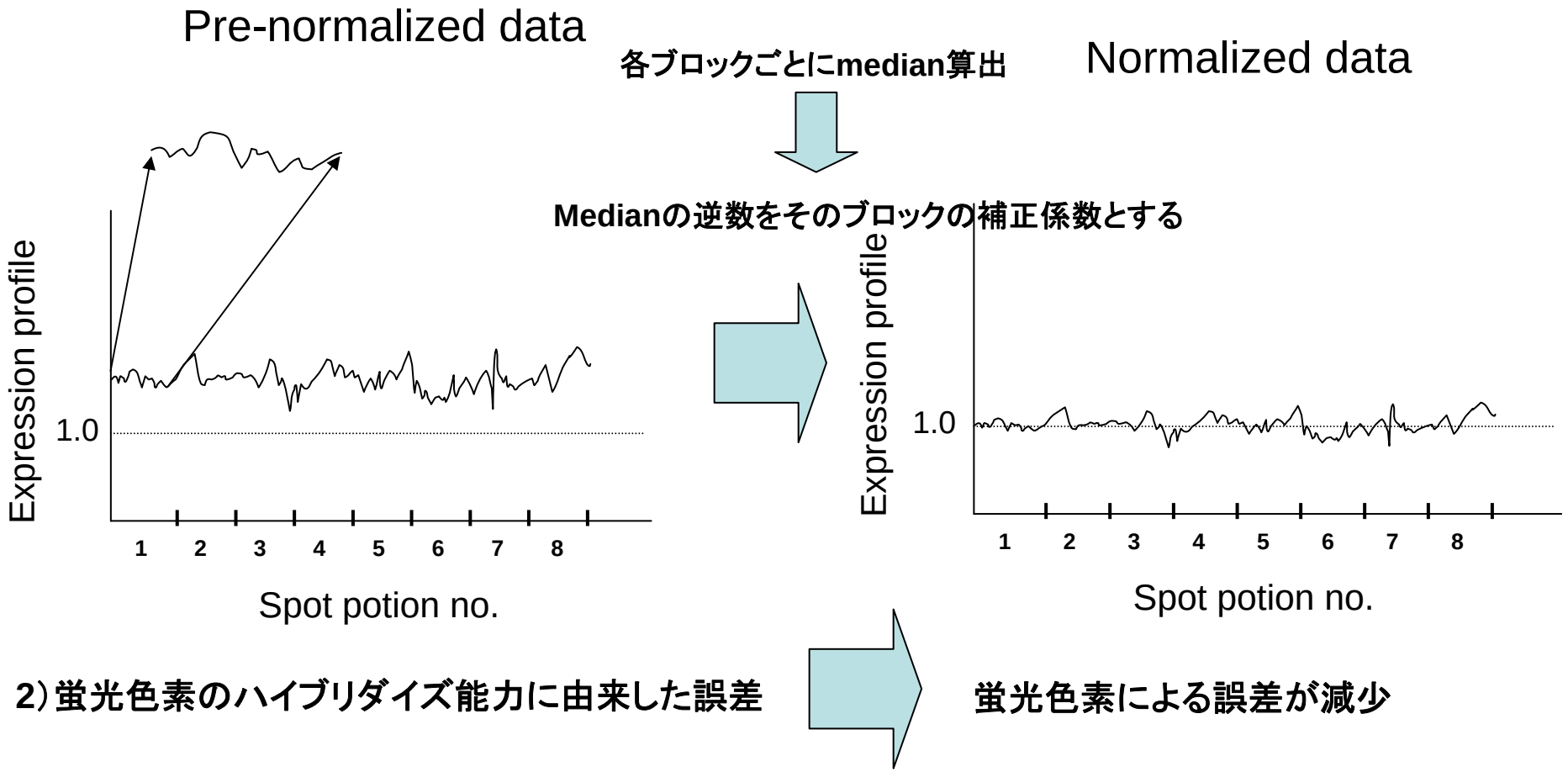


1）スライドガラスの傾きによる物理的な誤差

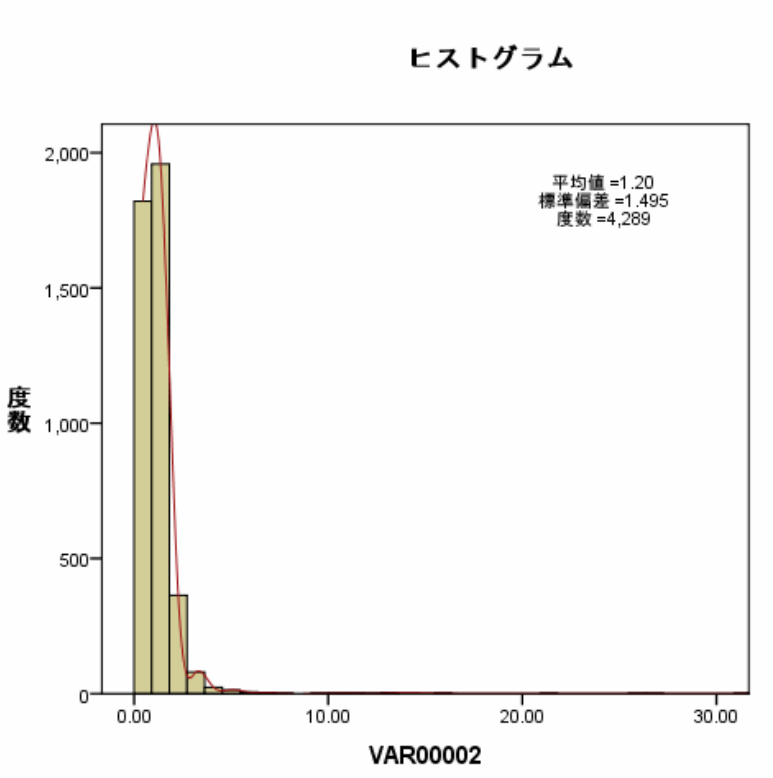
2）蛍光色素のハイブリダイズ能力に由来した誤差

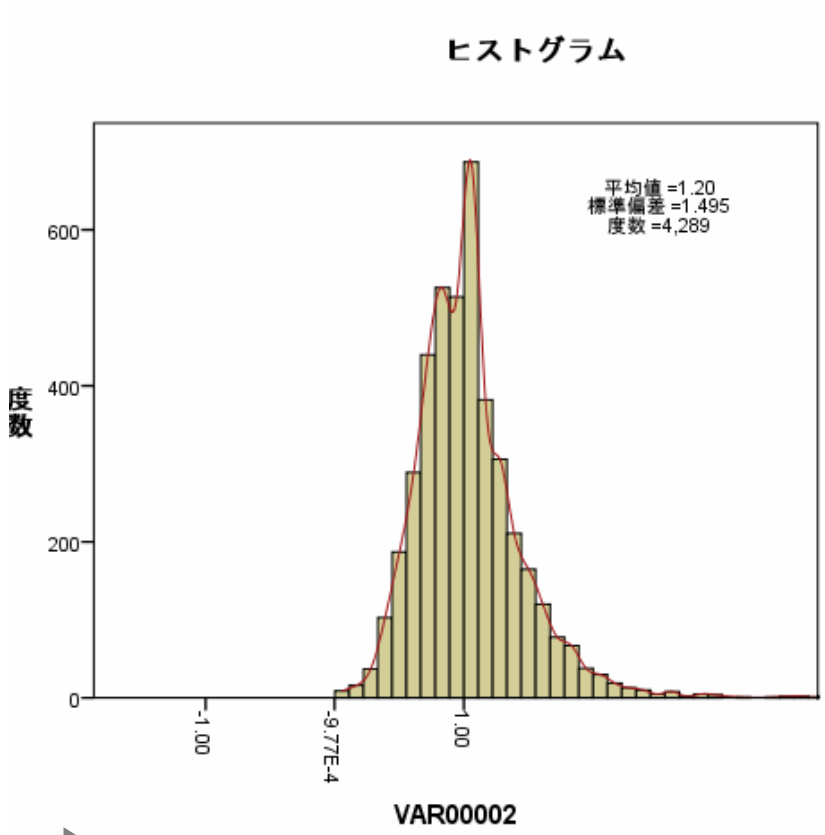
傾きによる傾向が消失

データ正規化(4/7)
-normalization -



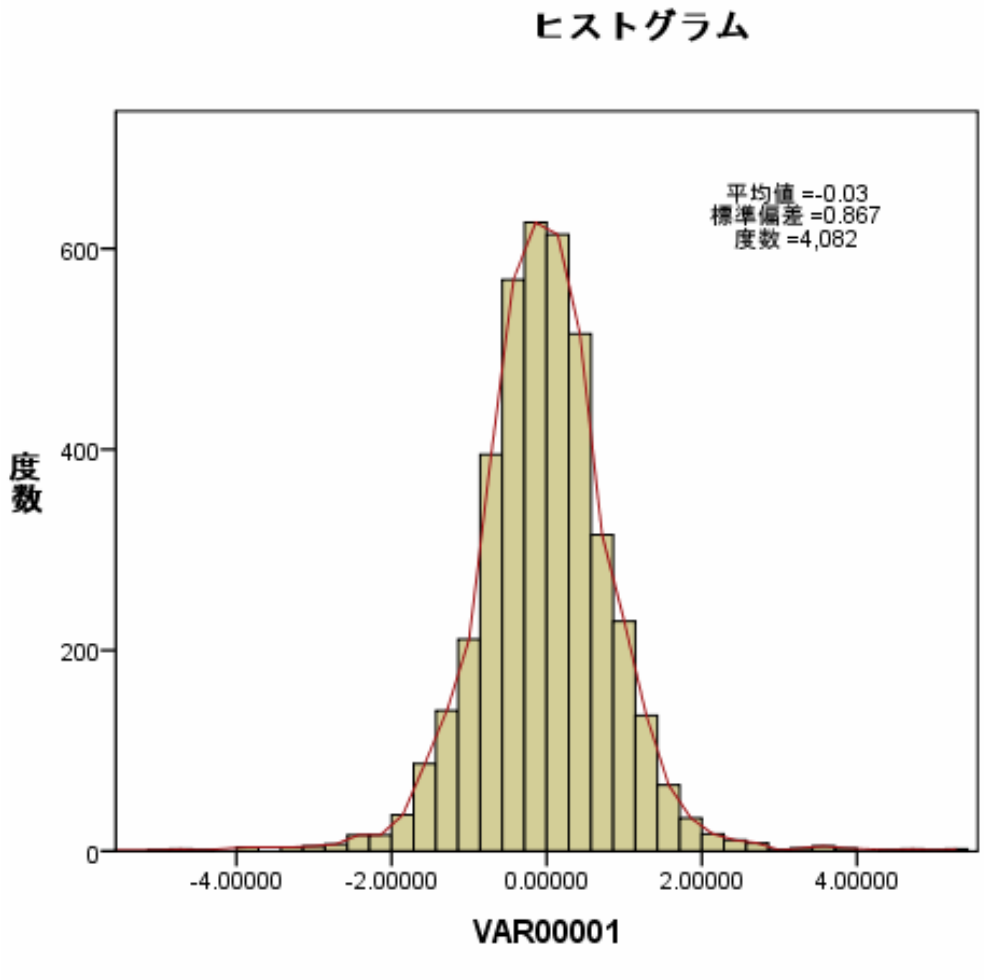
データ正規化(5/7)
-データ分布からの正規化-





対数変換 → 正規分布近似

データ正規化(6/7)
-外れ値の除去-



Z変換

$$Z = \frac{x_i - \bar{x}}{s}$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

標準正規分布$N(0,1)$

外れ値の検出 （Grubbs検定）

データ正規化(7/7)
-種類と特徴-

- **1色法アレイ**
  - バックグラウンド補正
    - 基板特有の傾向の除去
  - 正規化　(quantile normalization)

- **2色法アレイ**
  - global normalization
    - ターゲットサンプルとリファレンスサンプルでの発現比の中央値(平均値）が実験間で同じになるように処理
    - 細胞・組織での経時変化など同一細胞・組織内での解析
  - internal control normalization
    - ターゲットとリファレンスの状態が大きく異なっている場合、実験間で発現変動していないハウスキーピング遺伝子の発現比が同じになるように処理
    - アポトーシスなど遺伝子の発現量が大きく変化する場合

クラスター解析(1/8)
-クラスター解析の必要性-

発現パターンが似ている遺伝子はどれか？

	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5
$gene_1$	**0.03**	0.65	**2.78**	1.01	**3.07**
$gene_2$	**0.50**	0.73	**4.08**	0.89	**2.06**
$gene_3$	1.00	**2.00**	1.05	5.00	**0.04**
	…	…	…	…	…
$gene_n$	…	…	…	…	…

問題点

- 何をもって発現パターンの類似を決定するか？
- 発現パターンが類似した遺伝子としていない遺伝子の区別はどこでつけるか？

クラスター解析(2/8)
-クラスター解析とは？-

クラスター解析とは？

個々の遺伝子の発現データをもとに、**遺伝子のグループ分けを行う統計的手法**。

クラスター解析の手法

階層的クラスタリング

非階層的クラスタリング　K-mean法
SOM

クラスター解析の目的

遺伝子機能予測

プロモーター領域における調節要素の探索

データの統合

クラスター解析(3/8)
-階層的クラスタリングの方法-

群平均法 (group average method)
UPGMA (Unweighted Pair-Group Method using arithmetic averages

$$d(P,Q) = \frac{1}{|P||Q|}\sum_{p\in P}\sum_{q\in Q} d(p,q)$$

最短距離法 (nearest neighbor method)
単連結法　(Single Linkage Clustering method)

$$d(P,Q) = \min_{p\in P, q\in Q} d(p,q)$$

最長距離法 (furthest neighbor method)
完全連結法 (Complete Linkage Clustering method)

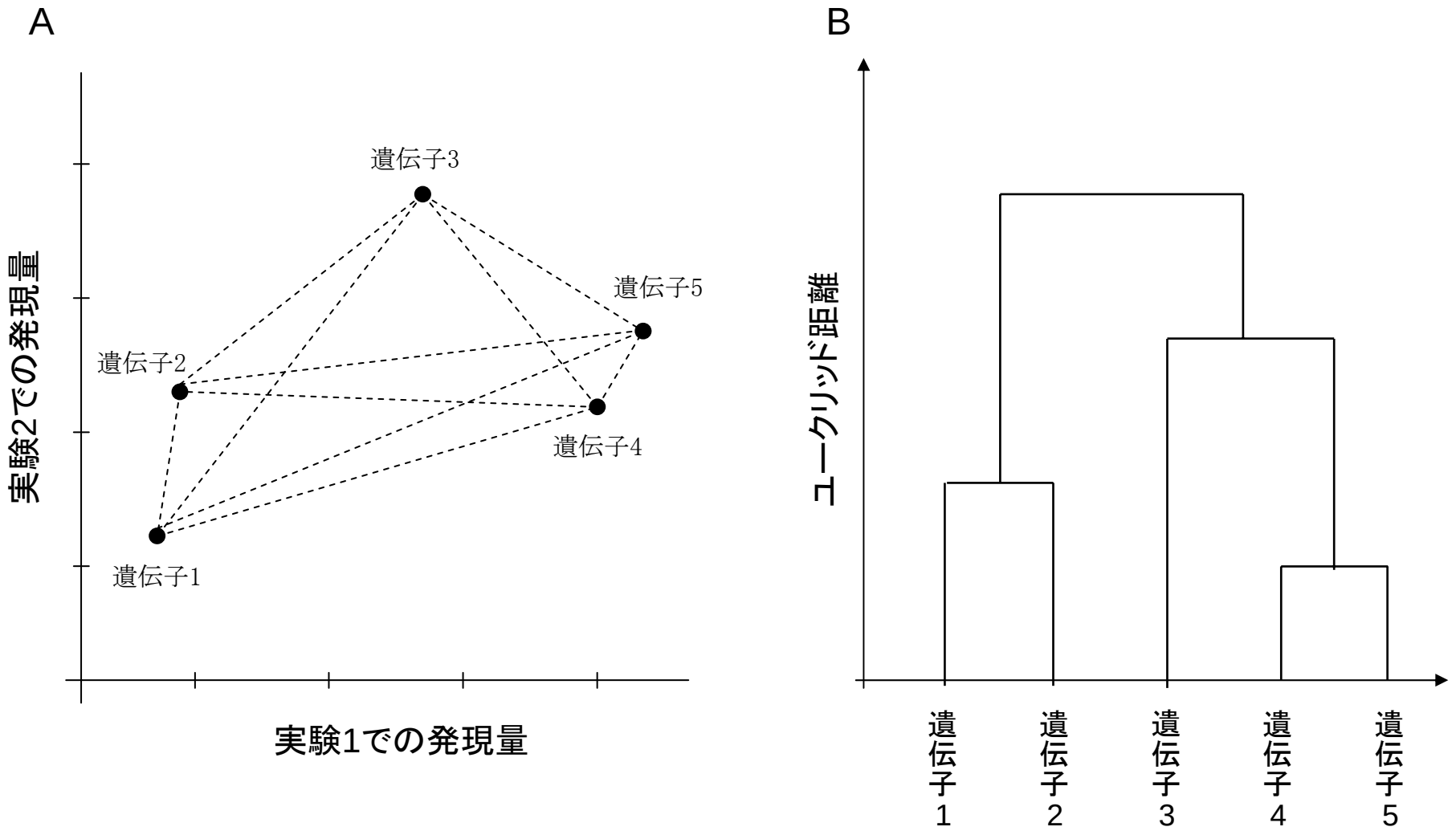
$$d(P,Q) = \max_{p\in P, q\in Q} d(p,q)$$

ウォード法 (Ward's method)

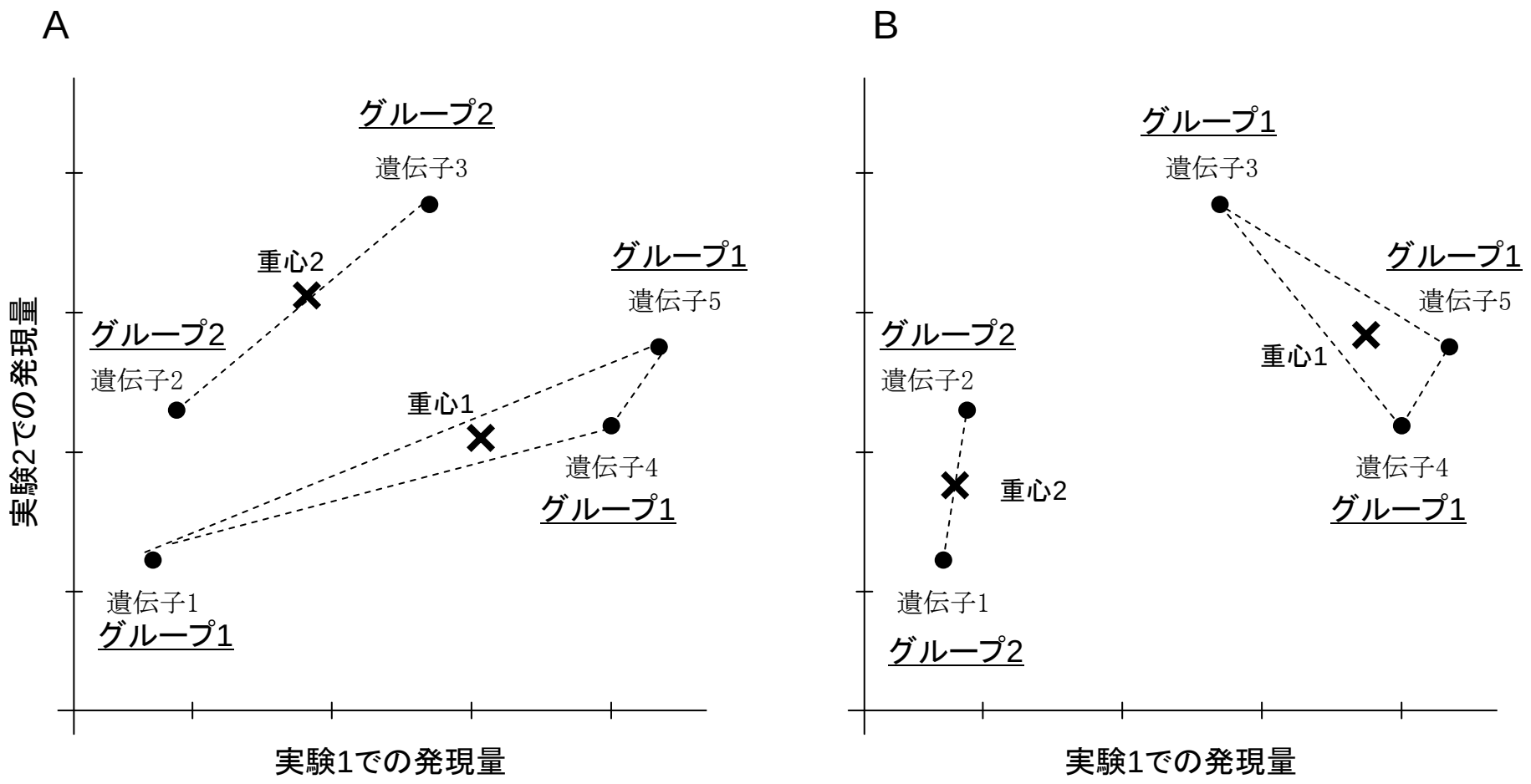
$$d(P,Q) = E(P\cup Q) - E(P) - E(Q)$$

$$E(P) = \sum_{p\in P} (d(p-c))^2$$

クラスター解析(4/8)
-階層的クラスター解析-



クラスター解析(5/8)
-非階層的クラスター解析（k-mean法）-

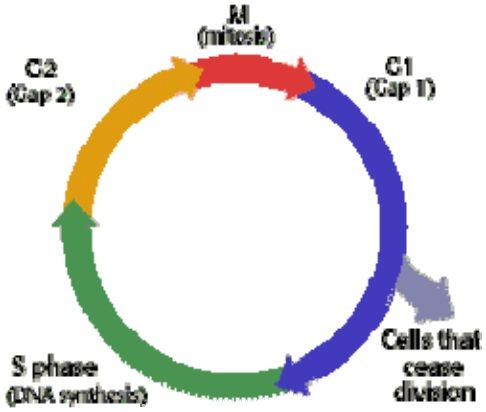


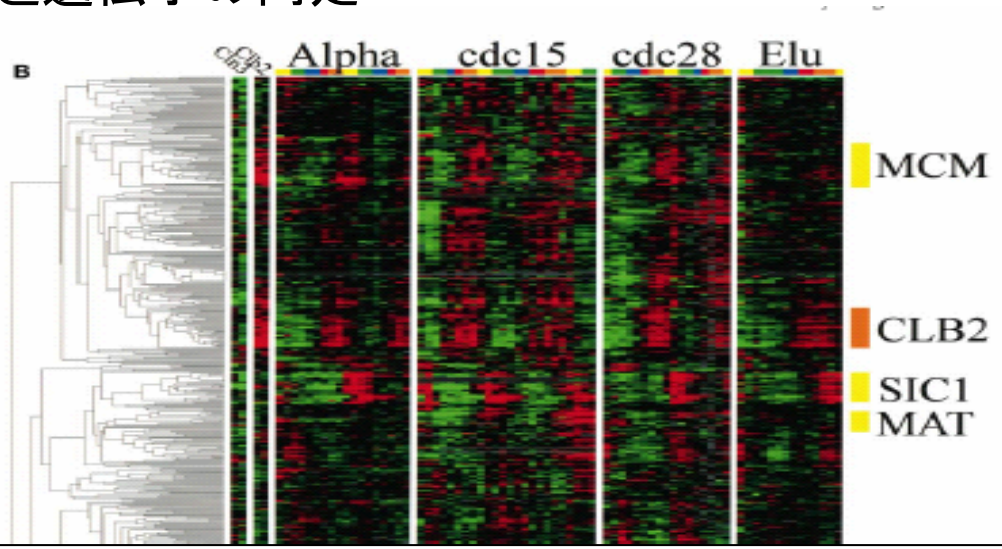
クラスター解析(6/8)
-Stanfordが開発したツール-

XCluster	Michael Eisen	Hierarchical clustering Self-organizing map
SAM	Rob Tibshirani	Data Mining
ScanAlyze	Michael Eisen	Fluorescent Image
SMD Package	SMD Staff	Database
TreeView	Michael Eisen	Graphically Browse
XCluster	Gavin Sherlock	Hierarchical clustering Self-organizing map
KNNimpute	Olga Troyanskaya	Estimation of missing value

クラスター解析(7/8)
-クラスター解析例　細胞周期-

*S.cerevisiae*のcell cycle関連遺伝子の同定





⇒約800個の細胞周期関連遺伝子群の同定
⇒各phaseにおいて発現している遺伝子群の同定

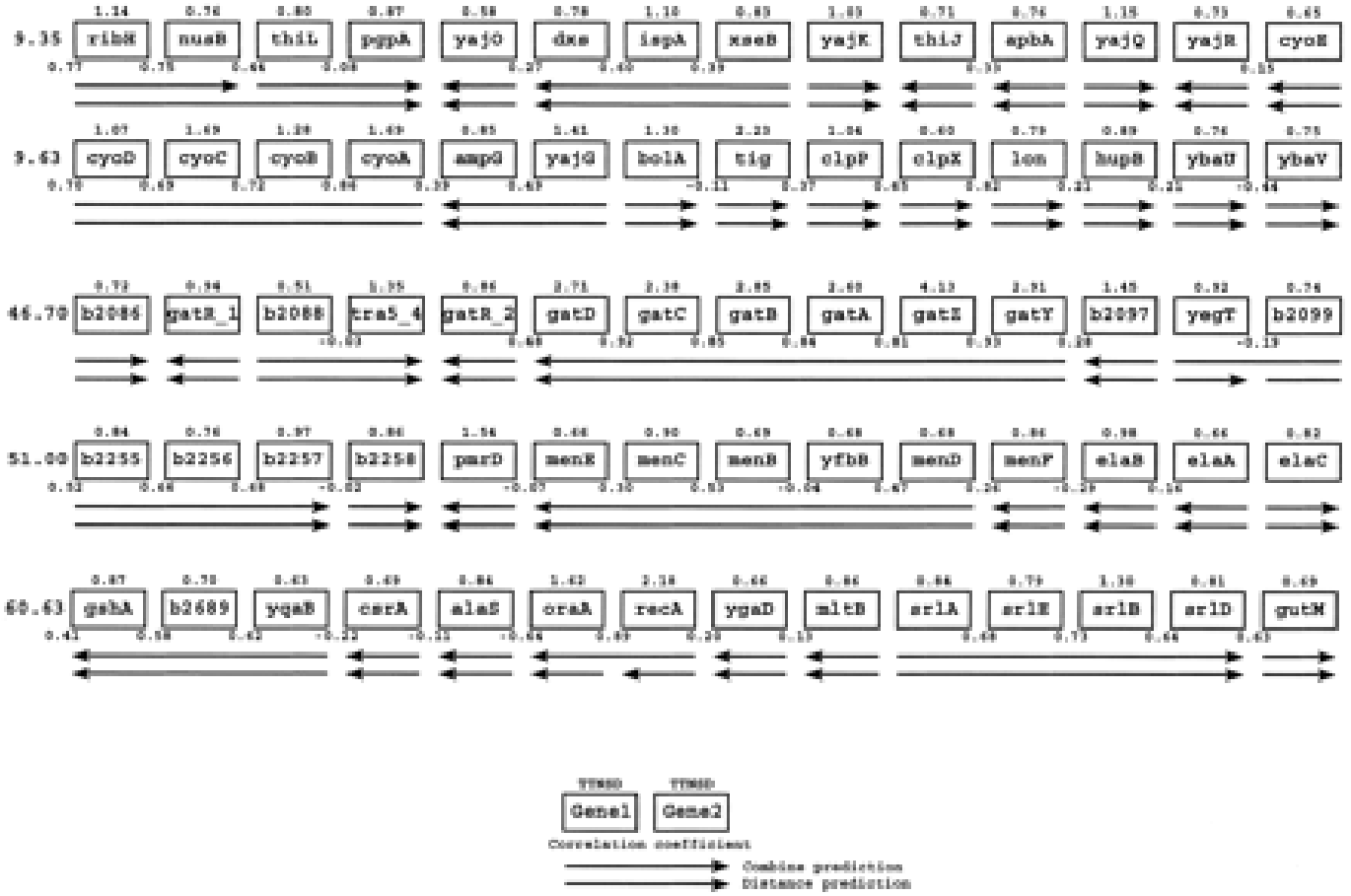
細胞周期のメカニズム解明への一歩

Figure 1. (cont).

Spellman, P.T. *et al. Mol. Cel. Bio.* Vol.9 1998:3273-97

クラスター解析(8/8)
-クラスター解析の応用例　オペロン予測-

*E.coli*のoperon prediction



Savatti, S., et al. *Nucleic acids Res.* 2002:2886-2893