

---

# バイオデータマイニング

森下  
情報生命科学専攻

講義ノート URL

<http://mlab.cb.k.u-tokyo.ac.jp/~moris/lecture/>

# 予定

---

- 4/10, 17, 24, 5/1 森下
- 5/8, 15, 22, 29, 6/5, 12 中谷
- 6/19, 7/3, 10 森下 (6/26 休講)
- 7/17 筆記試験

森下の講義は以下のトピックを扱います

- 導入 人工知能の歴史とデータマイニング
- 発現データ、表現型データのクラスタリング
- クラス分類(教師つき学習)の代表的アルゴリズム  
決定木、Boosting, Support Vector Machine
- 大規模ゲノムアセンブリ
- 脊椎動物の比較ゲノム用アルゴリズム
- ゲノムを利用した特異的配列設計

# 人工知能の歴史

---

- 全能の神のような技術？
- 人工的な知能をつくることの難しさを知る研究の歴史
- 魅力的な言葉の創造
  - 1970年代      人工知能 (Artificial Intelligence)
  - 1980年代      エキスパートシステム (Expert System)  
機械学習 (Machine Learning)
  - 1990年代      知識発見技術 (Knowledge Discovery)  
データマイニング (Data Mining)
  - 2000年代      Web Mining
- 実際は、人工的な知能が実現できそうな問題をさがす

# 人工知能の流れ — 決定不能問題

---

- 定理の証明のように高度な知能が要求される活動を、機械により置き換えることができるか？
- 残念ながら機械的に証明できる問題の範囲は非常に狭い

帰納法による証明が必要な自然数に関する基本的性質でさえ自動証明は不可能

1931年 ゲーデルの不完全性定理

# 人工知能の流れ — 決定不能問題

---

- デバッグを自動化できるか？
  - プログラムの性能保障を自動化することは可能か？
  - どのような入力に対しても必ず停止することを保証できるか？  
仕様どおりに動作するか？
  - 決定不能問題
- 
- 入門書 Douglas R. Hofstadter: *Gödel, Escher, Bach: An Eternal Golden Braid*
  - 専門書 Elliot Mendelson: *Introduction to Mathematical Logic, Fourth Edition*

# 手におえないほど計算時間がかかる問題

---

- しらみつぶし的に解けば、明らかに自動的に解ける問題であれば、計算機は答えをだしてくれるか？
- 問題のサイズが大きくなると  
手におえないほど計算時間がかかる問題が存在

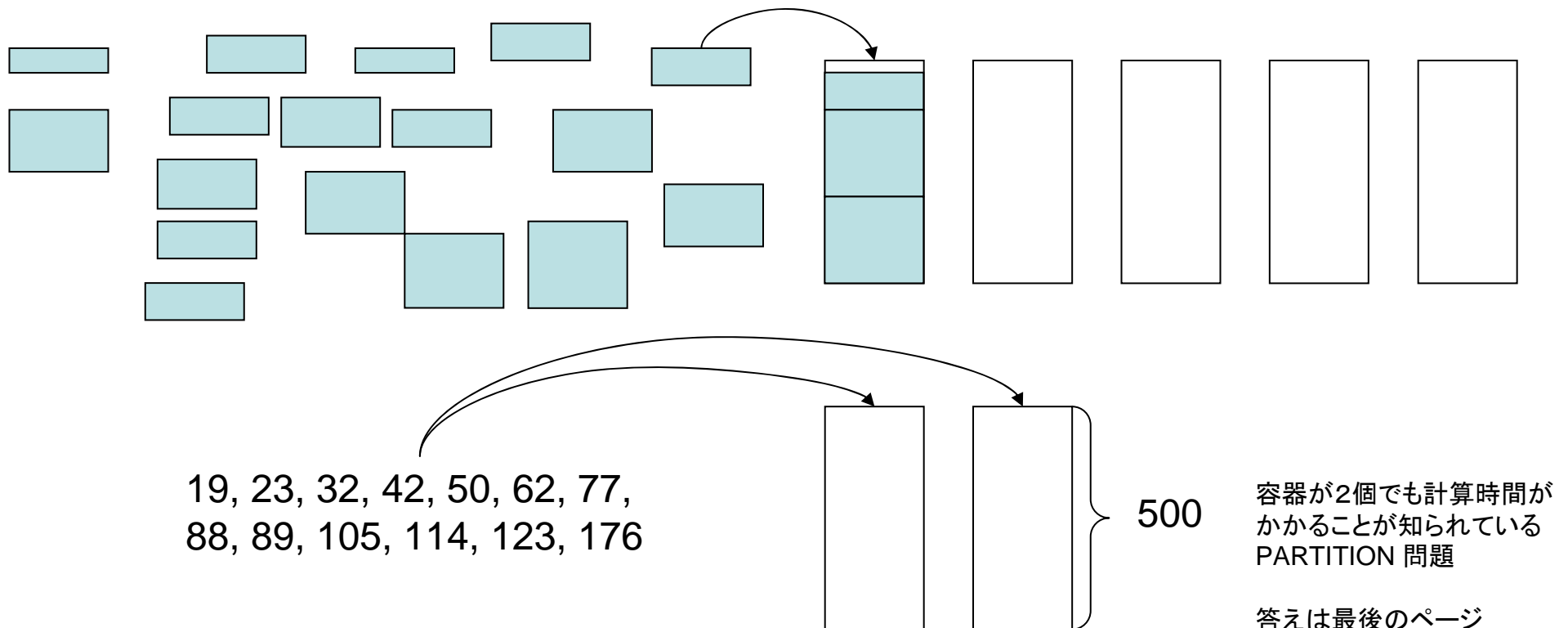
1972年 Cook によるクラスの提示  
Karp による様々な例の提示

- 専門書 Michael R. Garey and David S. Johnson: *Computers and Intractability - A Guide to the Theory of NP-Completeness*
- 効率的なアルゴリズムが存在するか否か (P vs NP 問題)  
クレイ数学研究所が2000年に挙げた7つの未解決問題の1つ

# データ数が増えると 手におえないほど計算時間がかかる問題

- Bin Packing: 長さが正の整数  $\text{size}(u)$  である積木  $u$  の集合  $U$  を  $k$  個の部分集合  $U_i (i=1, \dots, k)$  に分ける。

各部分集合を正の整数  $B$  の大きさの容器に格納できるように分けられるか？  
 $\sum \{ \text{size}(u) \mid u \in U_i \} \leq B$



# バイオインフォマティクス関連の問題① 配列断片アセンブリ

- Shortest (Common) Superstring Problem

文字の集合  $\Sigma$ 、文字列の集合を  $R$  とする。  
 $R$  に含まれるどの文字列も、連続した部分列(string) として含む  
文字列で最短なものは？

- $\Sigma = \{A, T\}$   
 $R = \{\text{AAA}, \text{AAT}, \text{ATA}, \text{ATT}, \text{TAA}, \text{TAT}, \text{TTA}, \text{TTT}\}$

AAAAATATAATTTAATATTTATTT  
AAAAATATAATTAAATATTTATTT

AAATAATTATTT  
AAA ATT  
AAT TTA  
ATA TAT  
TAA TTT

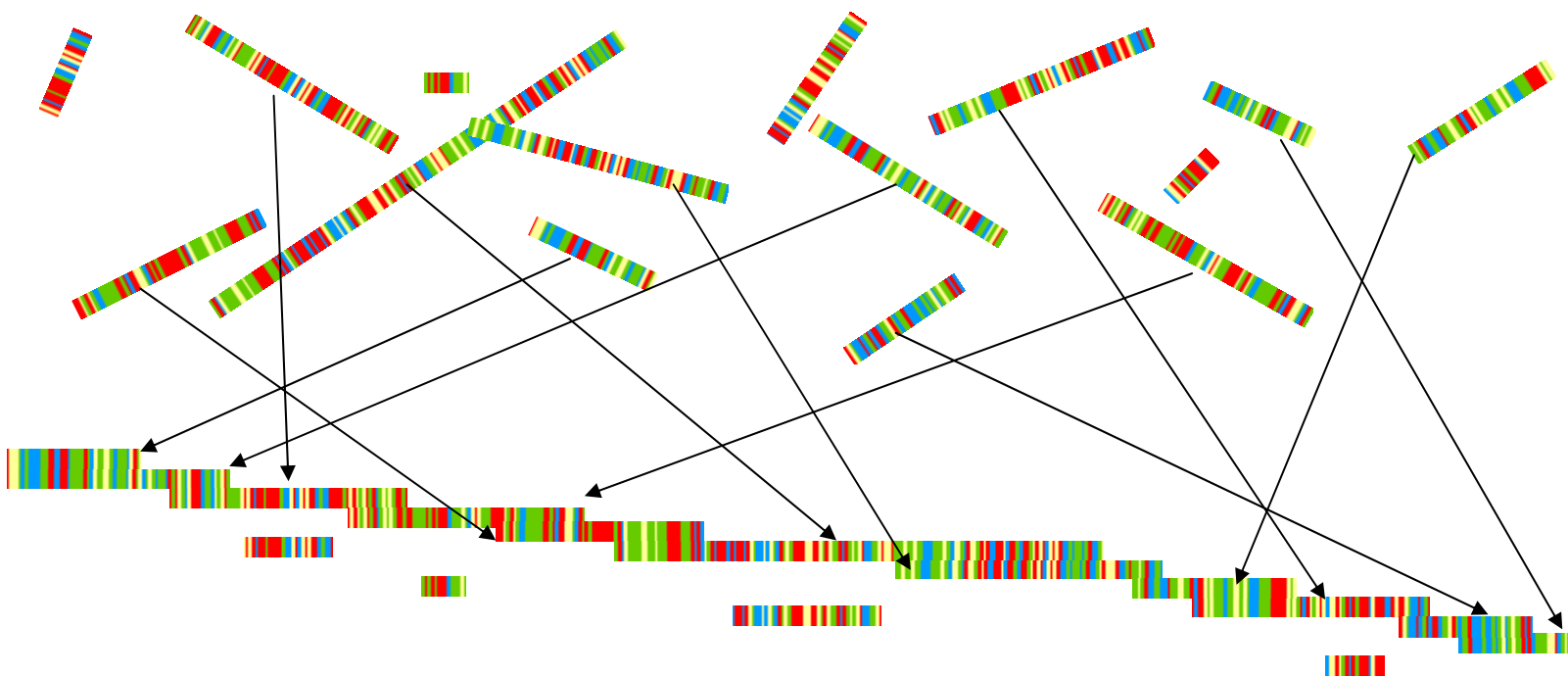
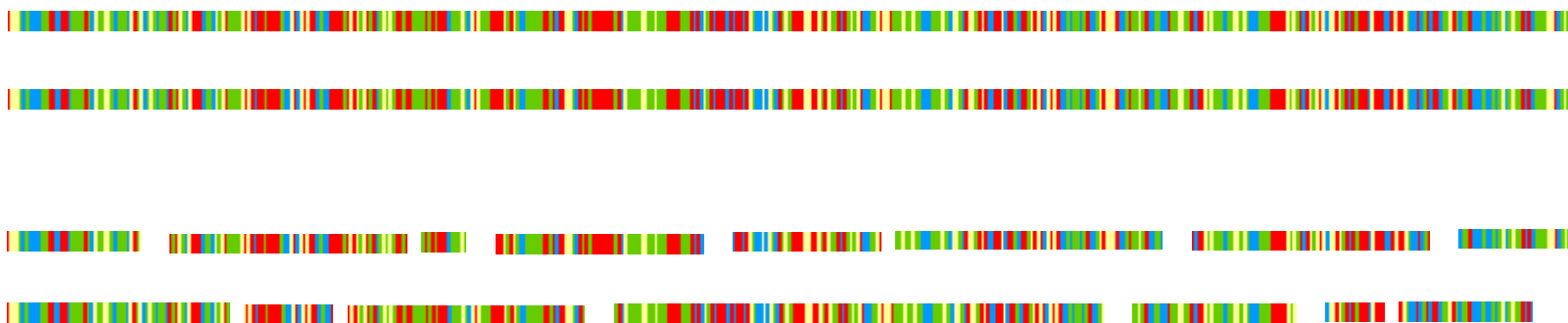
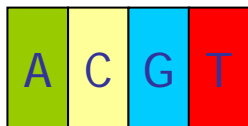
- NP困難問題:  $R$  の元の数を  $n$  とすれば  $2^n$  に比例する計算時間がかかるアルゴリズムしか知られてない



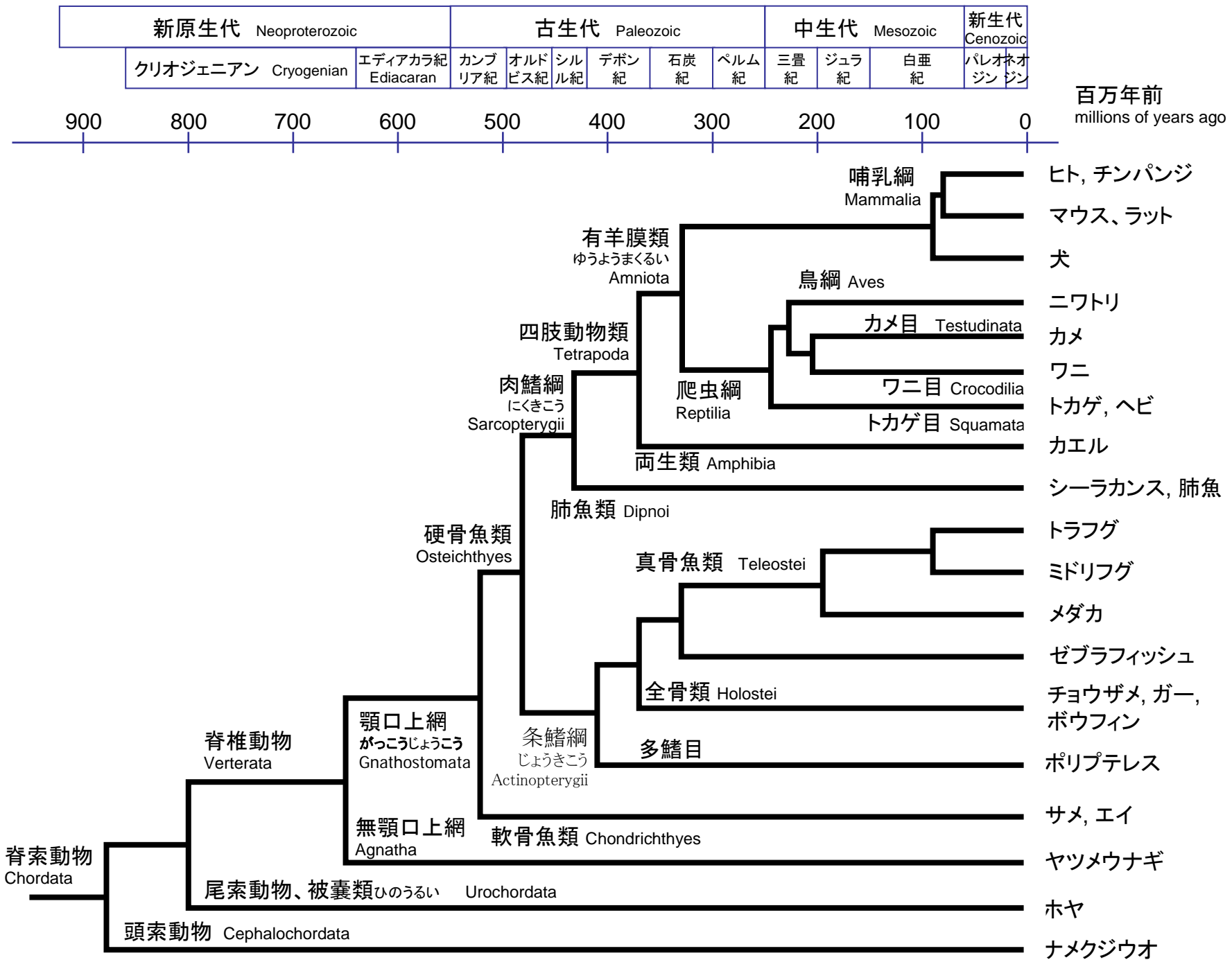
## バイオインフォマティクス関連の問題① 配列断片アセンブリ

---

- ゲノムアセンブリ: 断片配列の集合  $R$  から元の配列を復元
- Shortest Superstring 問題として捉えることもできる
- しかし、このように定式化して最適解を求めようとすると、集合  $R$  が大きくなるつれ手におえないほどの計算時間がかかる
- 現実的な時間で解けるような定式化が必要！



元の絵がわからないジグソーパズル(数百万～数千万ピース)



# 大規模ゲノムアセンブリの状況

論文発表 種	総塩基数	アセンブリ方式 (赤字は WGS 方式と開発機関)	
2001 / 2	ヒト	29億	clone-by-clone 国際チーム
			Celera アメリカ
2002 / 4	イネ(染色体地図なし)	4.7億	RePS (Beijing Genomics) 中国
2002 / 7	トラフグ(染色体地図なし)	3.6億	Jazz (JGI) アメリカ
2002 / 12	マウス	25億	Arachne (MIT) + clone by clone アメリカ
2004 / 2	カイコ(染色体地図なし)	5億	Ramen (東大) 農業生物資源研究所 日本
			RePS (Beijing Genomics) 中国
2004 / 4	ラット	25億	Atlas (Baylor College) + clone by clone アメリカ
2004 / 10	ミドリフグ	3.4億	Arachne (MIT) フランス, アメリカ
2004 / 12	チキン	10億	PCAP (Wash. U.) アメリカ
2005 / 8	イネ	3.9億	clone by clone 農業生物資源研究所 日本
2005 / 9	チンパンジ	29億	PCAP, Arachne アメリカ
2005 / 12	ドッグ	24億	Arachne (MIT) アメリカ
2007 / ?	メダカ	7億	Ramen (東大) 国立遺伝学研究所 日本
2007 / ?	ナメクジウオ	6億?	Jazz (JGI) アメリカ
2008 / ?	ゼブラフィッシュ	16億	Phesion (Sanger Ctr.) イギリス
2008 / ?	アフリカツメガエル	16億	Jazz (JGI) アメリカ

## バイオインフォマティクスの問題② マルチプルアラインメント

- Longest Common Subsequence

文字の集合  $\Sigma$ 、文字列の集合を  $R$  とする。  
 $R$  のどの文字列にも含まれ、必ずしも連続していない  
文字列のなかで最長のものは？

ATATATA

TATTAAT

ATAATAT

ATTATAA

ATATATA

TATTAAT

ATAATAT

ATTATAA

ATATATA

TATTAAT

ATAATAT

ATTATAA

-AT-A-TATA

TATTAAT---

-AT-AATAT-

-ATTAT-AA-

## バイオインフォマティクスの問題② マルチプルアラインメント

- $R$  の元が多くなると時間がかかる

NP困難問題:  $R$  の元の数  $n$  とすれば  $2^n$  に比例する計算時間がかかるアルゴリズムしか知られていない

- アラインメント、マルチプル・アラインメント、配列モチーフの発見を Longest Common Subsequence として定式化すると手におえない計算時間  
⇒ やはり現実的な時間で解ける問題に

- $R$  の元が2つに固定されていると動的計画法で解ける

Needleman-Wunsch, Smith-Waterman

**A**T**A**T**A**T**A**  
**T**A**T**T**A**A**T**

# ゲノム進化

CTACCCTTGCATACACCCACCCAGGGGCTTCAAGCCTTGCAGCTGAGTAAACACAGAAA  
CTACCCTTGCATACACCCACCCAGGGGCTCCAAGCCTTGCAGCTGAGTAAACACAGAAA  
CTACCCTTGCATACACCC-ACCCAGGGGCTCCAAGCCTTGCAGCTGAGTAAACAGAGGAC  
CTACCCTTGCATACACCC-ACCCAGGGGCTCCAAGCCTTGCAGCTGAGTAAACACAAGAC

ヒト	C	T	C	G	A
チンパンジー	C	C	C	G	A
ゴリラ	-	C	G	G	C
オラウータン	-	C	C	A	C

MSH2 gene, promoter region nucleotides 61 – 120

[http://sayer.lab.nig.ac.jp/~silver/our\\_data.html](http://sayer.lab.nig.ac.jp/~silver/our_data.html) 齊藤成也先生のHPより

homeobox A1 protein    **Refseq:** [NM\\_005522](#)    **Protein:** [P49639](#) (aka HXA1\_HUMAN)

Human	GACAATGCAAGAATGAACTCCTTCCTGGAATACCC---CATA
Chimp	GACAATGCAAGAATGAACTCCTTCCTGGAATACCC---CATA
Mouse	GACAATGCAAGAATGAACTCCTTT <b>T</b> CTGGAATACCC---CAT <b>C</b>
Rat	GACAATGCAAGAATGAACTCCTTT <b>T</b> CTGGAATACCC---CAT <b>C</b>
Dog	GACAATGCAAGAATGAG <b>G</b> CTCCTTCCTGGAATACCC---CAT <b>C</b>
Chicken	GACAAT <b>A</b> CTAG <b>G</b> ATGAACTCCTTC <b>TT</b> AGAGTAT <b>G</b> C--- <b>A</b> ATT
Fugu	-ACAATGCC <b>C</b> ACAATGAG <b>C</b> AGCTTC <b>TT</b> AGATTACT <b>C</b> --- <b>T</b> GT <b>G</b>
Zebrafish	GA <b>A</b> GATG <b>A</b> CACAATGAG <b>C</b> ACATTCT <b>TT</b> AGATTTT <b>T</b> CGTCCATA

アミノ酸  
同義置換

M	N	S	F	L	E	Y	P	I
ATGAACTCCTTCCTGGAATACCCC								ATA
AGCTTT <b>TT</b> AGAGTAT								ATC
								ATT
S		T			D	F	V	
AG <b>C</b> AC <b>A</b>				GATTTT		GT <b>G</b>		

Human	GACAATGCAAGAATGAACTCCTTCCTGGAATACCC---CATA
	xx       xx   x           x   x   x       x   x     x   xxx
Zebrafish	GAAGATGACACAATGAGCACATTCTTAGATTTTTCGTCCATA
Fugu	-ACAATGCCACAATGAGCAGCTTCTTAGATTACTC---TGTG
	xx       x                     xx               xx     xx   x
Zebrafish	GAAGATGACACAATGAGCACATTCTTAGATTTTTCGTCCATA



# 化石記録と分子時計

1962年

ポーリング ズッカーカンドル

- ヘモグロビン  $\alpha$  鎖中の141個のアミノ酸の置換数を調査
- 化石から分かる分岐時期と置換率は高く相関
- 置換速度はほぼ一定であることを示唆  
⇒置換率から分岐年代を測定

	ゴリラ	ウマ	イヌ	イモリ	コイ
ヒト	1	18	23	62	68

# ミトコンドリアDNAの利用

---

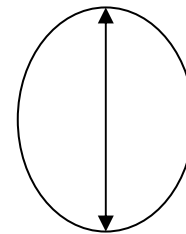
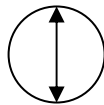
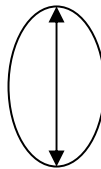
- ほとんど母系由来 家系をたどりやすい (※ 父系由来もある)
- 総塩基数 約16,500個
- ヒトとチンパンジ (約500万年前に分岐)  
ミトコンドリアDNAの遺伝子上で約9%のちがい  
(核DNAの遺伝子上では約1%)  
数万年スケールの分岐年代測定に威力
- 1987年 アラン・ウィルソン  
147人のミトコンドリアの系統樹  
14ー29万年前のアフリカ女性が祖先 (ミトコンドリア・イブ説)
- 1991年 アルプスで発見された約5000年前のミイラ  
(アイスマン)

## バイオインフォマティクスの問題③ クラスタリング

- 点の集合  $X$  における任意の2点  $x, y$  間には自然数の距離  $d(x, y)$  が定められているとする。

$X$  を  $k$  個のグループへ分割したとき  $(X_1, X_2, \dots, X_k)$ 、同じグループにある2点間の最大距離を、最小にする分割を計算したい。

$$\min \{ \max \{ d(x, y) \mid x, y \in X_i \} \mid i=1, 2, \dots, k \}$$

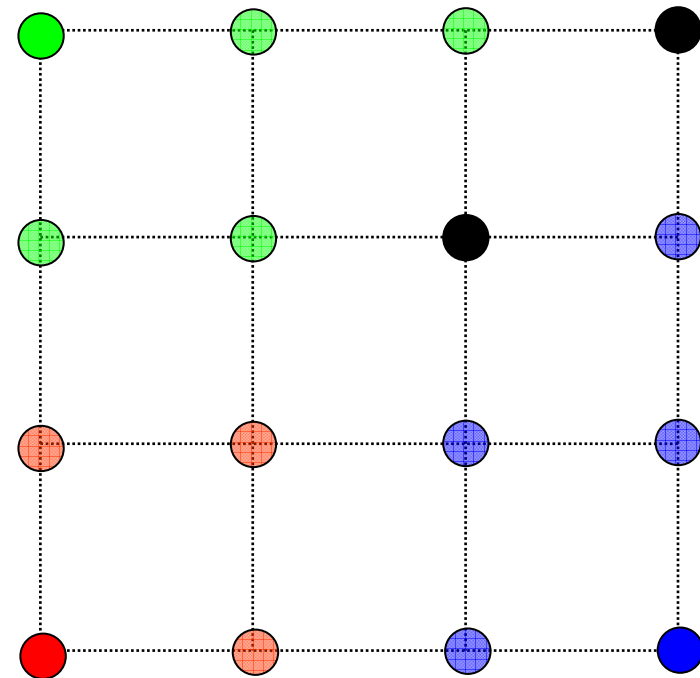
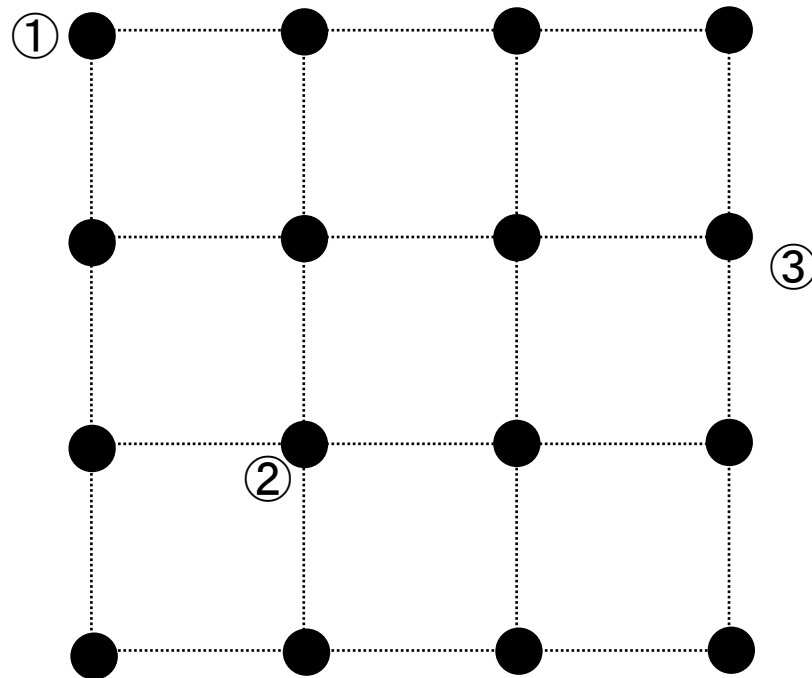


$k$  が3以上の場合 NP 困難問題

## バイオインフォマティクスの問題③ クラスタリング

3つのグループへ分けることを考える

マンハッタン距離 ①と②の距離は3, ①と③は4



最小化の例は？

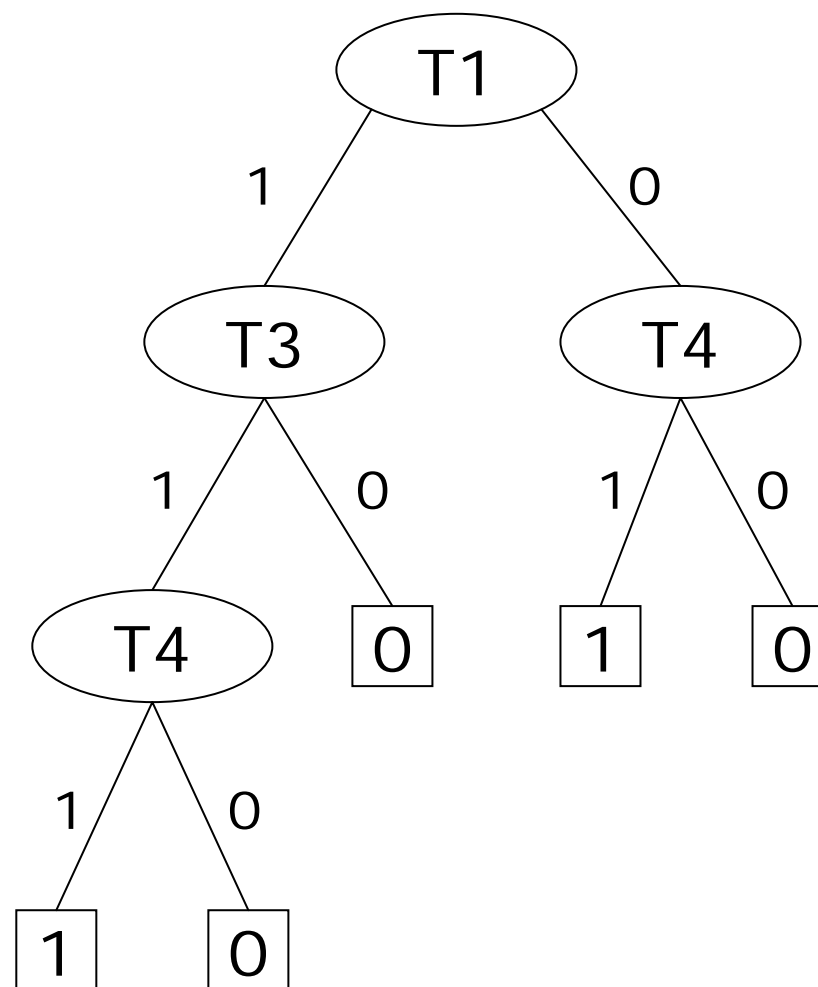
## バイオインフォマティクスの問題④ クラス分類

---

- 各データには目標属性が付随  
たとえばDNAを修復できない表現型をもつか否か？  
ALL か AML であるか？
- また目標属性を予測するための説明属性も付随  
たとえば、細胞が異常に大きいか否か？  
遺伝子が野生型より多く発現しているか否か？
- 説明属性の論理的な組合せで、目標属性を推定したい
- 組合せはコンパクトであってほしい

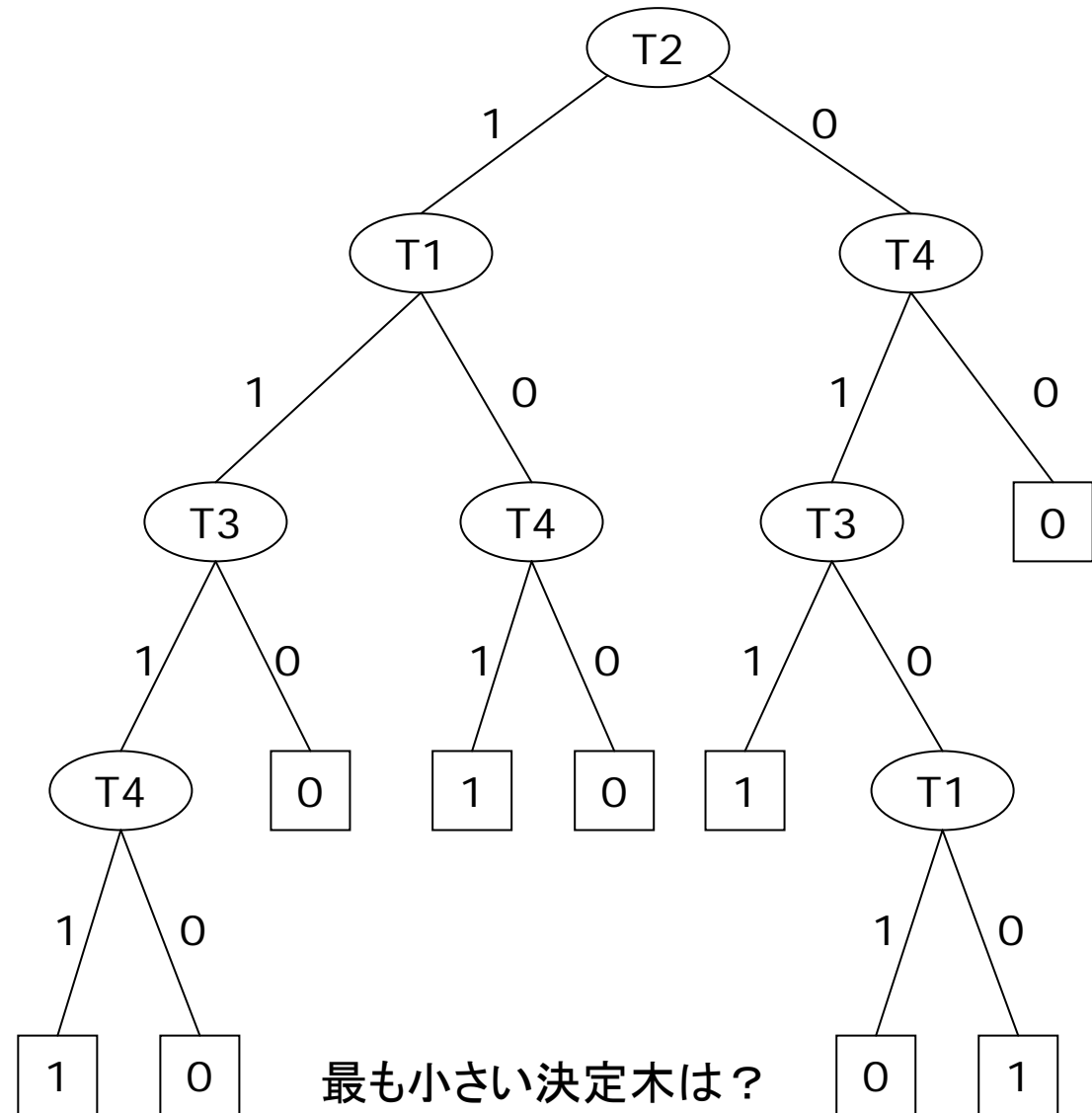
# バイオインフォマティクスの問題④ クラス分類

	T1	T2	T3	T4	目標属性
	1	0	1	1	1
	1	0	1	1	1
T4	1	1	1	1	1
	1	1	1	0	0
T3	1	0	1	0	0
	1	1	0	1	0
	1	0	0	1	0
T1	1	1	0	1	0
	0	1	0	1	1
	0	0	1	1	1
	0	1	0	1	1
	0	1	0	1	1
	0	0	0	1	1
T4	0	0	1	0	0
	0	1	0	0	0
	0	0	1	0	0



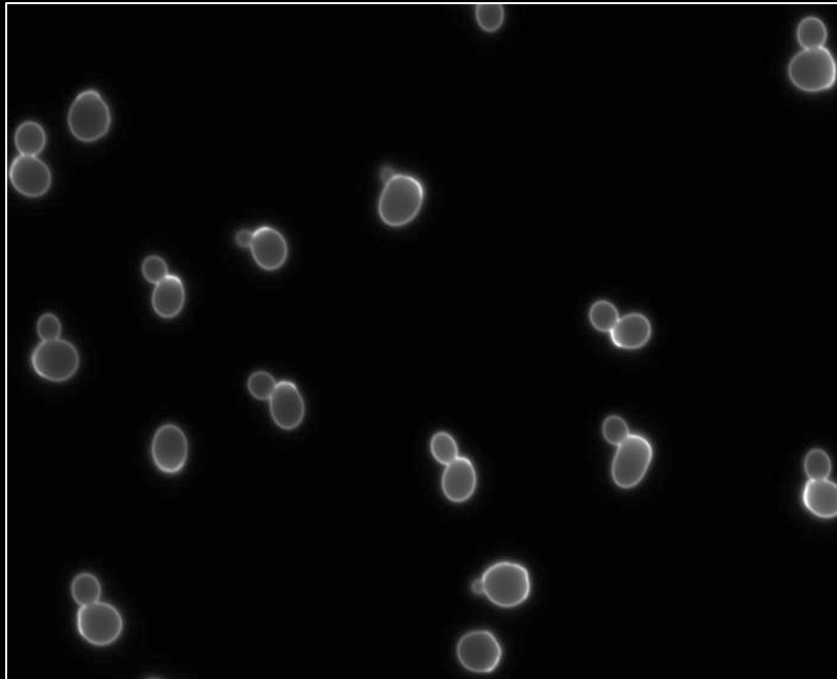
# バイオフィォマティクスの問題④ クラス分類

	T1	T2	T3	T4	目標属性
T4	1	1	1	1	1
T3	1	1	1	0	0
	1	1	0	1	0
T1	1	1	0	1	0
	0	1	0	1	1
	0	1	0	1	1
T4	0	1	0	1	1
	0	1	0	0	0
T2	1	0	1	1	1
	1	0	1	1	1
	0	0	1	1	1
T3	0	0	0	1	1
T1	1	0	0	1	0
T4	0	0	1	0	0
	1	0	1	0	0
	0	0	1	0	0

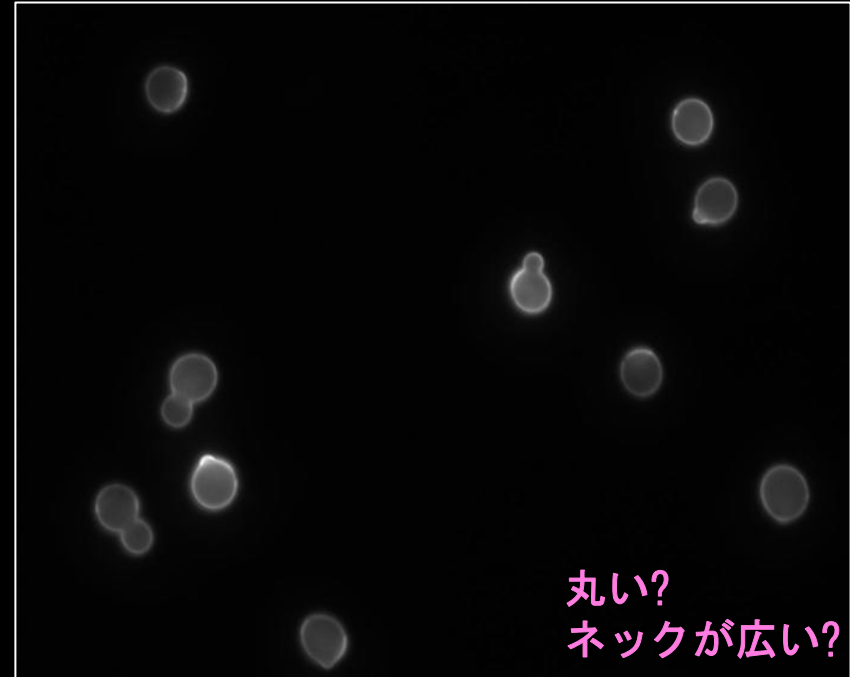


# 表現型からの遺伝子機能予測

野生株



*och1Δ*



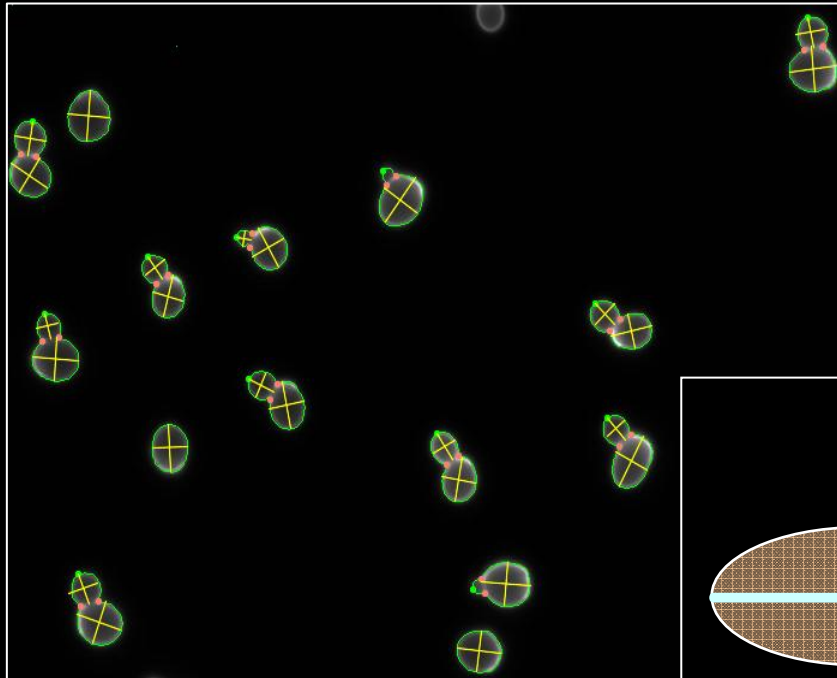
## 肉眼による比較

- ・ 主観的な記述になりがち
- ・ 変化の定量化が困難
- ・ 時間がかかる  
ハイスループット化は困難



# 表現型からの遺伝子機能予測

野生株

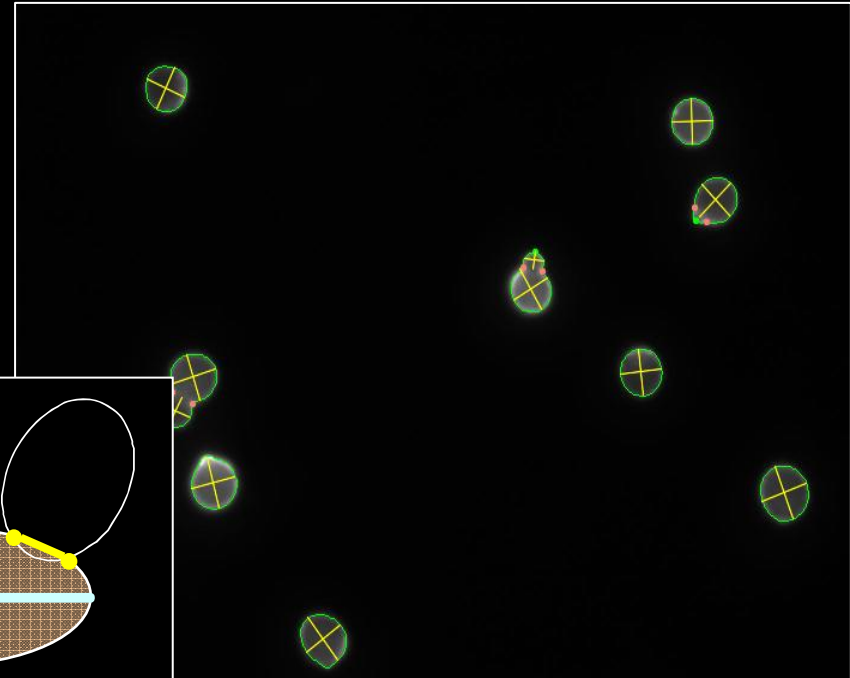


**1102.90 ± 301.34 pixels**

**1.21 ± 0.07 (n=397)**

**13.28 ± 1.70 pixels (n=246)**

*och1Δ*



**1332.03 ± 457 pixels**

**1.09 ± 0.05 (n=243)**

**17.417 ± 3.00 pixels (n=261)**

ソフトウェアによる解析

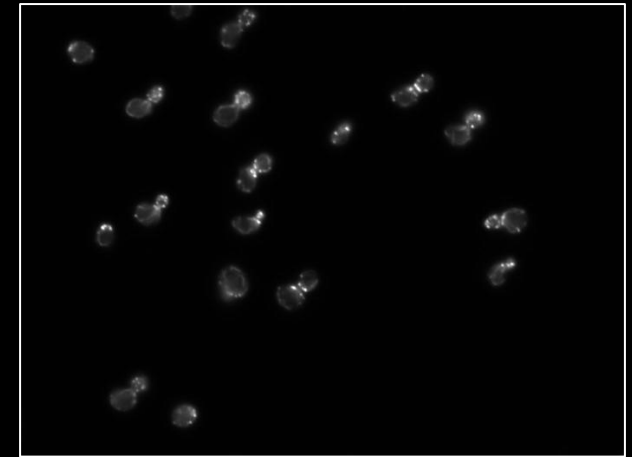
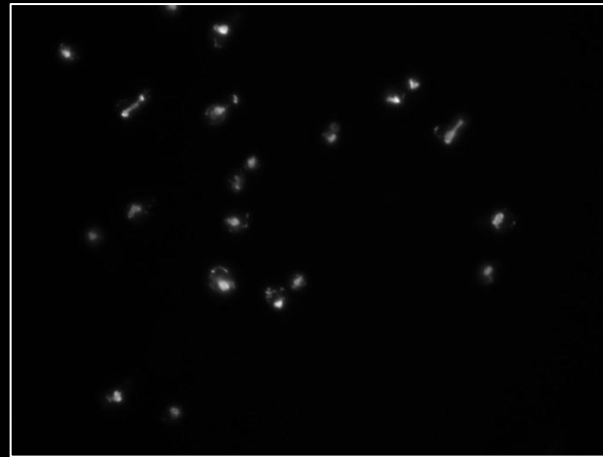
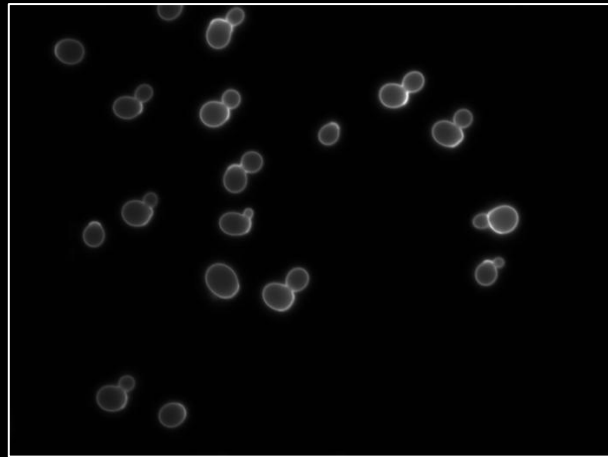


- 誤差の小さい定量化, 安定した再現性
- 大量処理可能 ⇒ 異常性の統計的検定

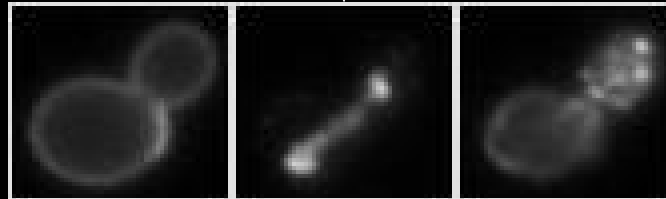
細胞壁 FITC-ConA

核 DAPI

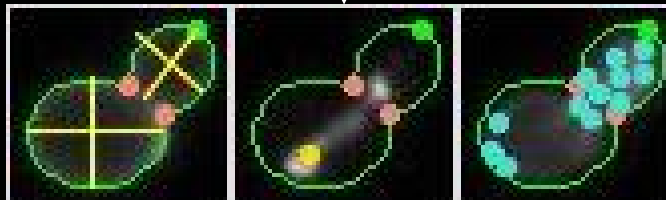
アクチン Rh-ph

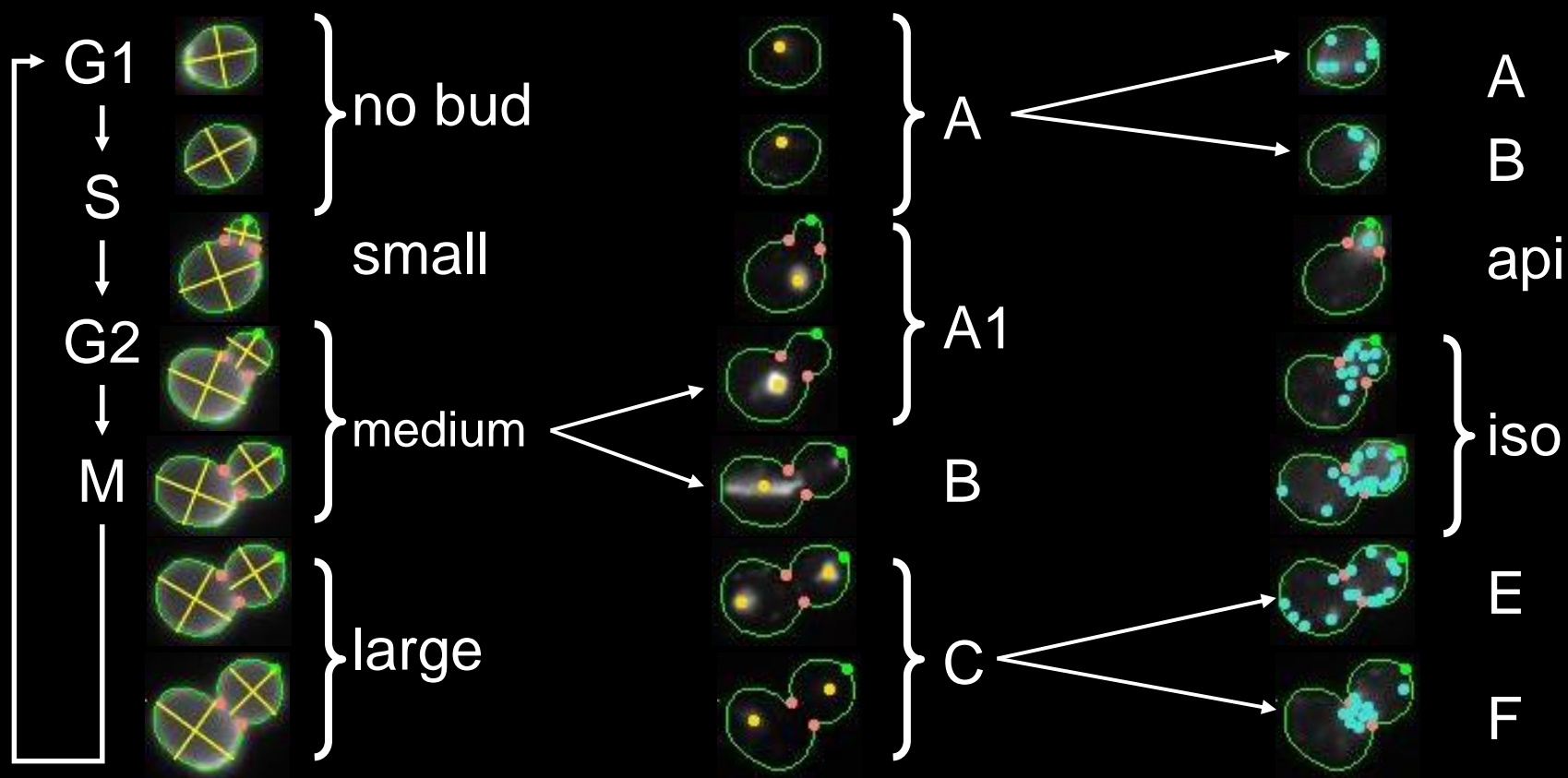
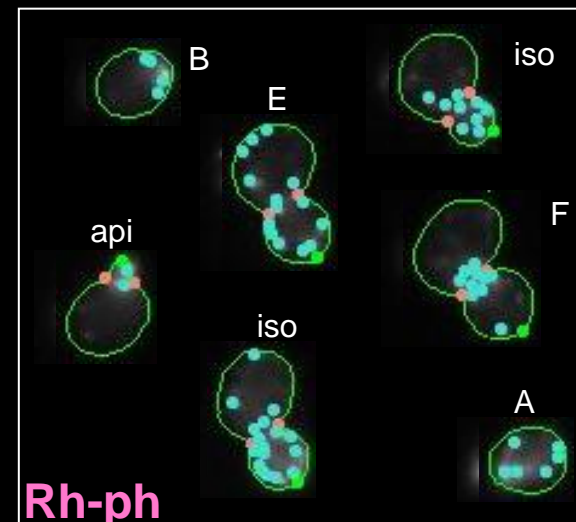
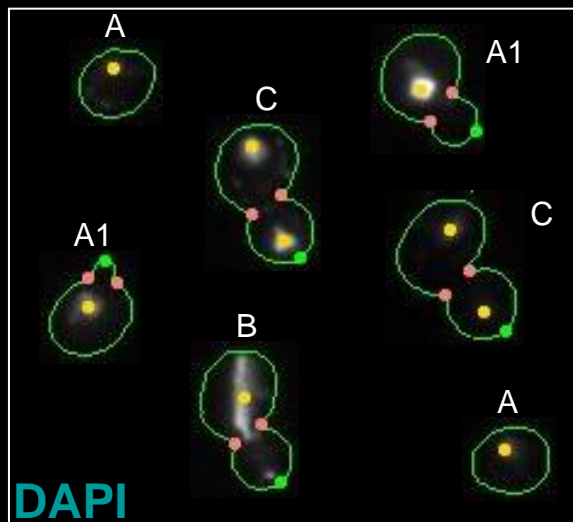
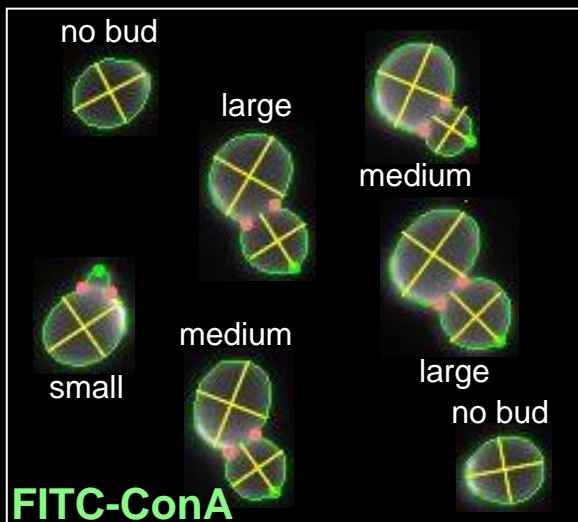


Superimposition



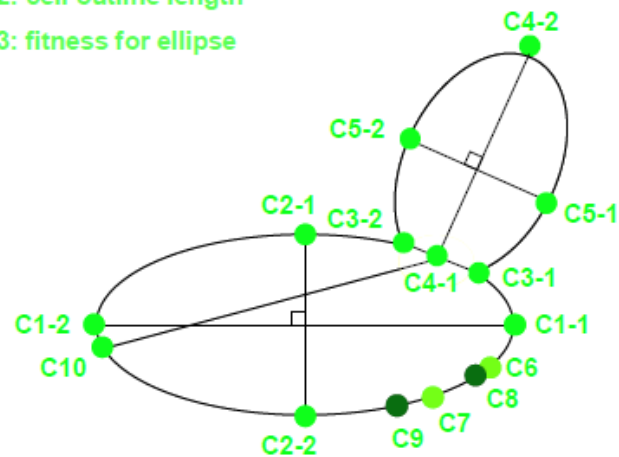
Cell-image Processing





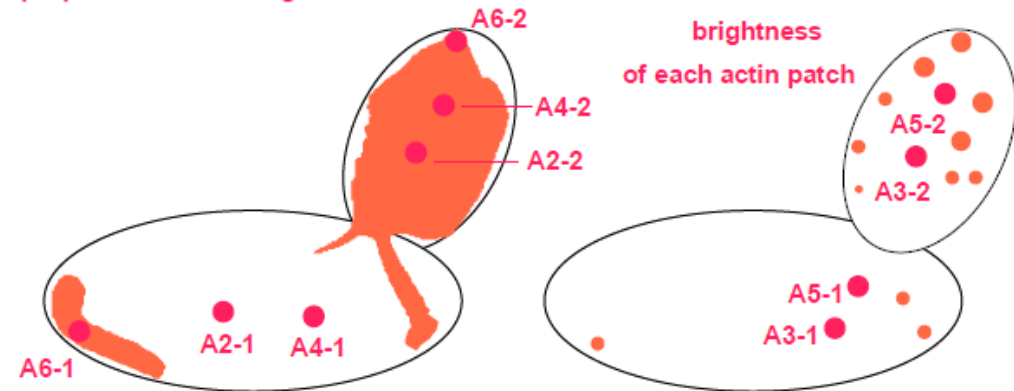
### From FITC-ConA image

C11: cell size  
C12: cell outline length  
C13: fitness for ellipse

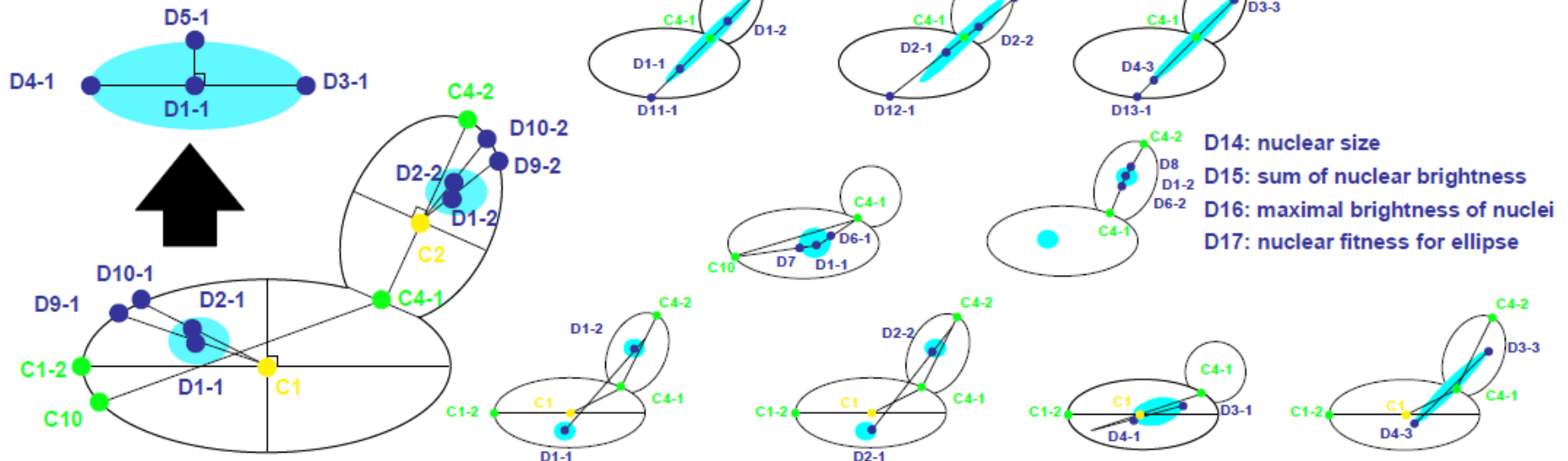


### From Rh-ph image

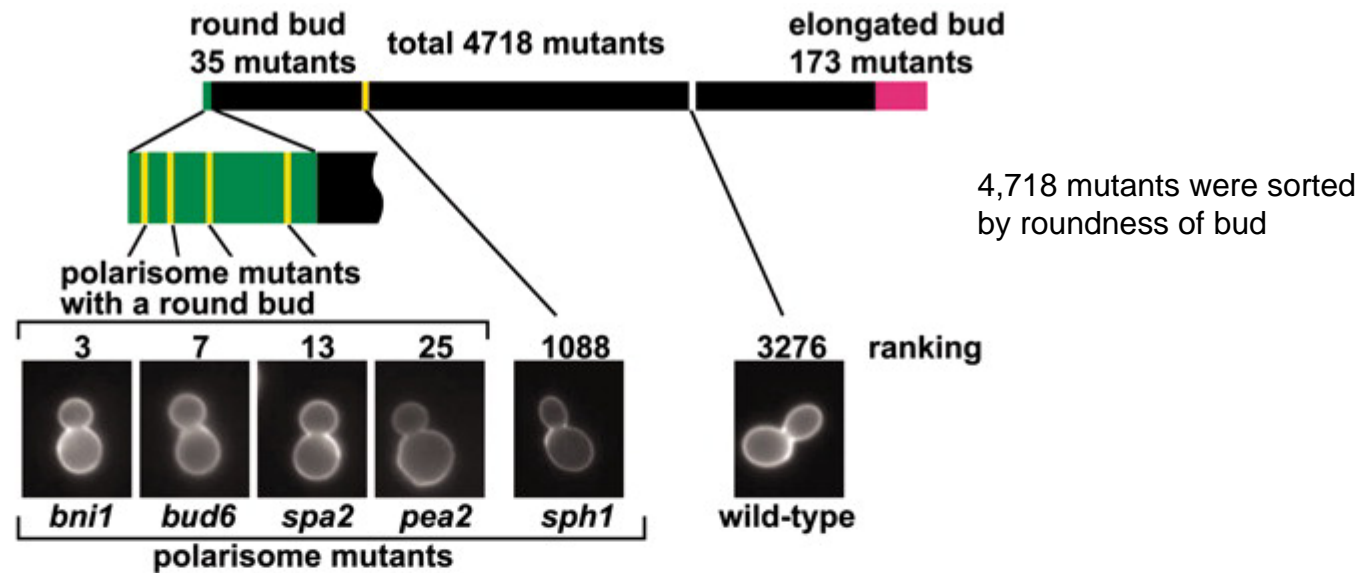
A7: size of actin region  
A8: total brightness of actin region  
A9: proportion of actin region at neck



### From DAPI image



A



B

parameter	GO term	gene in GO	distribution of mutants
mother axis ratio of pre-mitotic cell	chromatin assembly complex	<i>HTZ1</i> , <i>CAC2</i> , <i>ASF1</i> , <i>MSI1</i> , <i>RLF2</i>	
variation of long axis length of pre-mitotic cell	THO complex	<i>MFT1</i> , <i>HPR1</i> , <i>THP2</i> , <i>RLR1</i>	
variation of whole cell size of pre-mitotic cell	double-strand break repair via single-strand annealing	<i>RAD10</i> , <i>MRE11</i> , <i>SAE2</i> , <i>RAD59</i> , <i>RAD55</i> , <i>RAD57</i> , <i>RAD1</i> , <i>RAD50</i> , <i>RAD52</i> , <i>RAD51</i> , <i>RAD54</i>	
axis ratio of no bud cell	SWR1 complex	<i>SWC7</i> , <i>SWR1</i> , <i>VPS72</i> , <i>SWC5</i> , <i>ARP6</i> , <i>VPS71</i> , <i>SWC3</i> , <i>YAF9</i>	
direction and distance of bud tip of pre-mitotic cell	mitotic sister chromatid cohesion	<i>CHL1</i> , <i>CTF18</i> , <i>CTF8</i> , <i>DCC1</i> , <i>CTF4</i> , <i>TRF4</i>	
nucleus position of unbudded cell	RNA polymerase complex	<i>RPA34</i> , <i>RPA14</i> , <i>RPA49</i> , <i>RPA12</i> , <i>RPB9</i> , <i>RPB4</i>	
axis ratio of no bud cell	Pol II transcription elongation factor activity	<i>RTF1</i> , <i>CHD1</i> , <i>CDC73</i> , <i>SPT4</i> , <i>ELP2</i> , <i>ELP4</i> , <i>ELP6</i> , <i>LEO1</i> , <i>IKI3</i> , <i>ELP3</i> , <i>PAF1</i>	

# 手におえないほど計算時間がかかる問題への対処

---

- 最適解に近い近似解を高速にもとめる（近似アルゴリズム）
- 高い確率で最適解を高速にもとめる（確率的アルゴリズム）
- 手におえないほど計算時間がかかる問題として定式化しない（わざわざ問題を難しくしない）

現実の時間で解ける問題で妥協する

- しかし、問題を定式化したとき、手におえない問題であるか否かを判断できる理論的素養をもつことは重要でしょう

# 答え

---

$$50 + 62 + 89 + 123 + 176 = 500$$

AAATTTATAA

AAA TTA

AAT TAT

ATT ATA

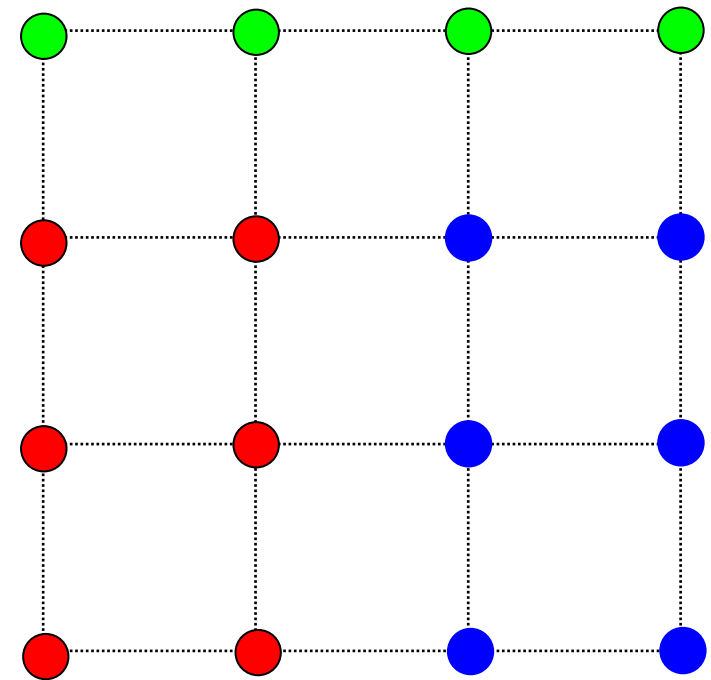
TTT TAA

ATATATA

TATTAAT

ATAATAT

ATTATAA



答え

	T1	T2	T3	T4	目標属性
	1	1	1	1	1
	1	0	1	1	1
	1	0	1	1	1
T3	0	0	1	1	1
	0	1	0	1	1
	0	1	0	1	1
	0	1	0	1	1
T1	0	0	0	1	1
	1	1	0	1	0
	1	1	0	1	0
	1	0	0	1	0
T4	0	0	1	0	0
	1	0	1	0	0
	0	0	1	0	0
	1	1	1	0	0
	0	1	0	0	0

