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

- 4/8, 15, 22, 5/20, 27, 6/3　　森下
- 4/29, 5/6 は休日　5/13 は休講
- 6/10, 17, 24, 7/1, 8, 15　　中谷
- 7/22　　筆記試験

森下分の講義

- 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, Naïve Bayes, Support Vector Machine
- 大規模ゲノムアセンブリ

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　ー　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

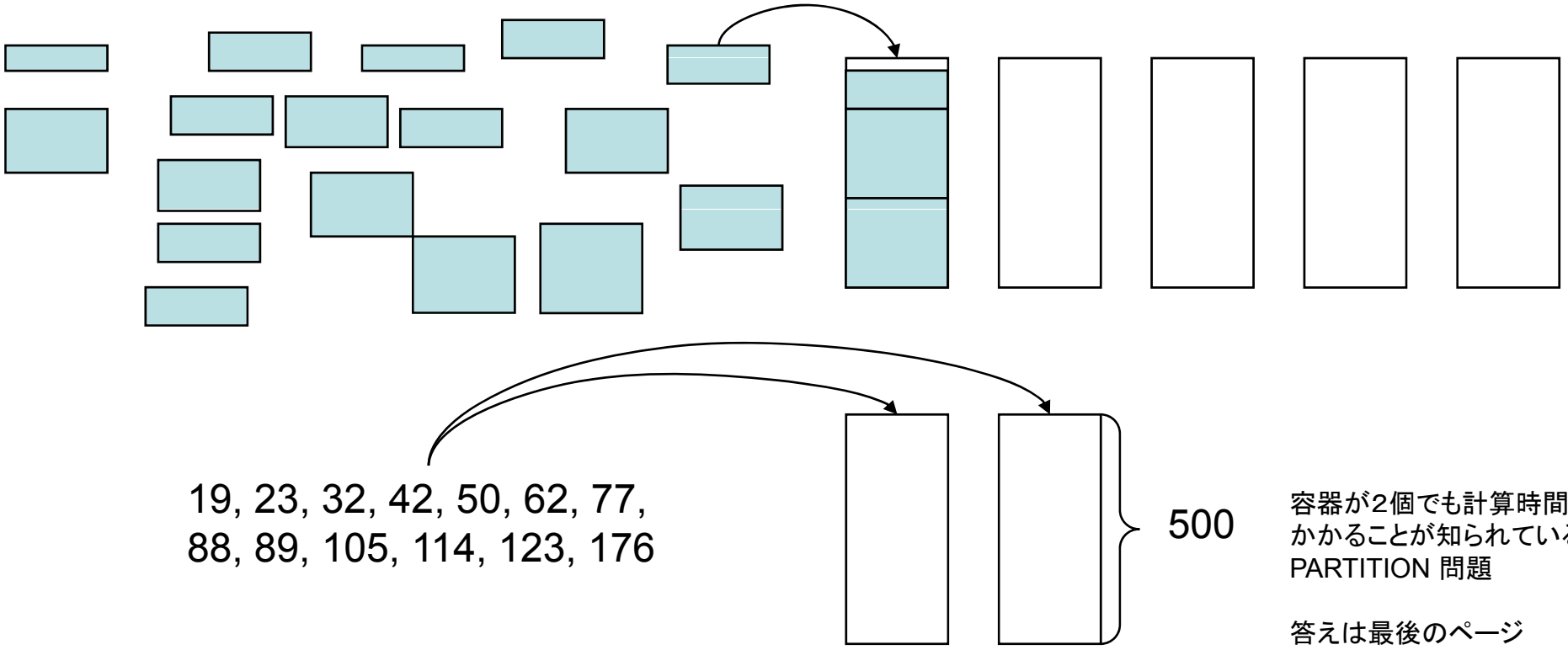
  1972年　Cook によるクラスの提示
  　　Karp による様々な例の提示

- 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 size(u) である積木 u の集合 U を k 個の部分集合 U_i (i=1,…,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？　　$\Sigma\{\text{size}(u) \mid u \in U_i\} \leqq B$



容器が2個でも計算時間がかかることが知られている PARTITION 問題

答えは最後のページ

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Σ、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？

- Σ＝{**A**, **T**}
  R = {**AAA**, **AAT**, **ATA**, **ATT**, **TAA**, **TAT**, **TTA**, **TTT**}

AAAAATATAATTTAATATTTATTT
AAAAATATAATTTAATATTTATTT

AAATAATTATTT
AAA　ATT
AAT　TTA
ATA　TAT
TAA　TTT

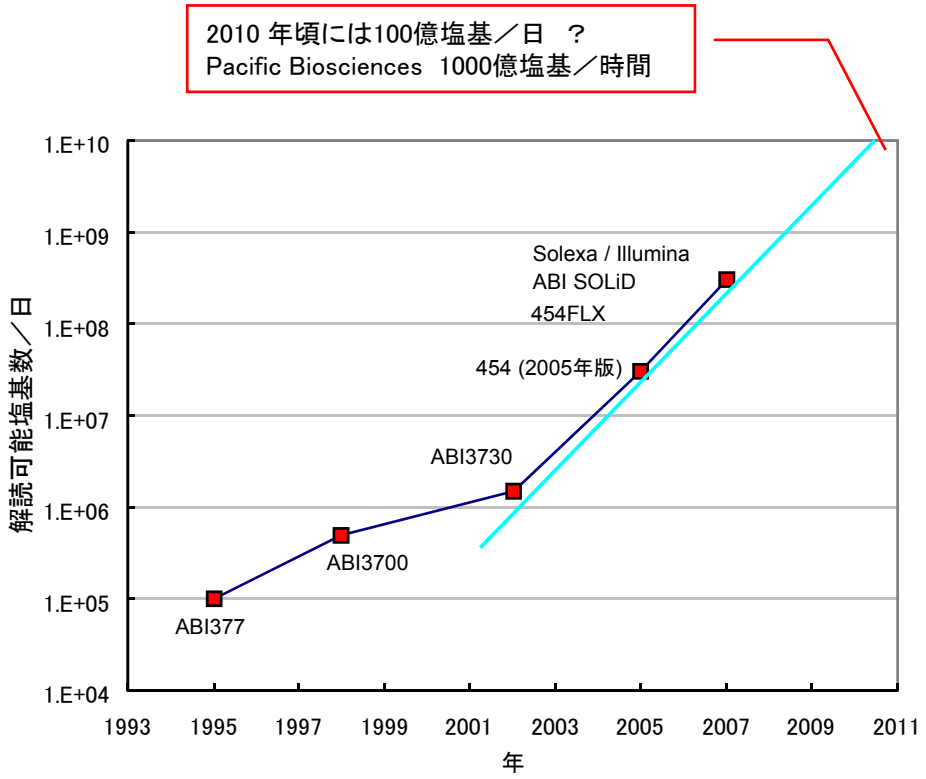
- NP困難問題：　*R* の元の数を *n* とすれば 2^n に比例する計算時間がかかるアルゴリズムしか知られてない

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！

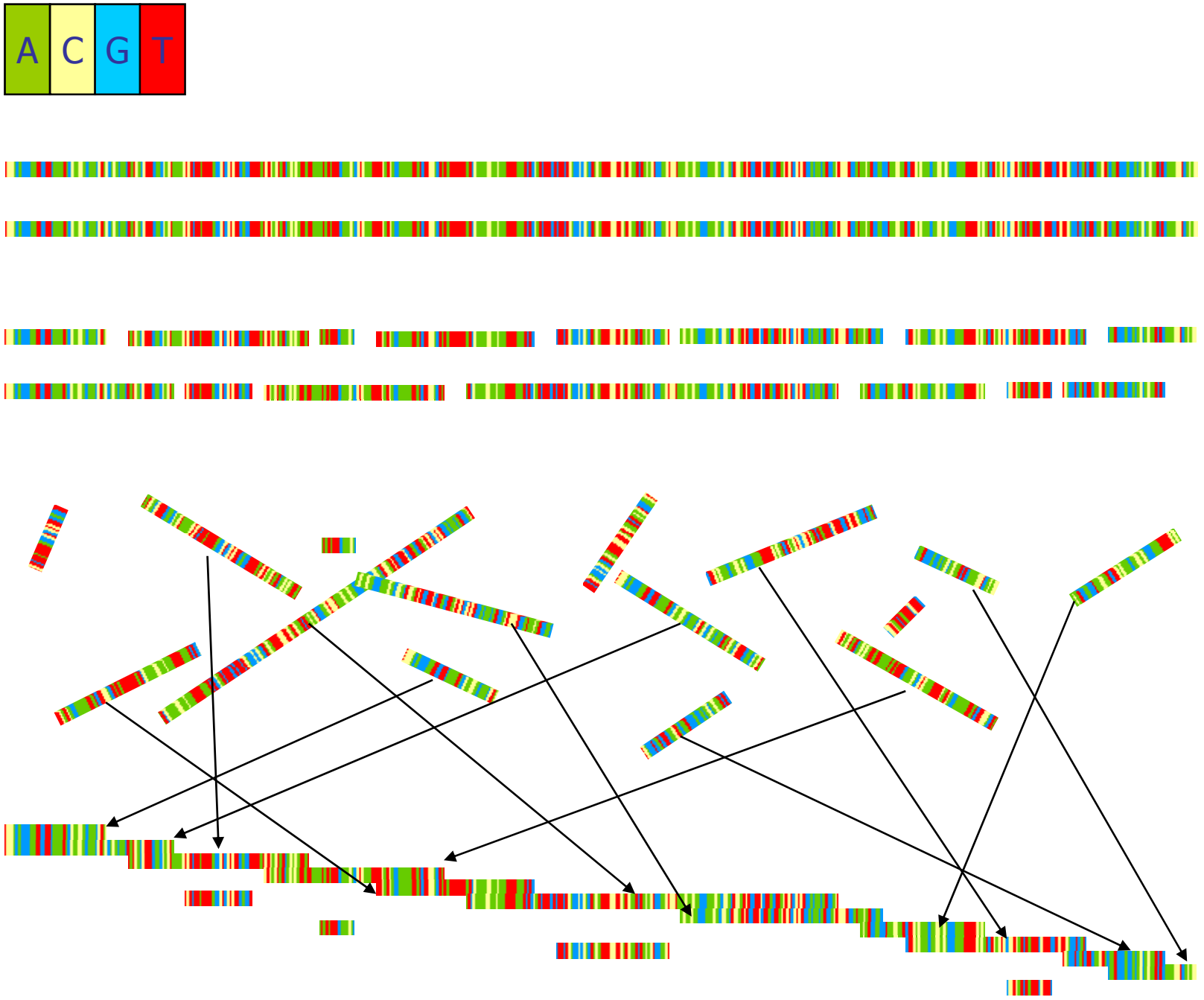
ゲノム解読の高速化

- ヒトゲノムプロジェクト（1991–2006）$2.7 billion, 16年
- 2004年　哺乳類ゲノム（30億塩基）の解読 $10–50 million
- 2004年2月　NIHファンド“$1000 genome project”
- 2005年から驚異的な高速化



シークエンサー	普及年	配列長(塩基数)	解析可能試料数*	総塩基数	30億塩基の収集
ABI SOLiD	2007	25 – 35	100,000,000/実験	– 30 億/8日	約 8日
SOLEXA	2007	25 – 36	60,000,000/実験	– 20 億/4日	約 6日
454FLX	2007	– 250	400,000/実験	– 2 億/日	約 15日
ABI 3730xl	2002	– 800	2304/ 日	– 0.02 億/日	約1,500日

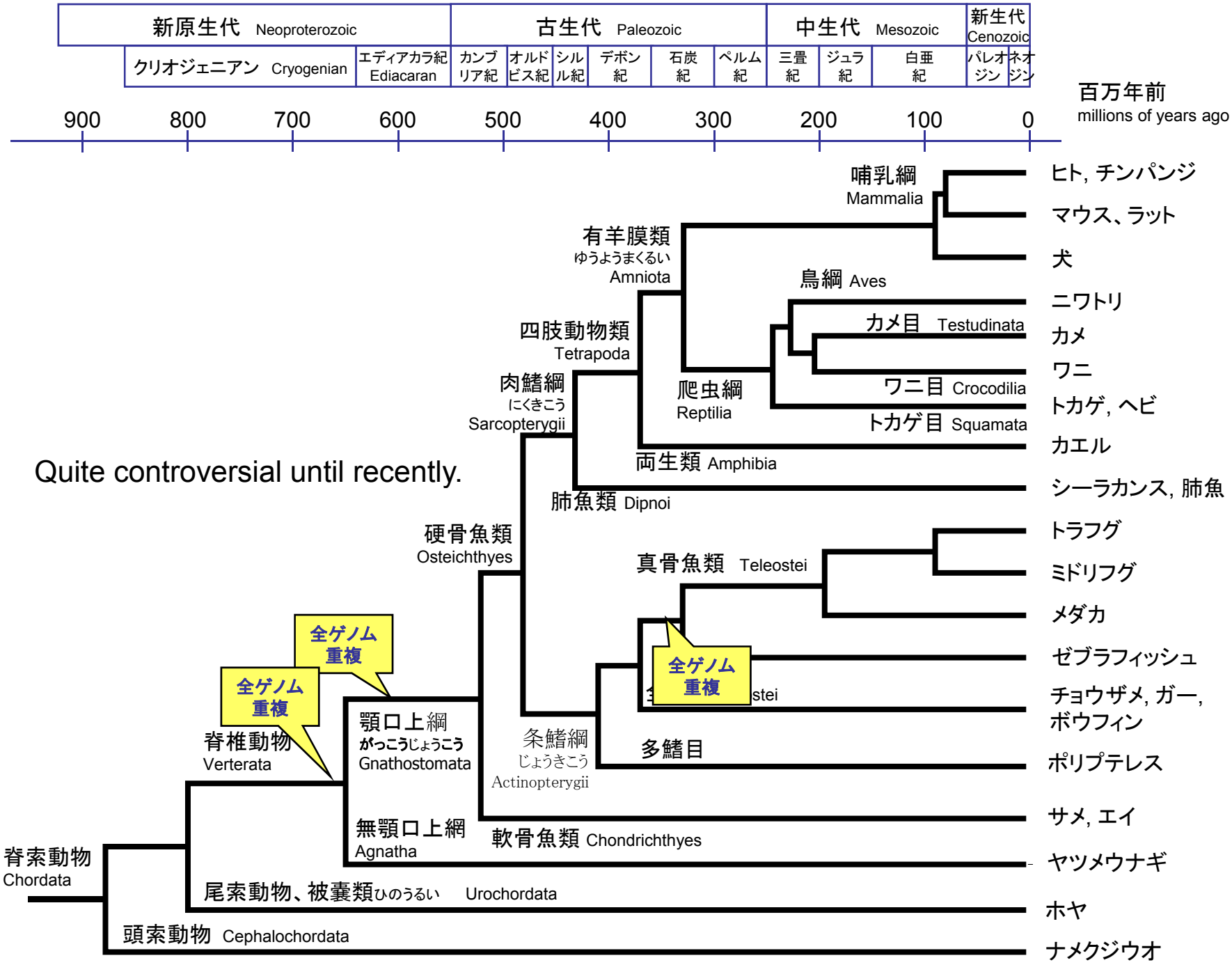
*SOLEXAは1回の実験に4日（コスト約100万円）*SOLiDは1回の実験に8日 *454FLXは1回の実験に7時間



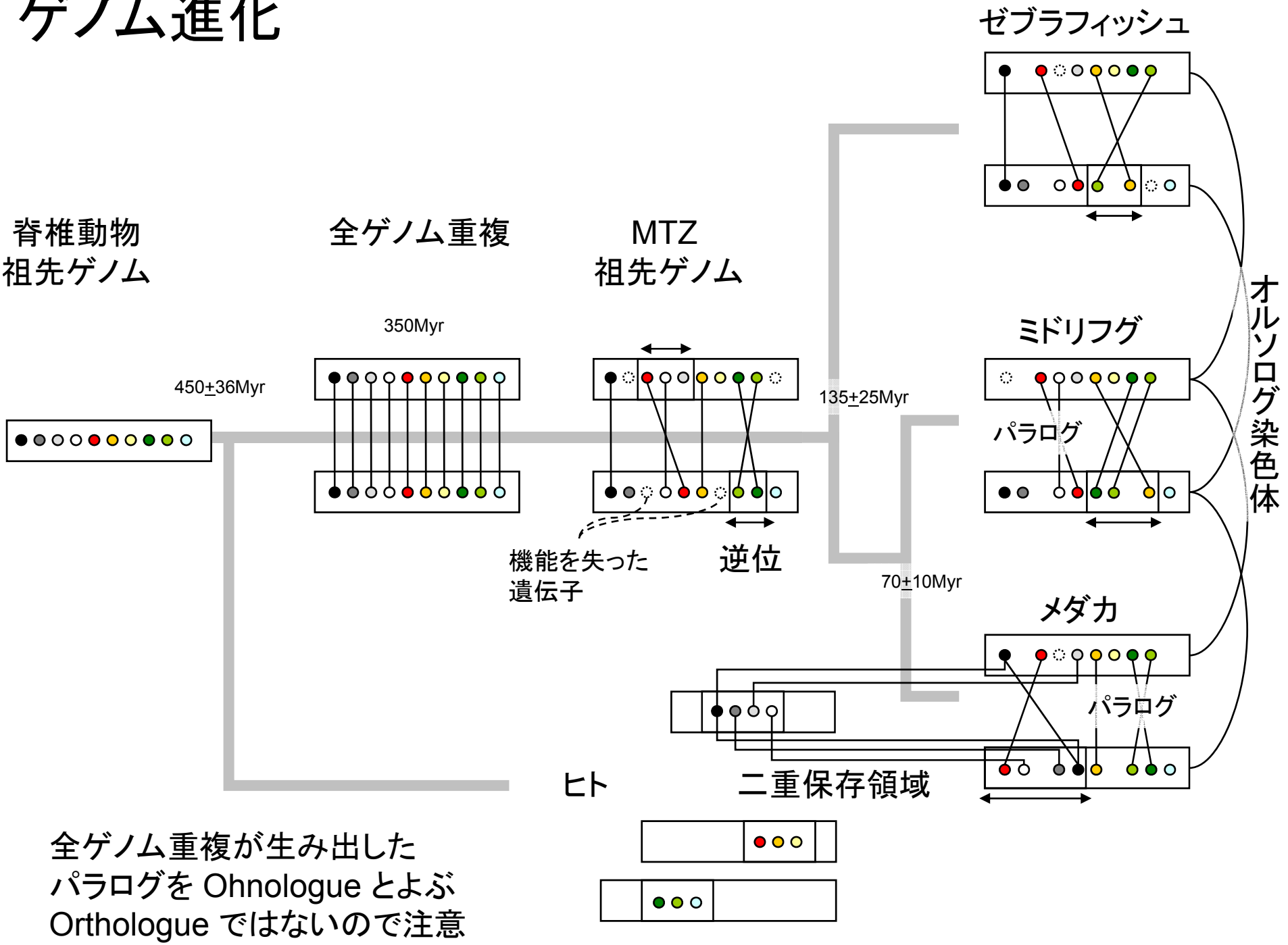
元の絵がわからないジグソーパズル（数百万～数千万ピース）

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 / 6	メダカ	7億	Ramen (東大) 国立遺伝学研究所	日本
2008 / ?	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

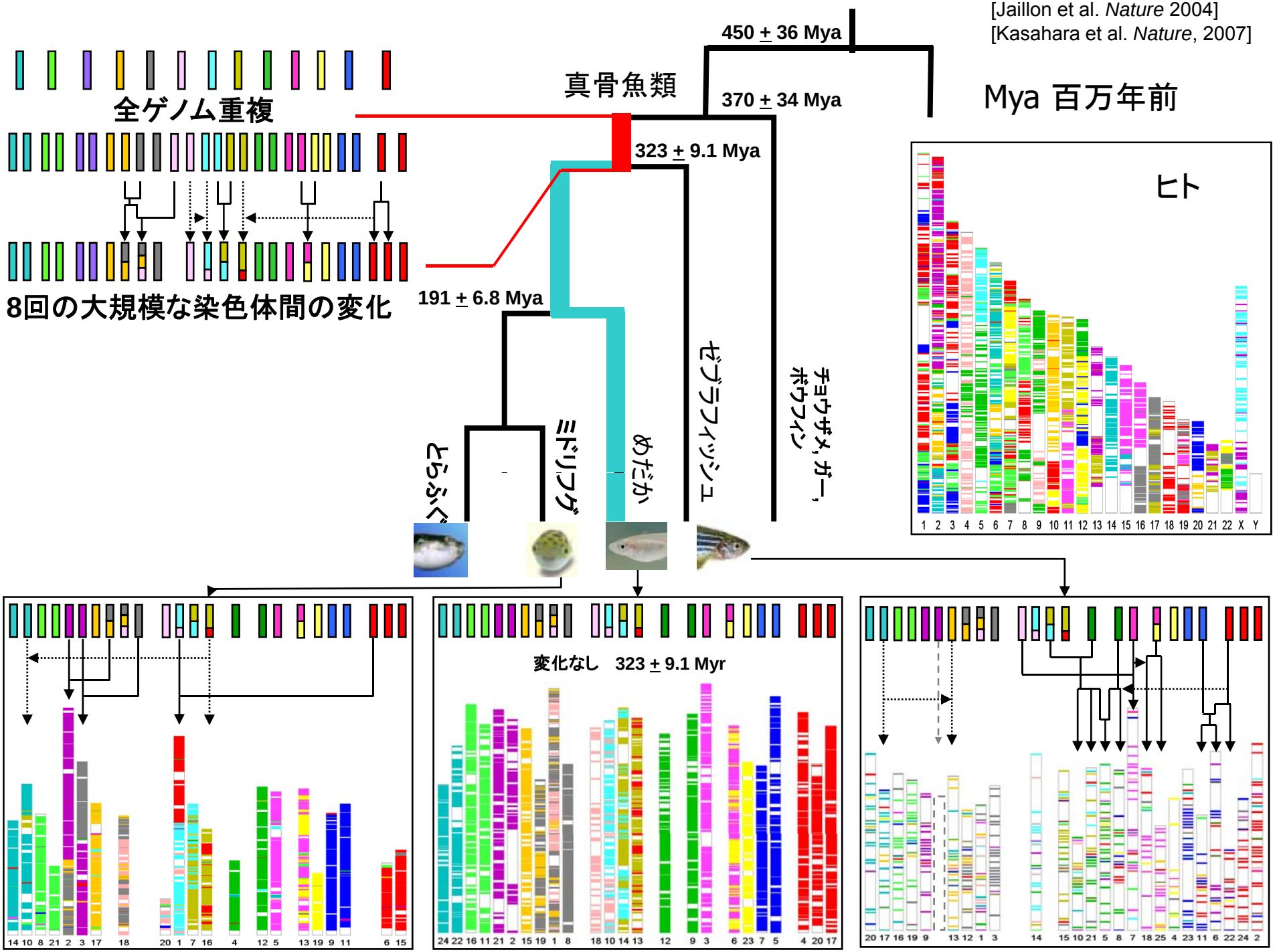

新原生代 Neoproterozoic
古生代 Paleozoic
中生代 Mesozoic
新生代 Cenozoic
クリオジェニアン Cryogenian
エディアカラ紀 Ediacaran
カンブリア紀
オルドビス紀
シルル紀
デボン紀
石炭紀
ペルム紀
三畳紀
ジュラ紀
白亜紀
パレオジン
ネオジン
900
800
700
600
500
400
300
200
100
0
百万年前
millions of years ago
哺乳綱 Mammalia
ヒト, チンパンジ
マウス、ラット
犬
有羊膜類 ゆうようまくるい Amniota
鳥綱 Aves
ニワトリ
カメ目 Testudinata
カメ
ワニ
四肢動物類 Tetrapoda
肉鰭綱 にくきこう Sarcopterygii
爬虫綱 Reptilia
ワニ目 Crocodilia
トカゲ, ヘビ
トカゲ目 Squamata
カエル
両生類 Amphibia
シーラカンス, 肺魚
肺魚類 Dipnoi
Quite controversial until recently.
硬骨魚類 Osteichthyes
トラフグ
真骨魚類 Teleostei
ミドリフグ
メダカ
全ゲノム重複
ゼブラフィッシュ
全ゲノム重複
チョウザメ, ガー, ボウフィン
全ゲノム重複
顎口上綱 がっこうじょうこう Gnathostomata
脊椎動物 Verterata
条鰭綱 じょうきこう Actinopterygii
多鰭目
ポリプテレス
サメ, エイ
無顎口上網 Agnatha
軟骨魚類 Chondrichthyes
ヤツメウナギ
脊索動物 Chordata
尾索動物、被嚢類 ひのうるい Urochordata
ホヤ
頭索動物 Cephalochordata
ナメクジウオ

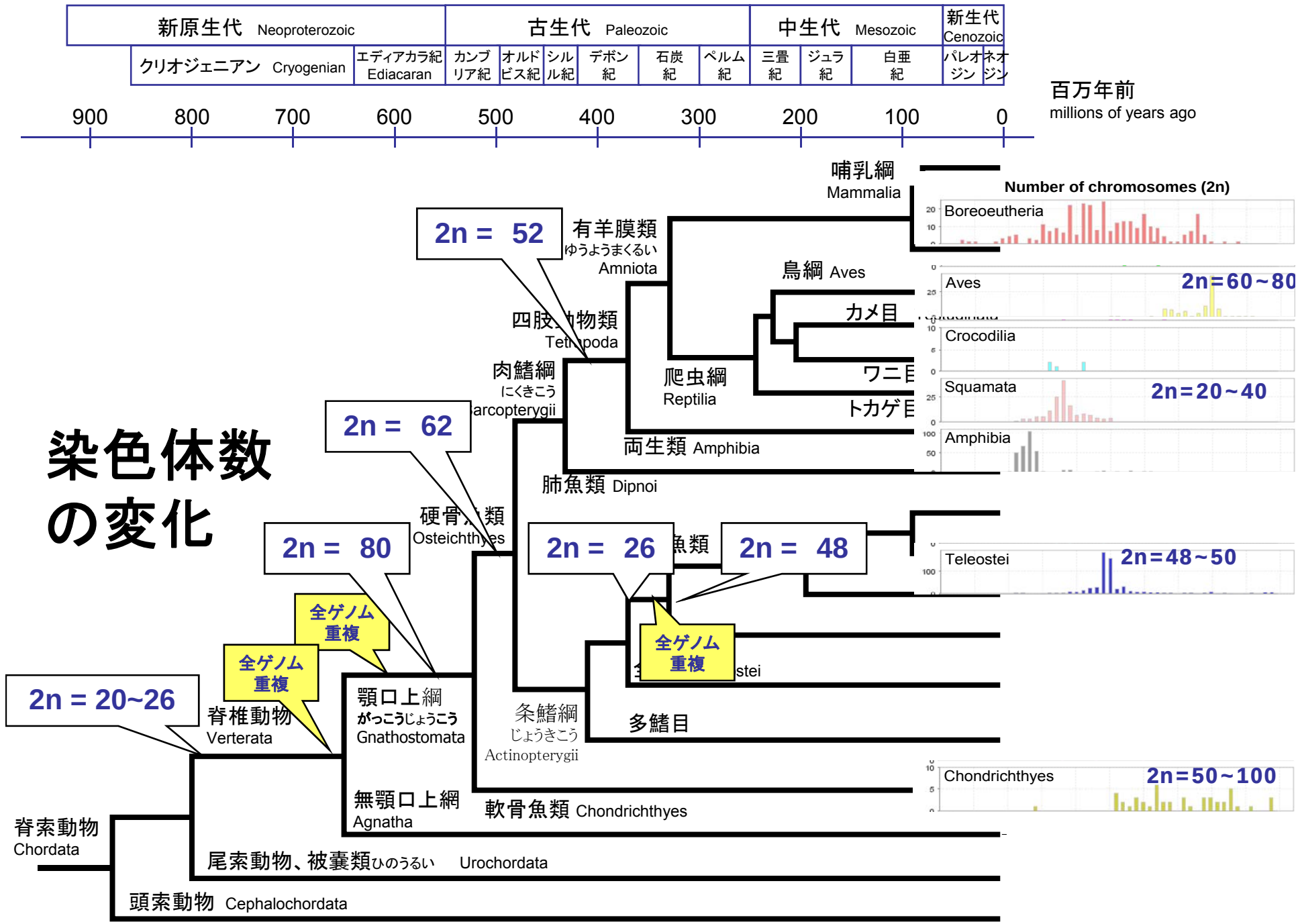
ゲノム進化



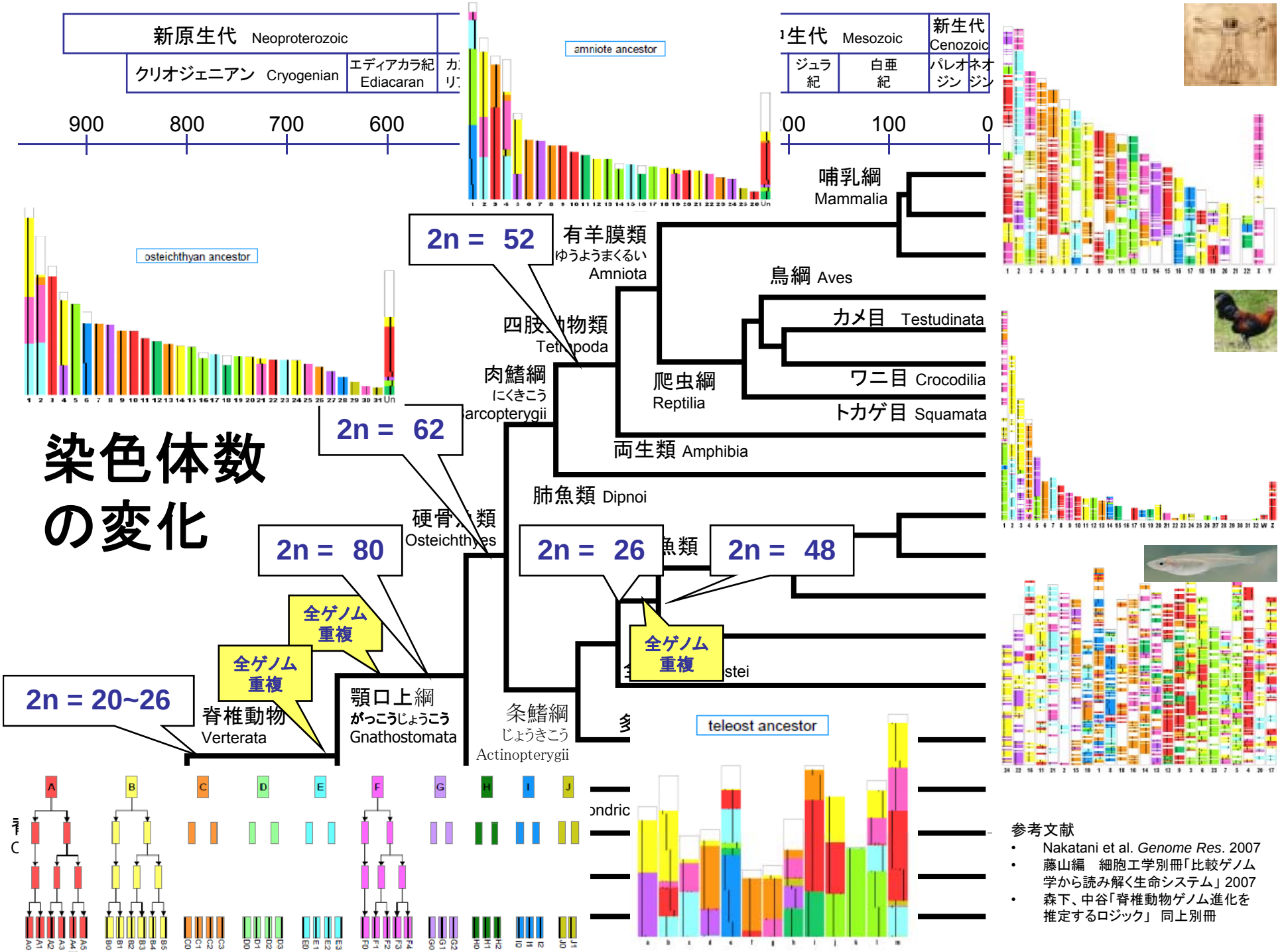
全ゲノム重複が生み出した
パラログを Ohnologue とよぶ
Orthologue ではないので注意

[Jaillon et al. *Nature* 2004]
[Kasahara et al. *Nature*, 2007]




新原生代 Neoproterozoic
古生代 Paleozoic
中生代 Mesozoic
新生代 Cenozoic
クリオジェニアン Cryogenian
エディアカラ紀 Ediacaran
カンブリア紀
オルドビス紀
シルル紀
デボン紀
石炭紀
ペルム紀
三畳紀
ジュラ紀
白亜紀
パレオジン
ネオジン
900 800 700 600 500 400 300 200 100 0
百万年前
millions of years ago
染色体数の変化
哺乳綱 Mammalia
Number of chromosomes (2n)
Boreoeutheria
2n = 52
有羊膜類 ゆうようまくるい Amniota
鳥綱 Aves
Aves
2n=60~80
カメ目
四肢動物類 Tetrapoda
Crocodilia
肉鰭綱 にくきこう Sarcopterygii
爬虫綱 Reptilia
ワニ目
Squamata
2n=20~40
トカゲ目
2n = 62
両生類 Amphibia
Amphibia
肺魚類 Dipnoi
硬骨魚類 Osteichthyes
2n = 80
2n = 26
2n = 48
Teleostei
2n=48~50
全ゲノム重複
全ゲノム重複
全ゲノム重複
2n = 20~26
脊椎動物 Verterata
顎口上綱 がっこうじょうこう Gnathostomata
条鰭綱 じょうきこう Actinopterygii
多鰭目
無顎口上綱 Agnatha
軟骨魚類 Chondrichthyes
Chondrichthyes
2n=50~100
脊索動物 Chordata
尾索動物、被嚢類 ひのうるい Urochordata
頭索動物 Cephalochordata

染色体数の変化



参考文献

- Nakatani et al. *Genome Res.* 2007
- 藤山編　細胞工学別冊「比較ゲノム学から読み解く生命システム」2007
- 森下、中谷「脊椎動物ゲノム進化を推定するロジック」 同上別冊

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Σ、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-
-ATTA-TAA-

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

ATATATA
TATTAAT

homeobox A1 protein **Refseq:** NM_005522 **Protein:** P49639 (aka HXA1_HUMAN)

Human	GACAATGCAAGA	ATG	AAC	TCC	TTC	CTG	GAA	TACCC	- - - CATA
Chimp	GACAATGCAAGA	ATG	AAC	TCC	TTC	CTG	GAA	TACCC	- - - CATA
Mouse	GACAATGCAAGA	ATG	AAC	TCC	TT**T**	CTG	GAA	TACCC	- - - CAT**C**
Rat	GACAATGCAAGA	ATG	AAC	TCC	TT**T**	CTG	GAA	TACCC	- - - CAT**C**
Dog	GACAATGCAAGA	ATG	A**G**	CTC	CTT	CTG	GAA	TACCC	- - - CAT**C**
Chicken	GACAAT**A**CT**T**AG**G**	ATG	AAC	TCC	TTC	**T**T**A**	GA**G**	TA**T**G	C - - - **A**AT**T**
Fugu	**-**ACAATGC**C**A**C**	A	TG	A**G**	C**AG**	CTTC	**T**T**A**	GA**T**	TAC**T**C - - - **T**G**T**G
Zebrafish	GA**AG**ATG**AC**A**C**	A	TG	A**G**	C**A**	C**A**	TTC	**T**T**A**	GA**T**T**TT**TC**GTC**CATA

アミノ酸	Met	Asn	Ser	Phe	Leu	Glu	Tyr	Pro	Ile
同義置換	ATG	AAC	TCC	TTC	CTG	GAA	TAC	CCC	ATA
			AGC	TTT	T**T**A	GA**G**	TA**T**		AT**C**
									AT**T**
		Ser	Thr			Asp	Phe		Val
非同義置換		A**G**C	A**C**A			GA**T**	T**TT**		**G**T**G**

```
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α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

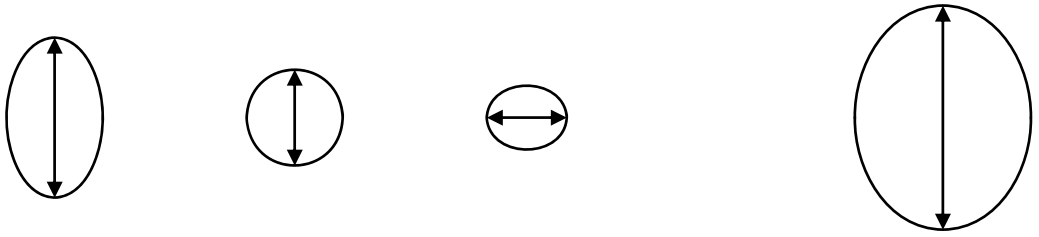
	ゴリラ	ウマ	イヌ	イモリ	コイ
ヒト	1	18	23	62	68

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 $(X_1, X_2, \ldots, X_k)$、
  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。

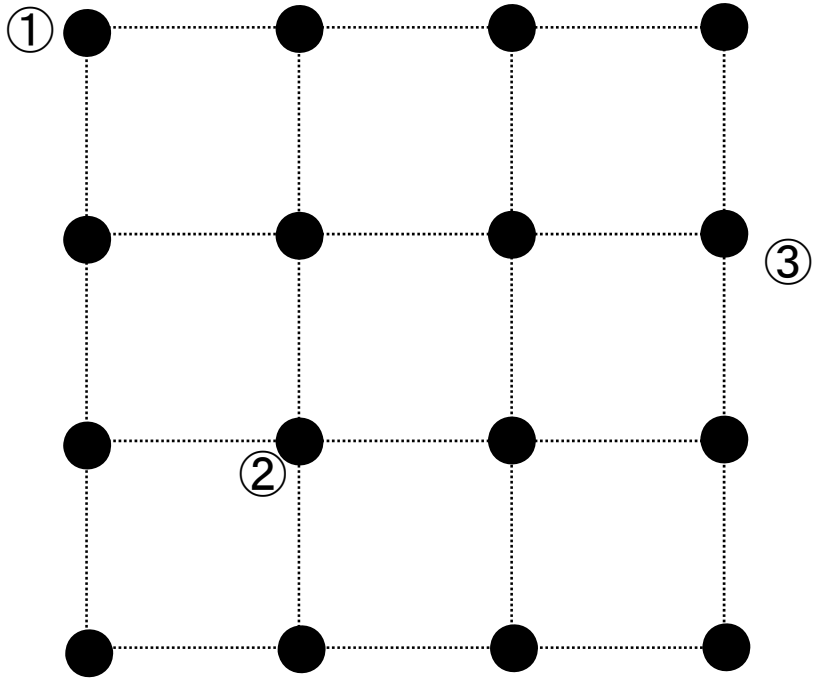
$$\min\{\ \max\{d(x,y) \mid x, y \in X_i\} \mid\ i=1,2,\ldots k\}$$

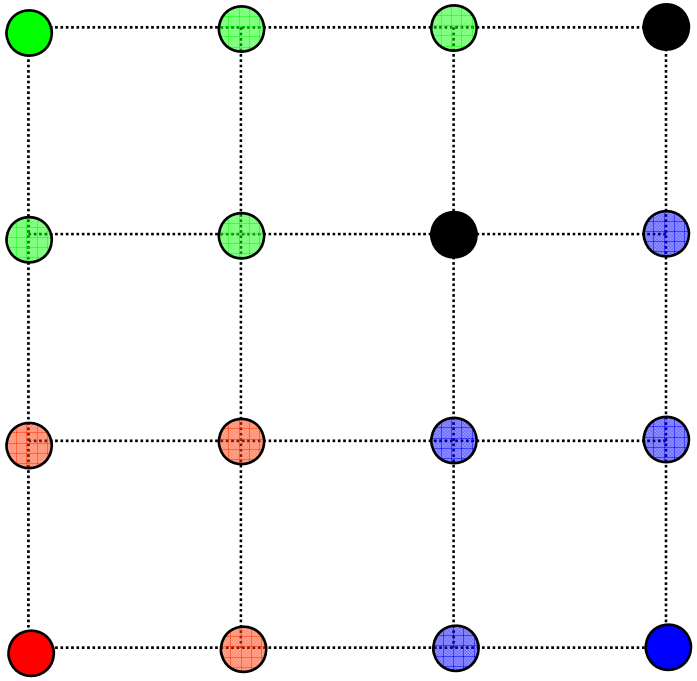


k が3以上の場合 NP 困難問題

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？

転写開始点周辺でのクロマチン構造

F. Ozsolak, J. S. Song, X. S. Liu, D. E. Fisher, *Nat Biotechnol* **25**, 244 (Feb, 2007).

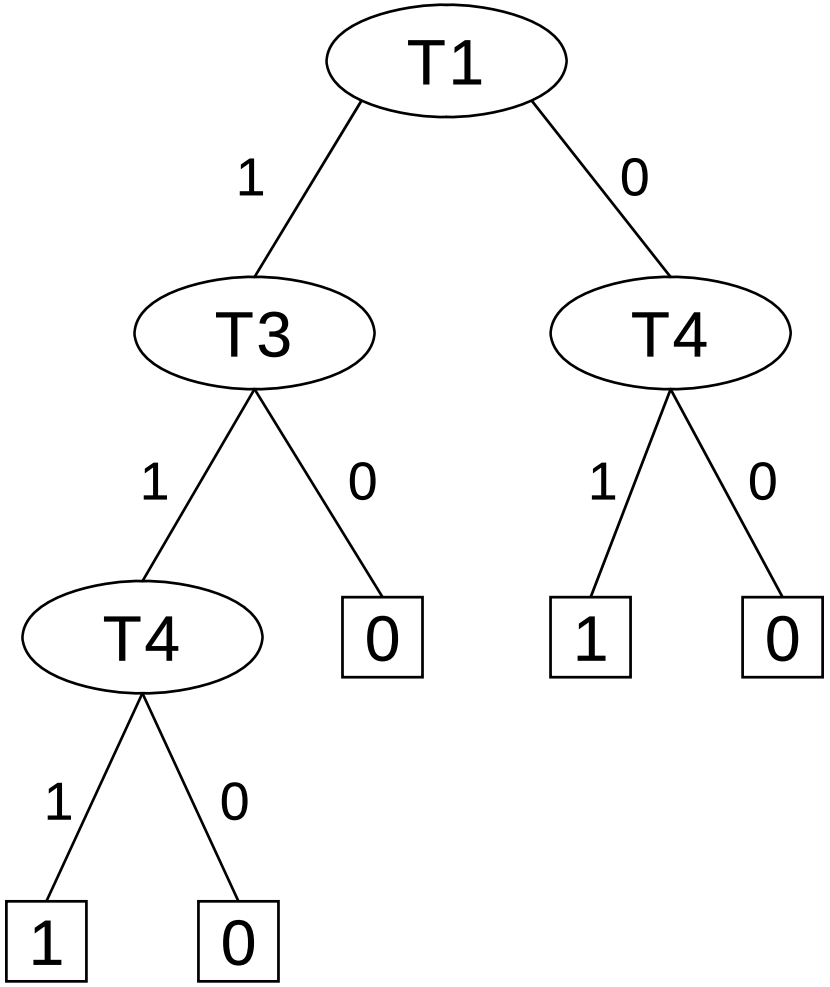
バイオインフォマティクスの問題④　クラス分類

- 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？
- 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

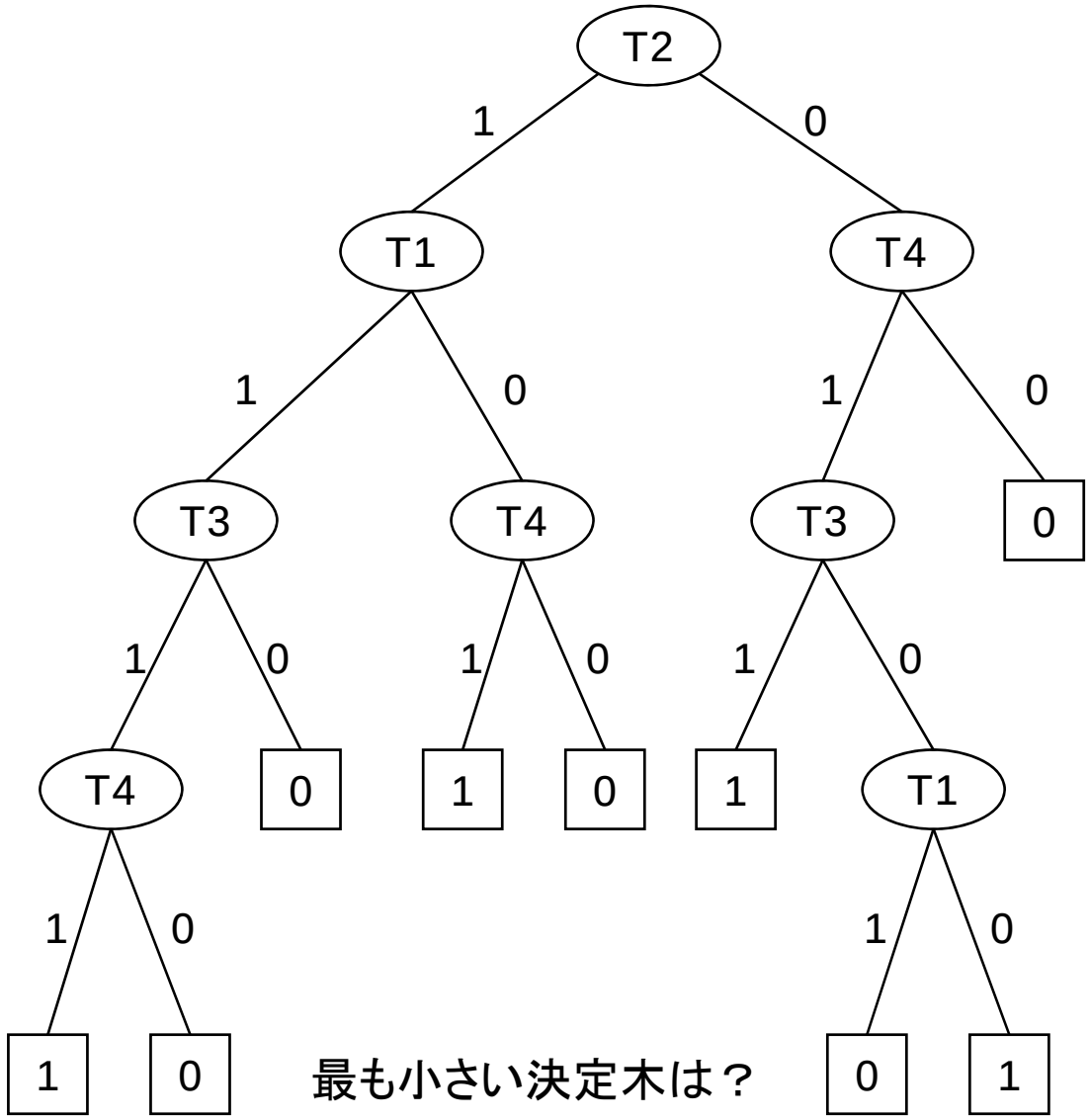
T1	T2	T3	T4	目標属性
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	0
1	0	1	0	0
1	1	0	1	0
1	0	0	1	0
1	1	0	1	0
0	1	0	1	1
0	0	1	1	1
0	1	0	1	1
0	1	0	1	1
0	0	0	1	1
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0

T4
T3
T1
T4



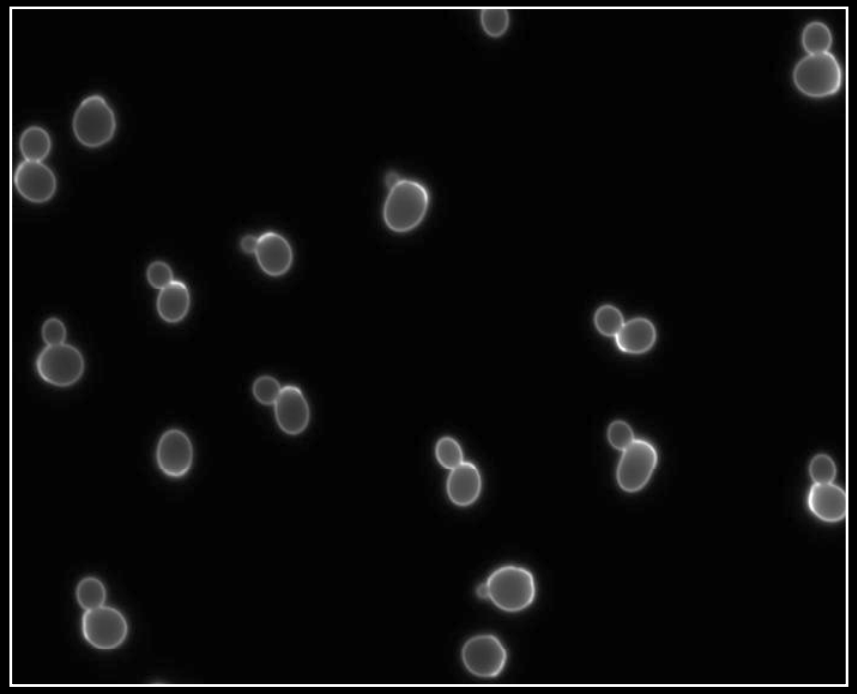
バイオインフォマティクスの問題④　クラス分類

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

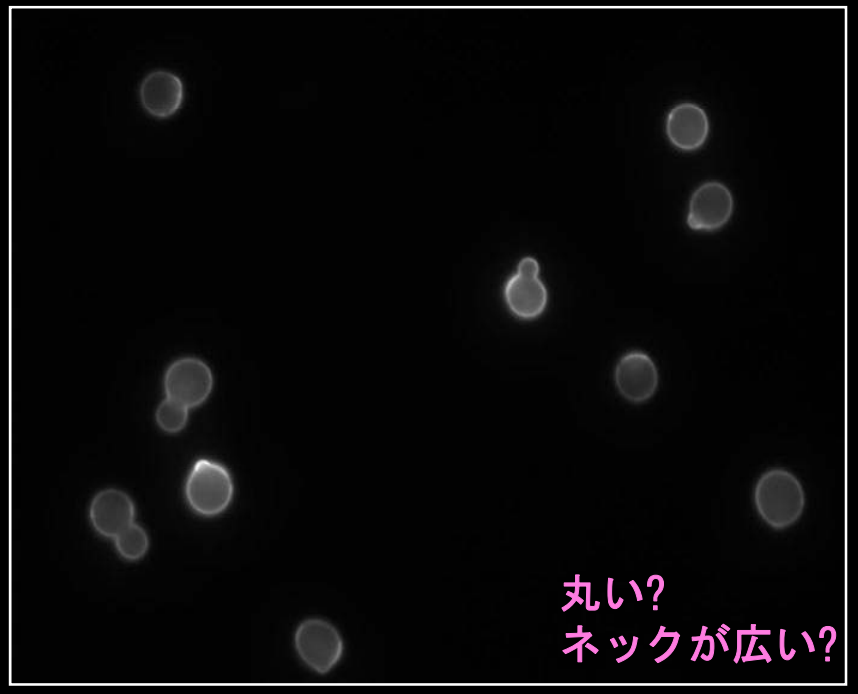


表現型からの遺伝子機能予測

野生株



och1Δ

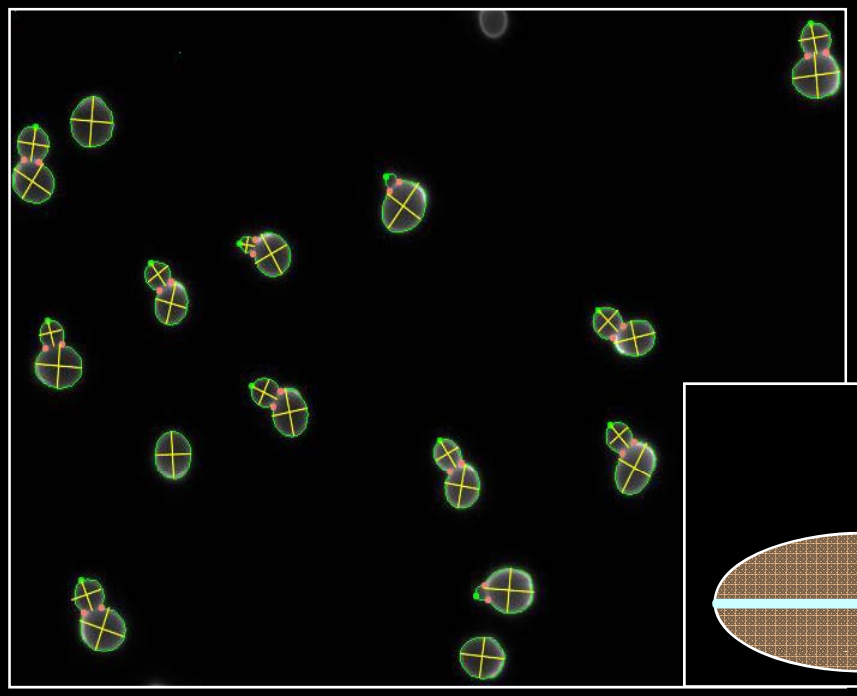


肉眼による比較

- 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

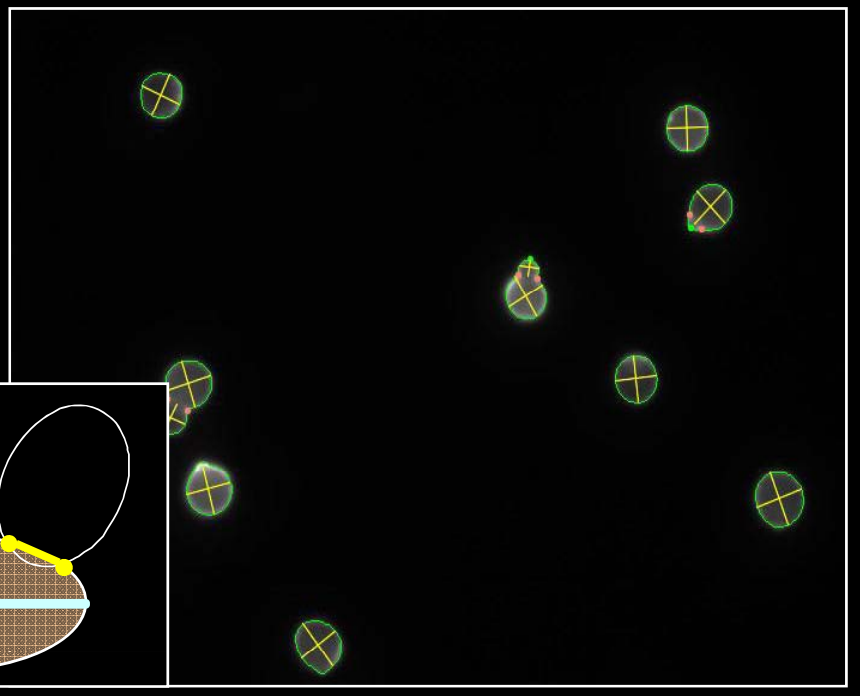


1102.90 ± 301.34 pixels

1.21 ± 0.07 (n=397)

13.28 ± 1.70 pixels (n=246)

och1Δ



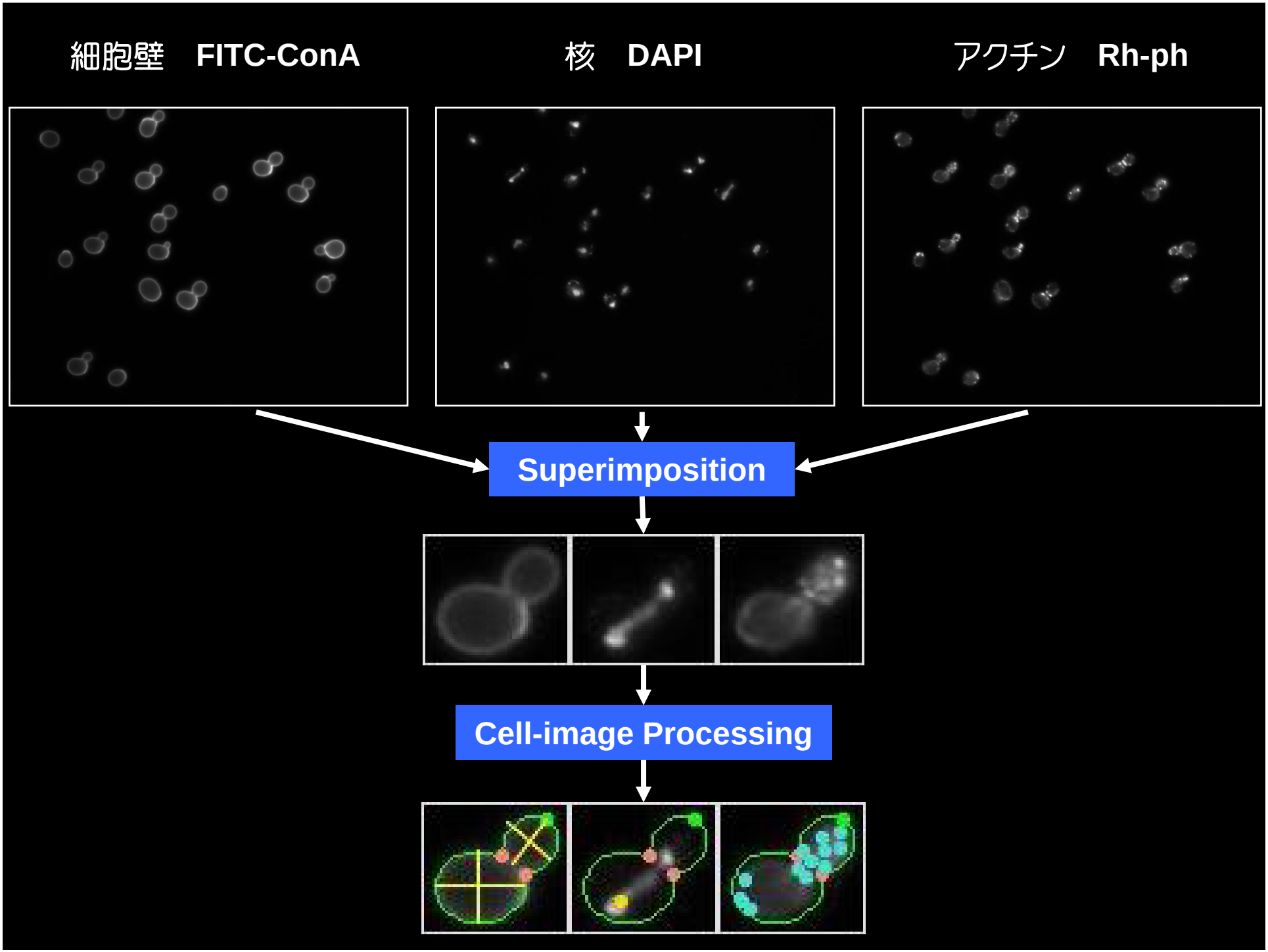
1332.03 ± 457 pixels

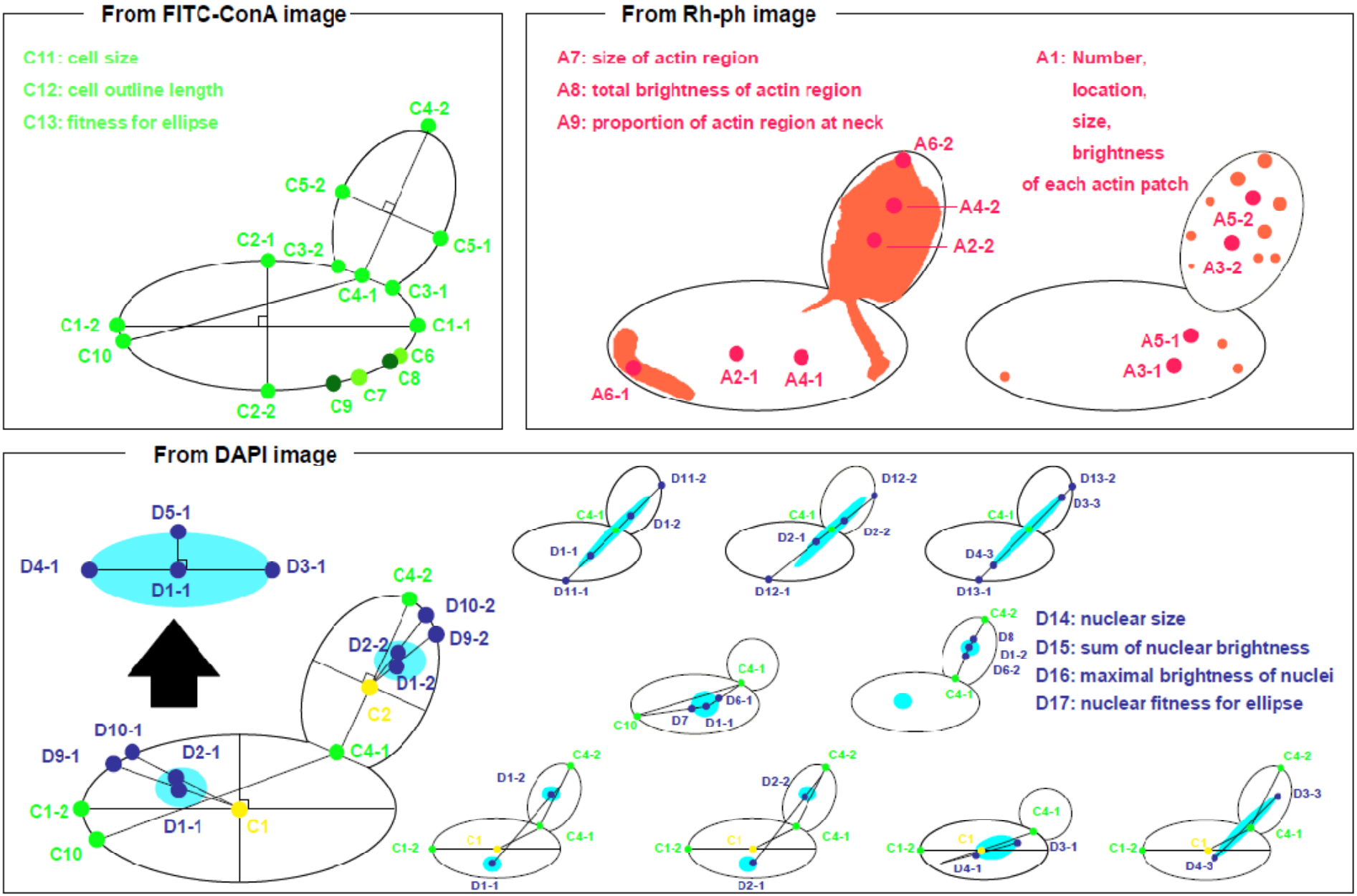
1.09 ± 0.05 (n=243)

17.417 ± 3.00 pixels (n=261)

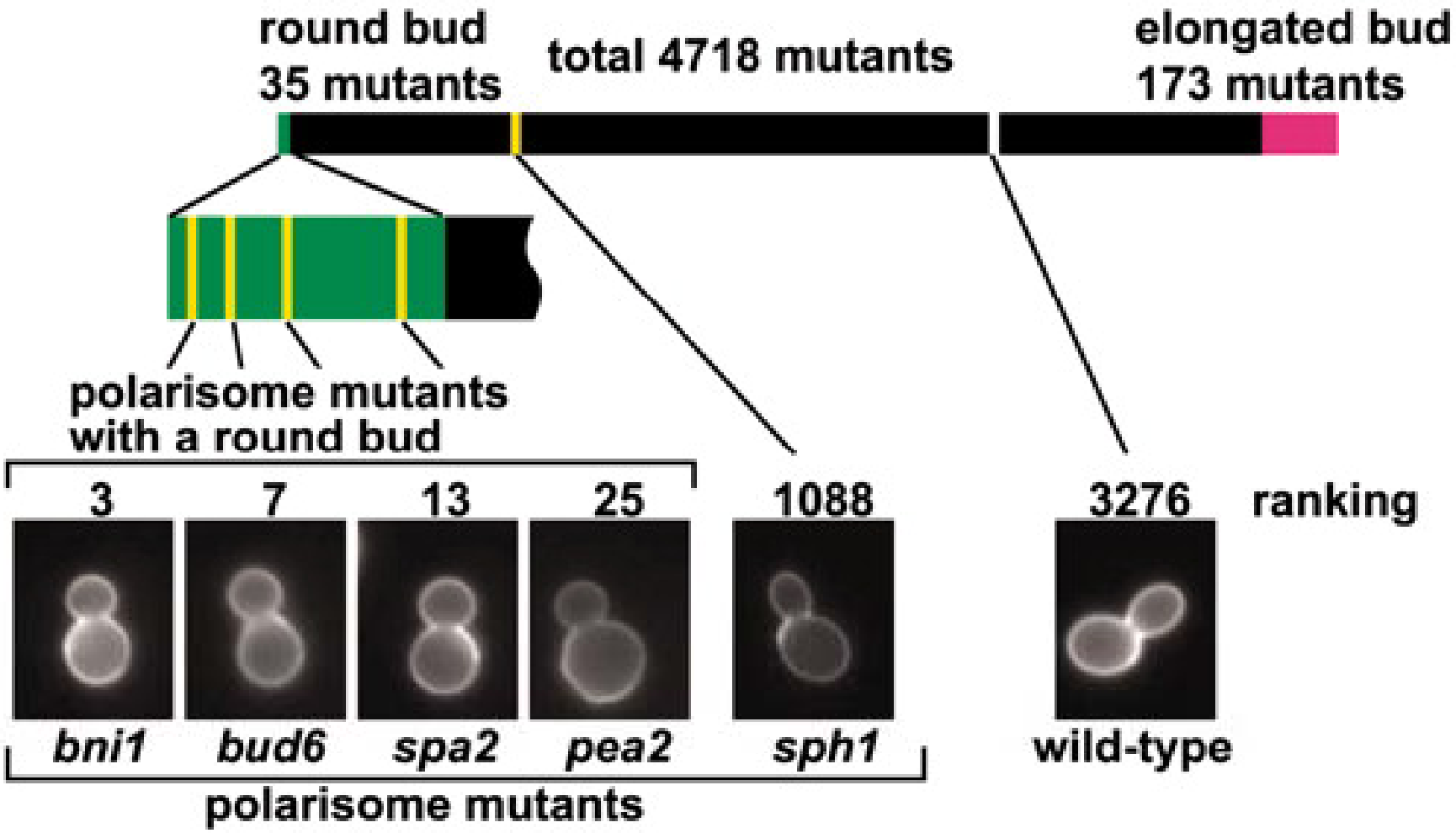
ソフトウエアによる解析

- 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



Ohya et al., High-dimensional and large-scale phenotyping of yeast mutants *Proc Natl Acad Sci U S A. 2005*


round bud
35 mutants
total 4718 mutants
elongated bud
173 mutants
polarisome mutants
with a round bud
3
7
13
25
1088
3276
ranking
bni1
bud6
spa2
pea2
sph1
wild-type
polarisome mutants

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

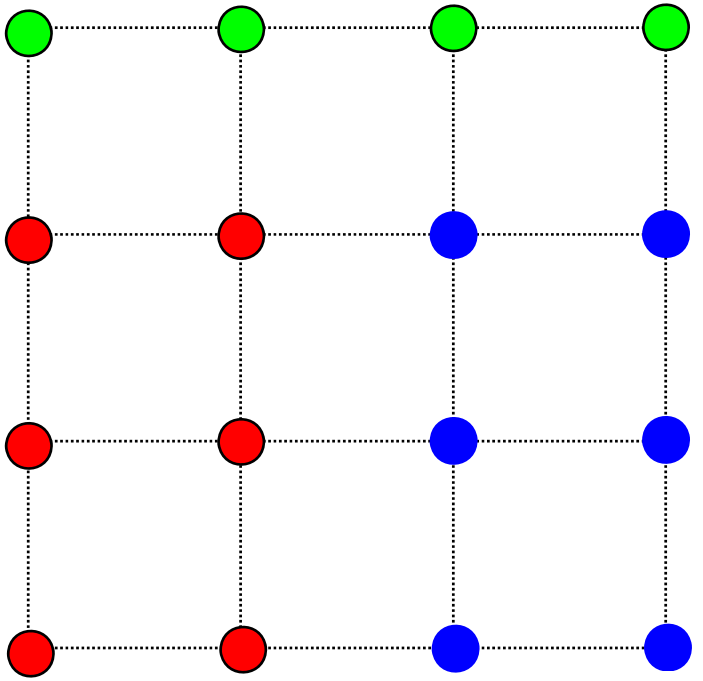
TTT TAA

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
0	0	1	1	1
0	1	0	1	1
0	1	0	1	1
0	1	0	1	1
0	0	0	1	1
1	1	0	1	0
1	1	0	1	0
1	0	0	1	0
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

T3
T1
T4