ゲノムとコンピュータ
Genome and Computing

森下　真一

Shinichi Morishita

染色体，クロマチン構造，ゲノム

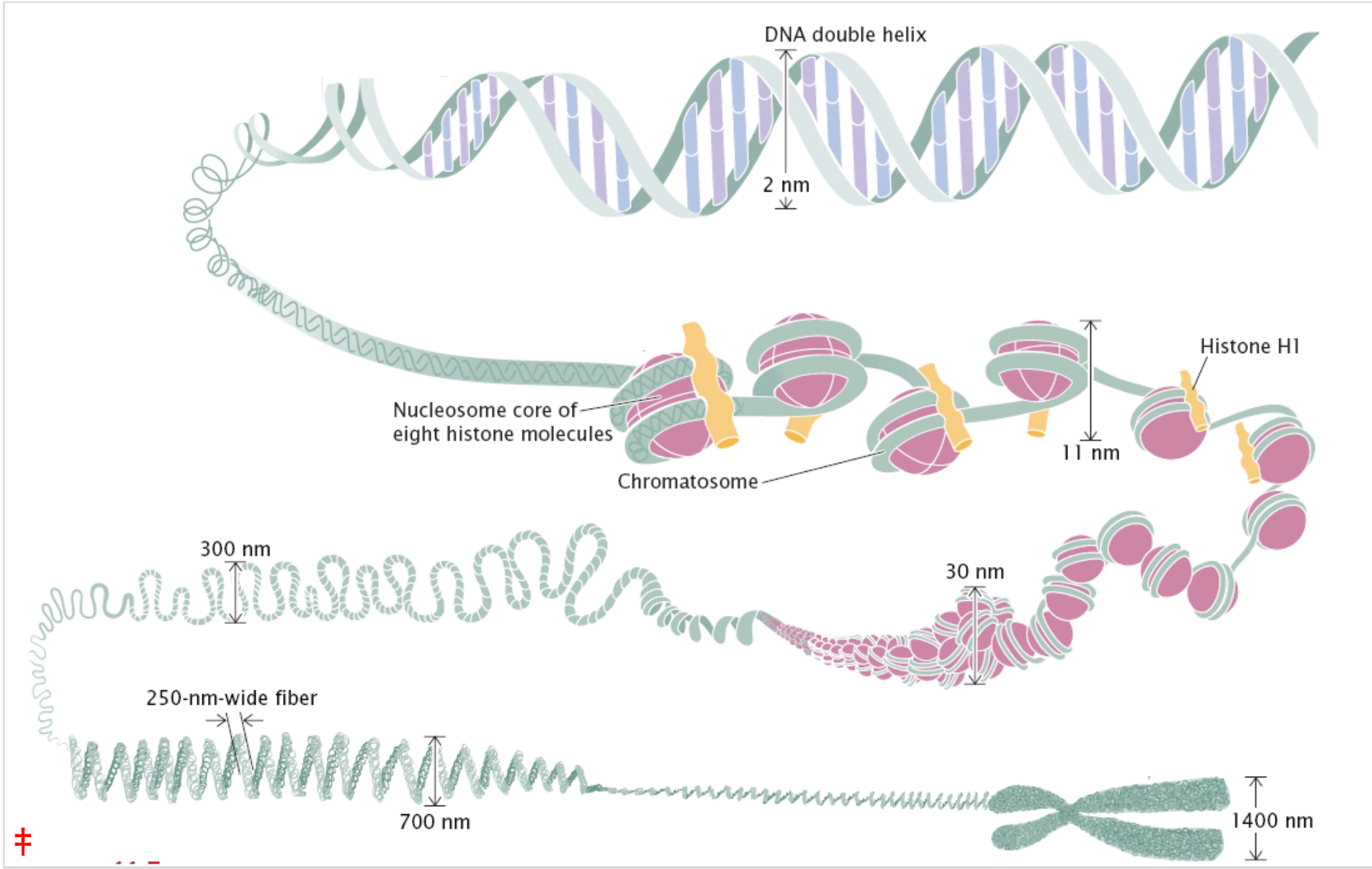


Figure 1 : Chromatin has highly complex structure with several levels of organization.

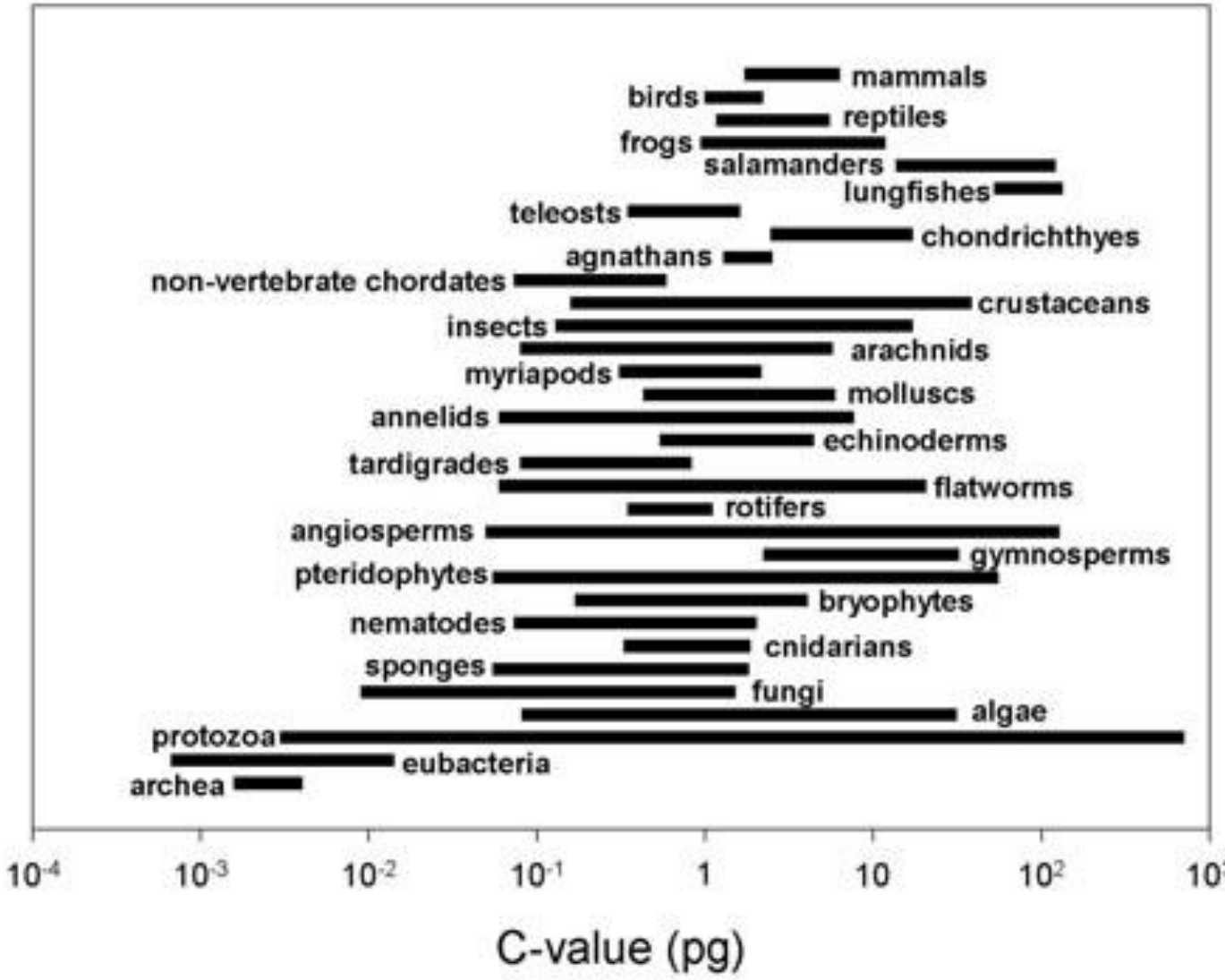
Ref: Annunziato, A. DNA packaging: Nucleosomes and chromatin. Nature Education 1(1), (2008)

ゲノムの解読

- ゲノムサイズと染色体数
  生物を特徴づけているか？
- ゲノムを解読すると何に利用できるか？
- ゲノムはどのようにして解読するか？
- 遺伝子コード領域はどのように見つけるか？
- 近年のゲノム解読装置の革命的進展とは
- クロマチン構造はどのように推定するか？

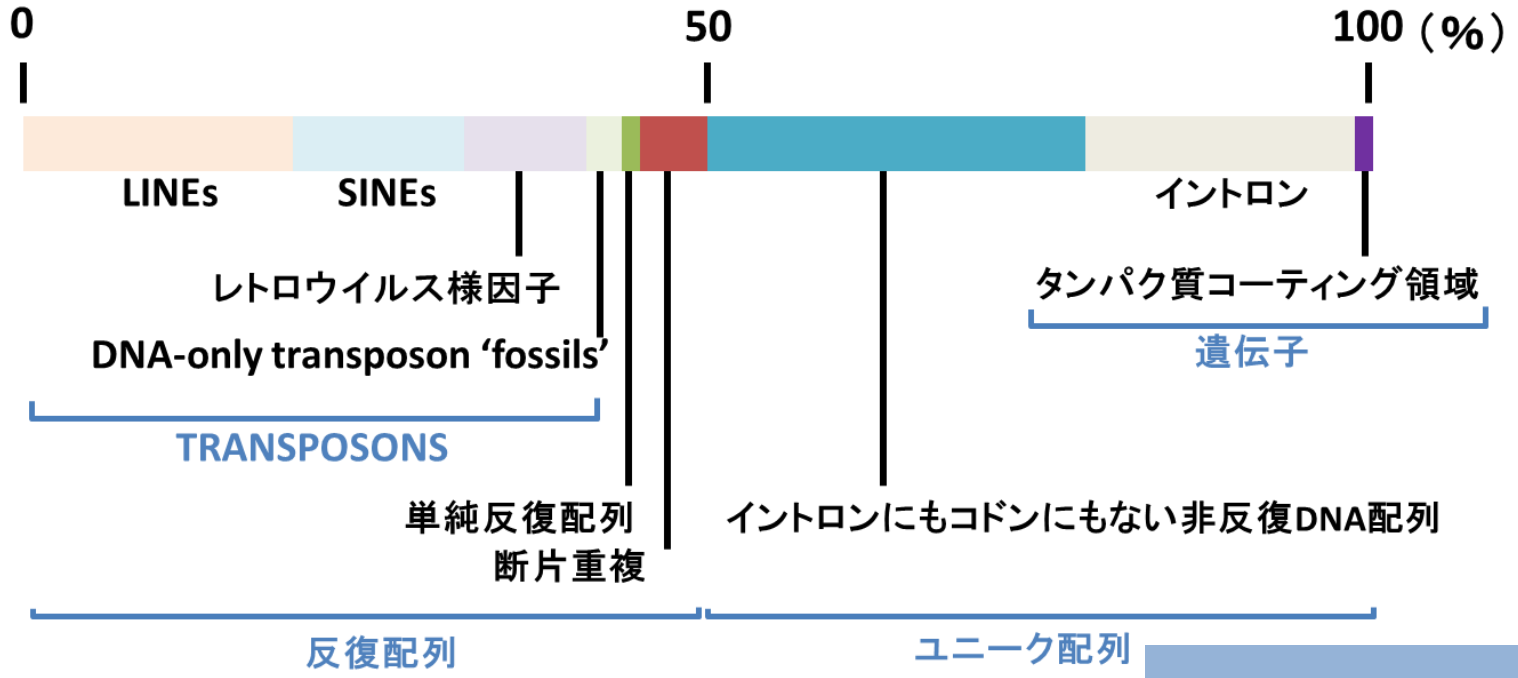
ゲノムサイズ (Genome Size)

1pg (10^{-12} g) ≒ 10億塩基 (正確には 9.78億塩基)



‡ Courtesy of Dr. T. Ryan Gregory
http://www.genomesize.com/statistics.php

なぜゲノムサイズがこれほど違うのか？

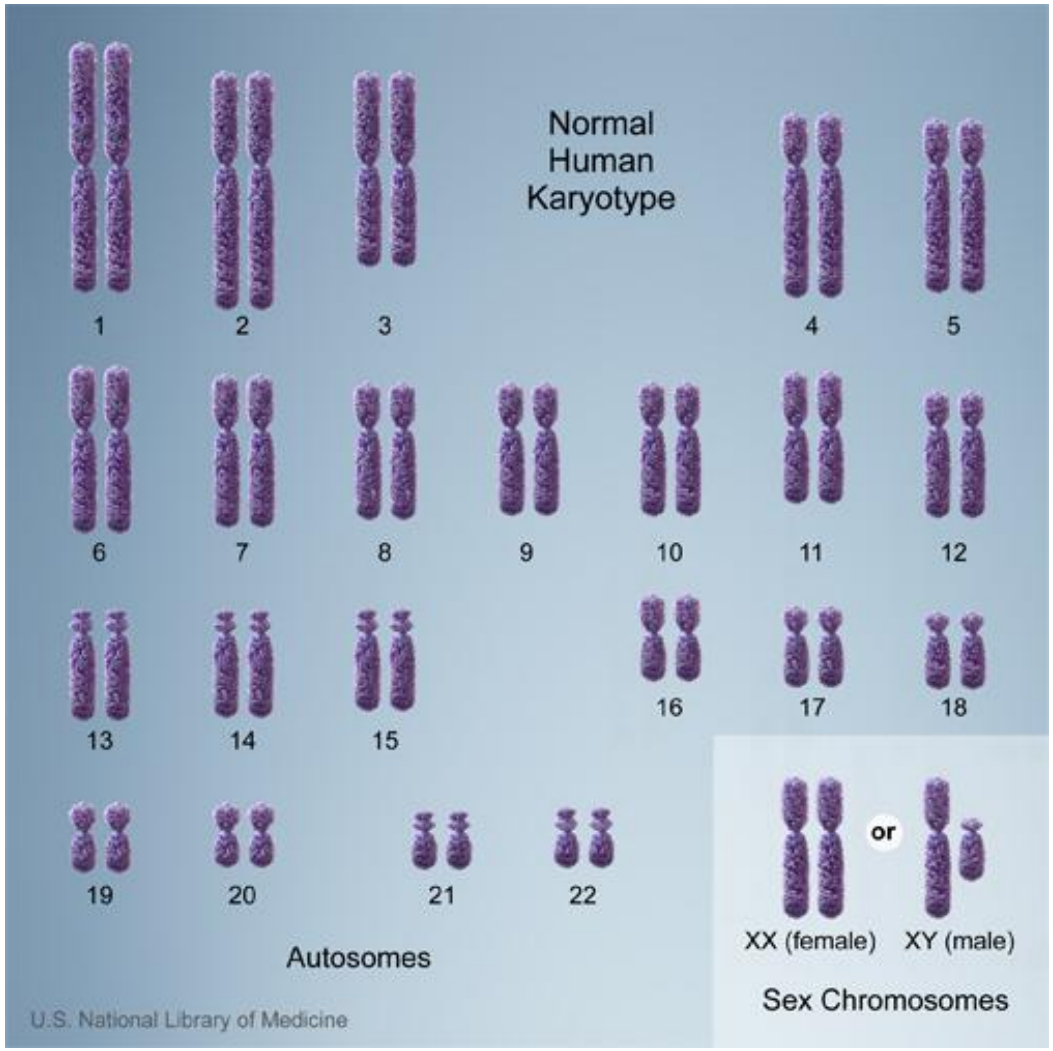


ヒトゲノムの構成

著作権の都合により、ここに挿入されていた画像を削除しました。

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 5－75

ヒト染色体 (Human Chromosomes)



‡ U.S. National Library of Medicine
http://ghr.nlm.nih.gov/handbook/illustrations/normalkaryotype

脊椎動物の染色体数

古生代 Paleozoic | 中生代 Mesozoic | 新生代 Cenozoic

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

百万年前
millions of years ago

500 400 300 200 100 0

脊椎動物 Verterata

無顎口上網 Agnatha

顎口上網 がっこうじょうこう Gnathostomata

軟骨魚類 Chondrichthyes

硬骨魚類 Osteichthyes

条鰭綱 じょうきこう Actinopterygii

肉鰭綱 にくきこう Sarcopterygii

四肢動物類 Tetrapoda

有羊膜類 ゆうようまくるい Amniota

哺乳綱 Mammalia

鳥綱 Aves

爬虫綱 Reptilia

両生類 Amphibia

肺魚類 Dipnoi

多鰭目

全骨類 Holostei

真骨魚類 Teleostei

カメ目 Testudinata

ワニ目 Crocodilia

トカゲ目 Squamata

ヒト, チンパンジ

マウス、ラット

犬

ニワトリ

カメ

ワニ

トカゲ, ヘビ

カエル

シーラカンス, 肺魚

トラフグ

ミドリフグ

メダカ

ゼブラフィッシュ

チョウザメ, ガーボウフィン

ポリプテレス

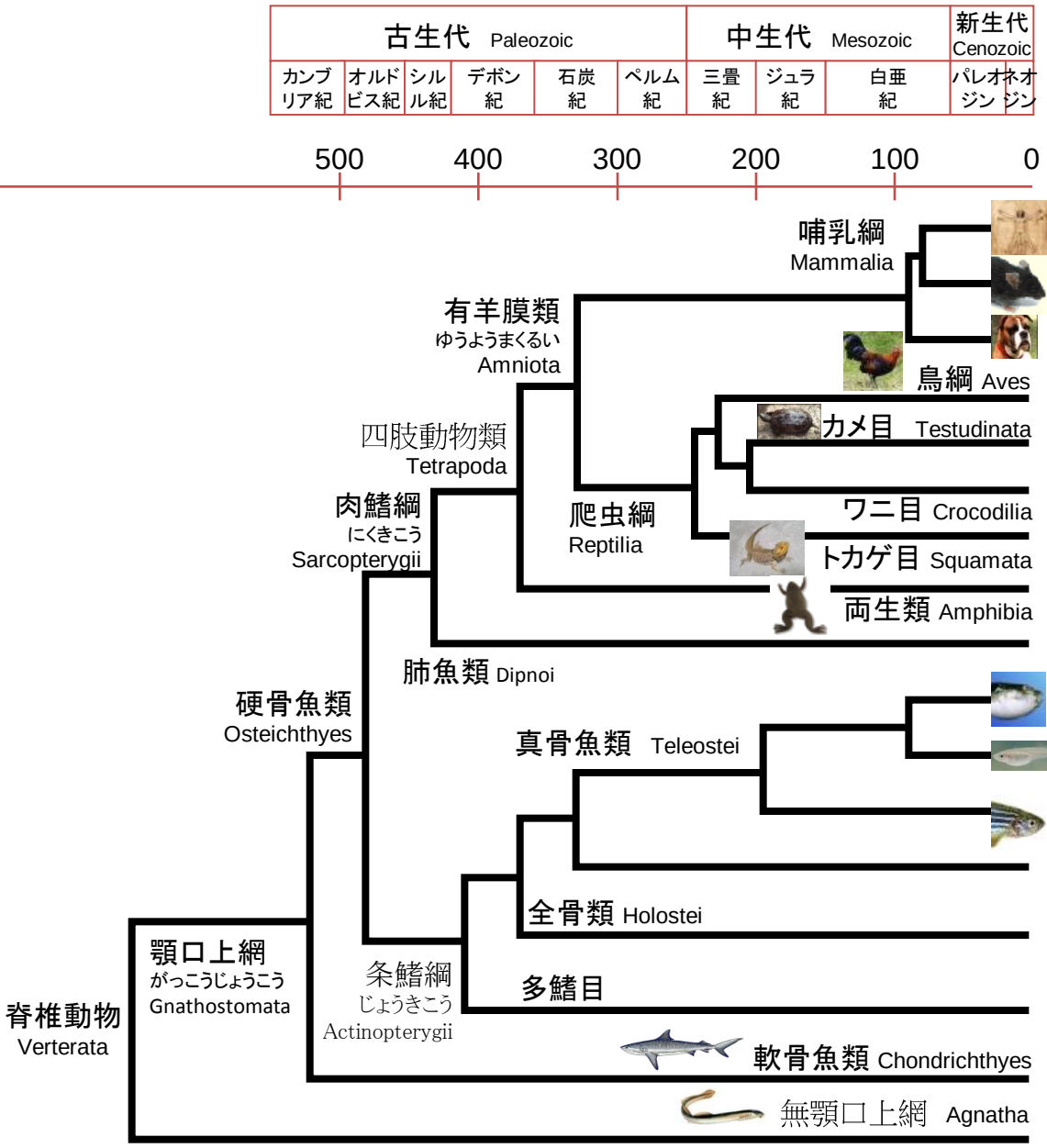
サメ, エイ

ヤツメウナギ

Nakatani et al., 2007 , Genome Res., 17, 1254-1265

染色体数の分布

縦軸：種の数　横軸：染色体数



著作権の都合により、ここに挿入されていた画像を削除しました。

Nakatani et al., 2007 , Genome Res., 17, 1254-1265 Figure6

Nakatani et al., 2007 , Genome Res., 17, 1254-1265

著作権の都合により、ここに挿入されていた画像を削除しました。

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 4-14

ゲノムはどのように利用するか？

- 遺伝子の有無を知る
- ニワトリゲノムの解読　2004年12月
- ニワトリは嗅覚がよくない？
- 嗅覚受容体（匂いの受容体遺伝子）と考えられる遺伝子が218個も予測された
- 飛ぶための遺伝子は？

ゲノムはどのように解読する？

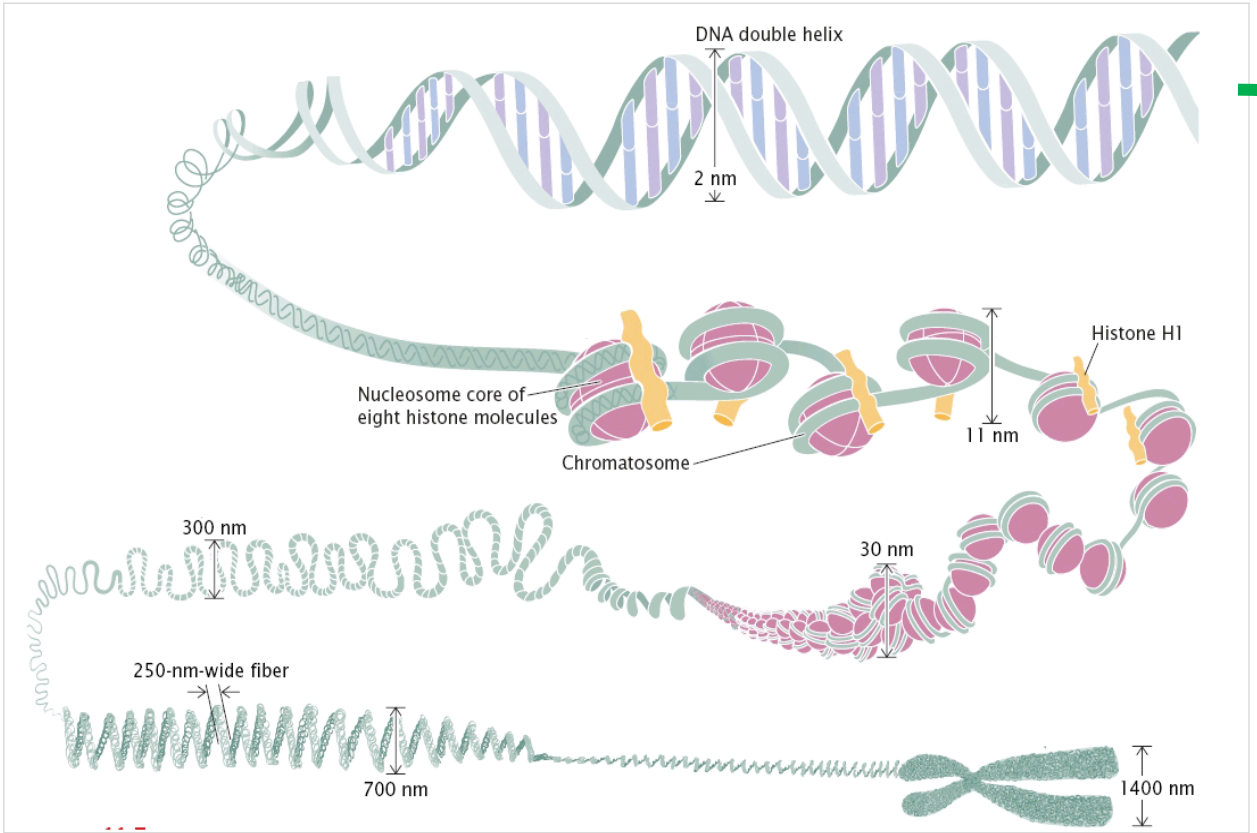
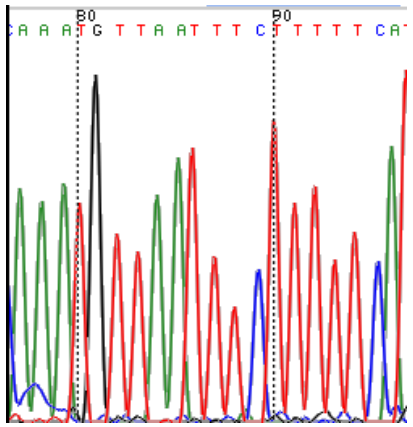


Figure 1 : Chromatin has highly complex structure with several levels of organization.

‡

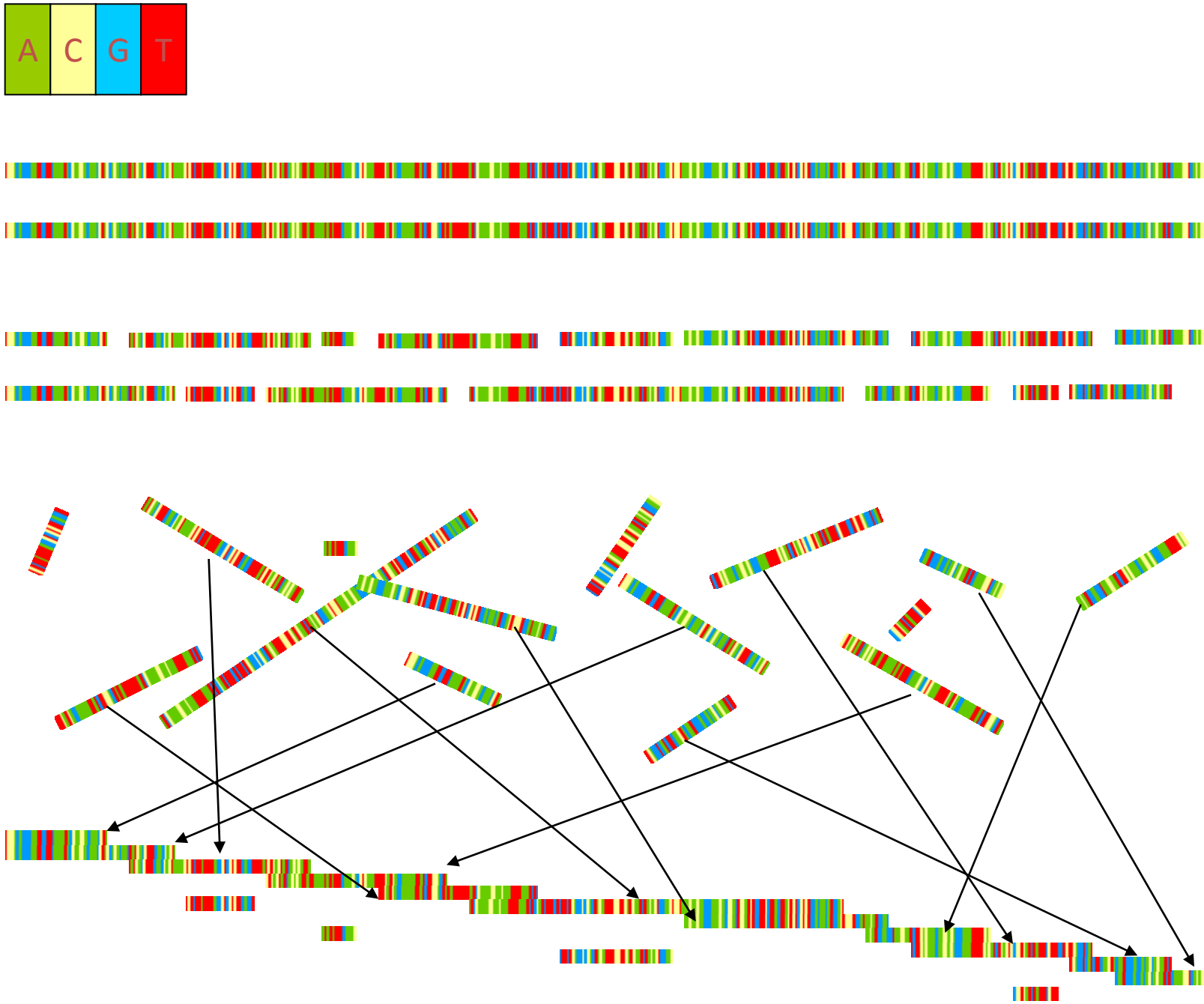
Ref: Annunziato, A. DNA packaging: Nucleosomes and chromatin. Nature Education 1(1), (2008)

‡ 日本アプライドバイオシステムズ



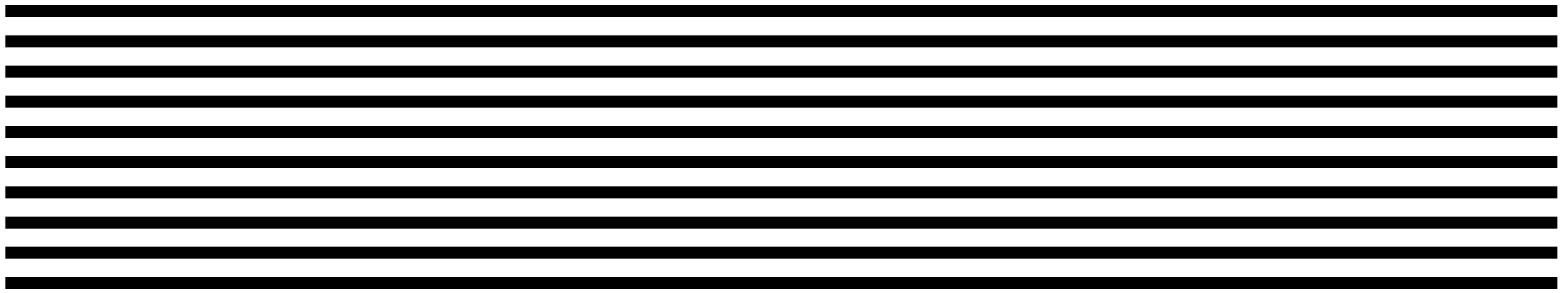
‡ 日本アプライドバイオシステムズ

サンガー法で読める長さは
500~800塩基 ....

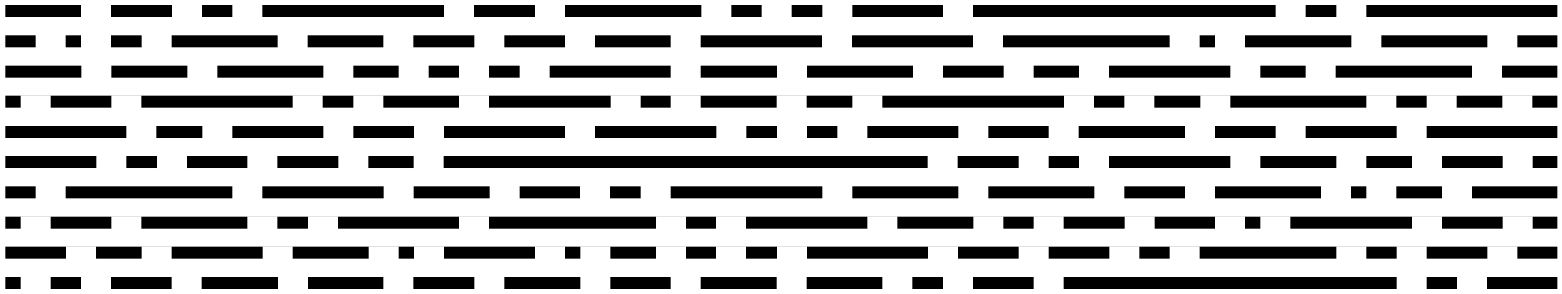


元の絵がわからないジグソーパズル（数百万～数千万ピース）

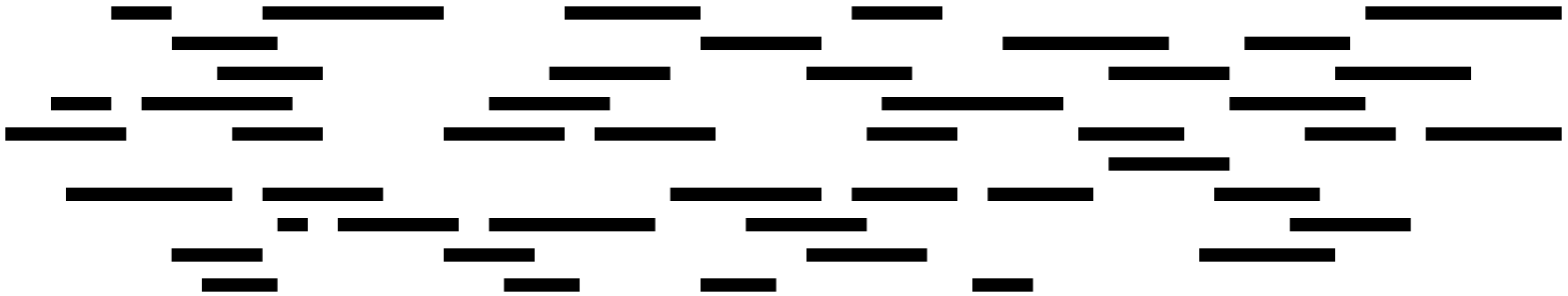
ゲノムをコピーしておく



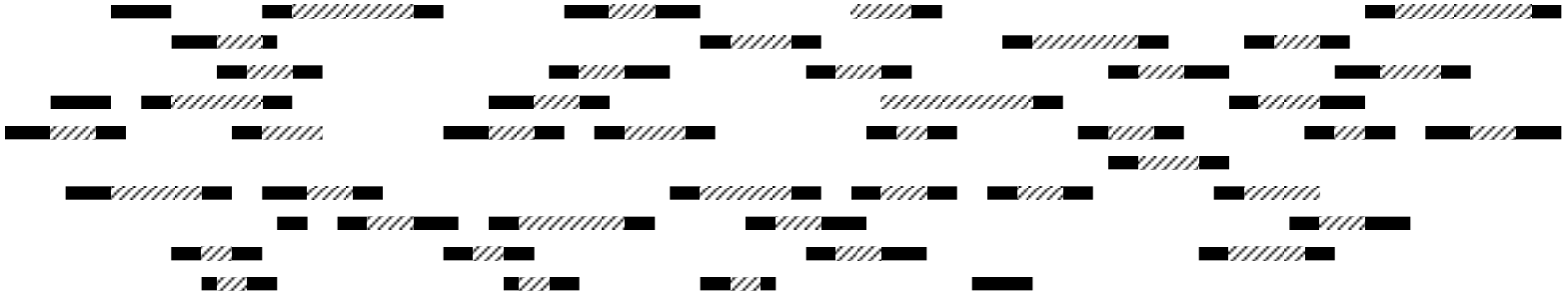
高速な水流でゲノムをランダムに断片化する



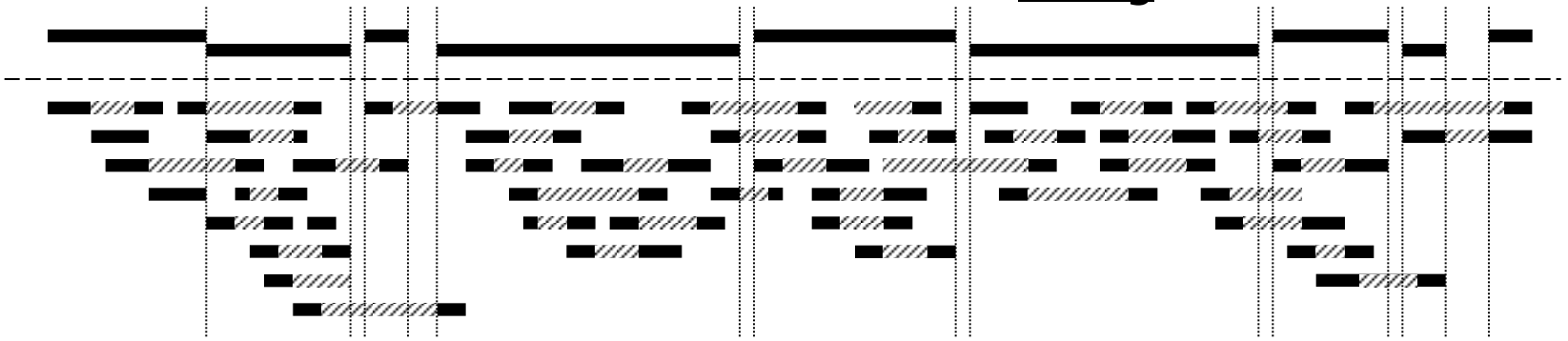
適切なサイズ（たとえば2000塩基対前後）の断片を集める



断片の両端 500～800 塩基を読む

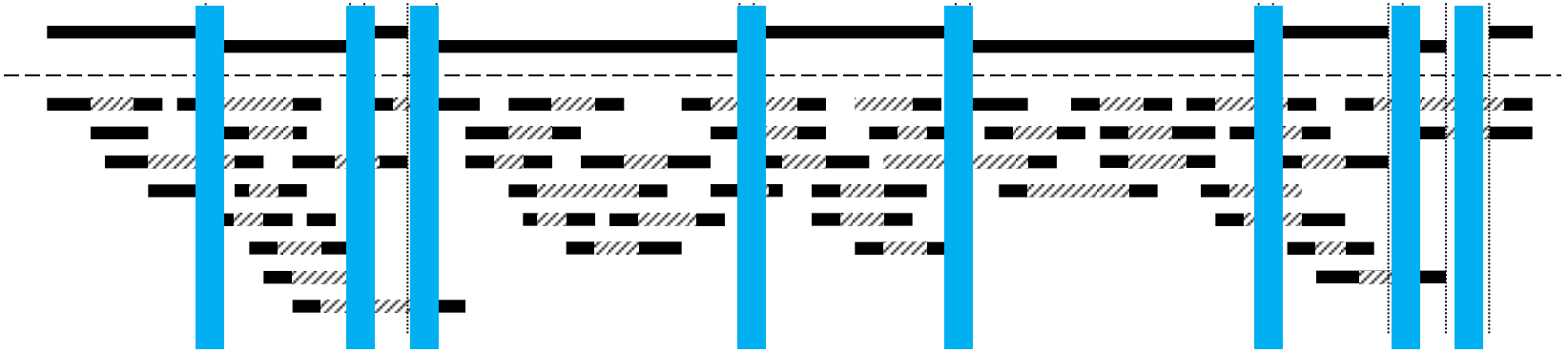


読んだ配列を繋げてゆきコンティグ（連続した配列 contiguous）を生成

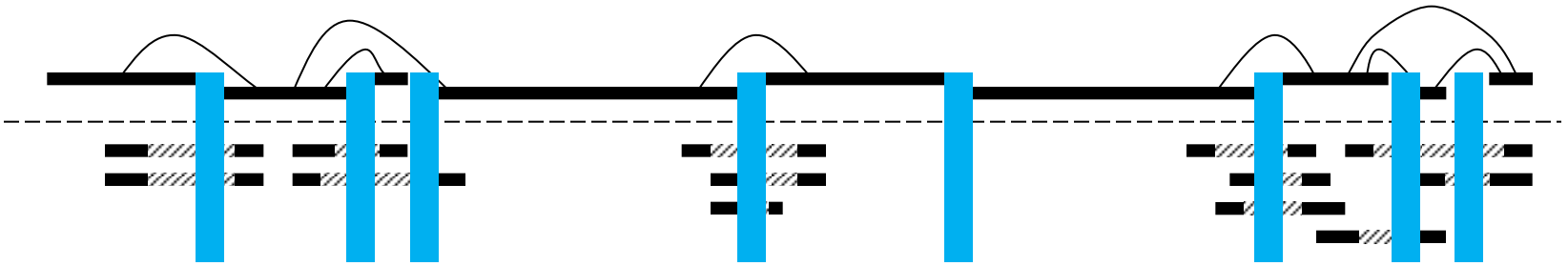


読んだ配列を唯一つの方法では繋げられない例

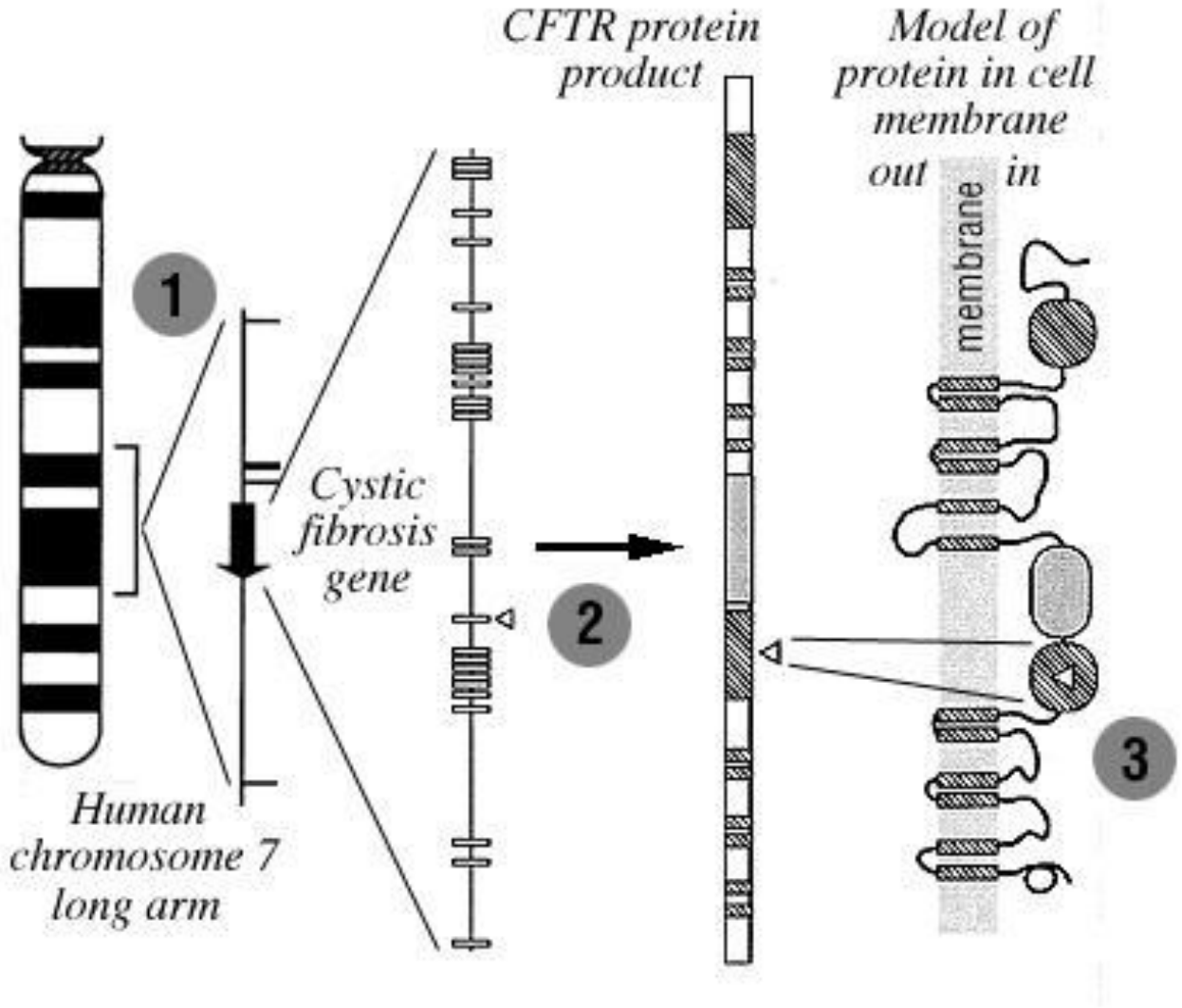
断片を繋げてゆきコンティグ（連続した配列 contiguous）を生成



コンティグを不連続に繋げてゆく

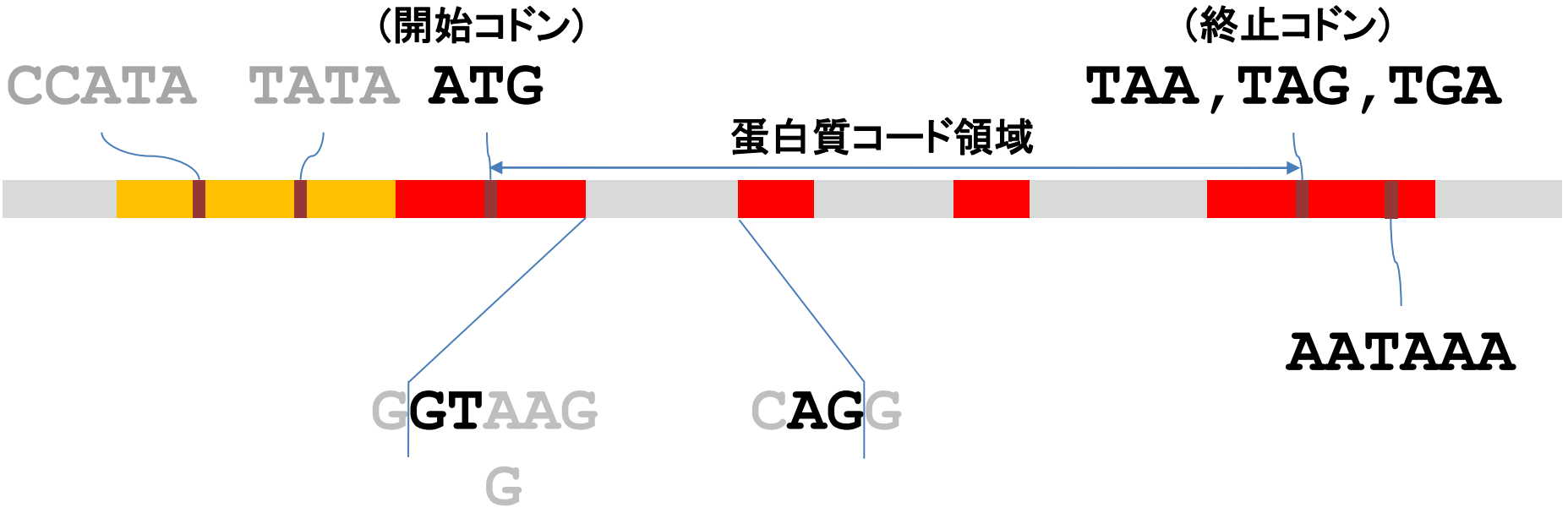


ゲノム中の遺伝子コード領域



‡ Barry Shell, www.science.ca
http://www.science.ca/scientists/scientistprofile.php?pID=19&pg=1

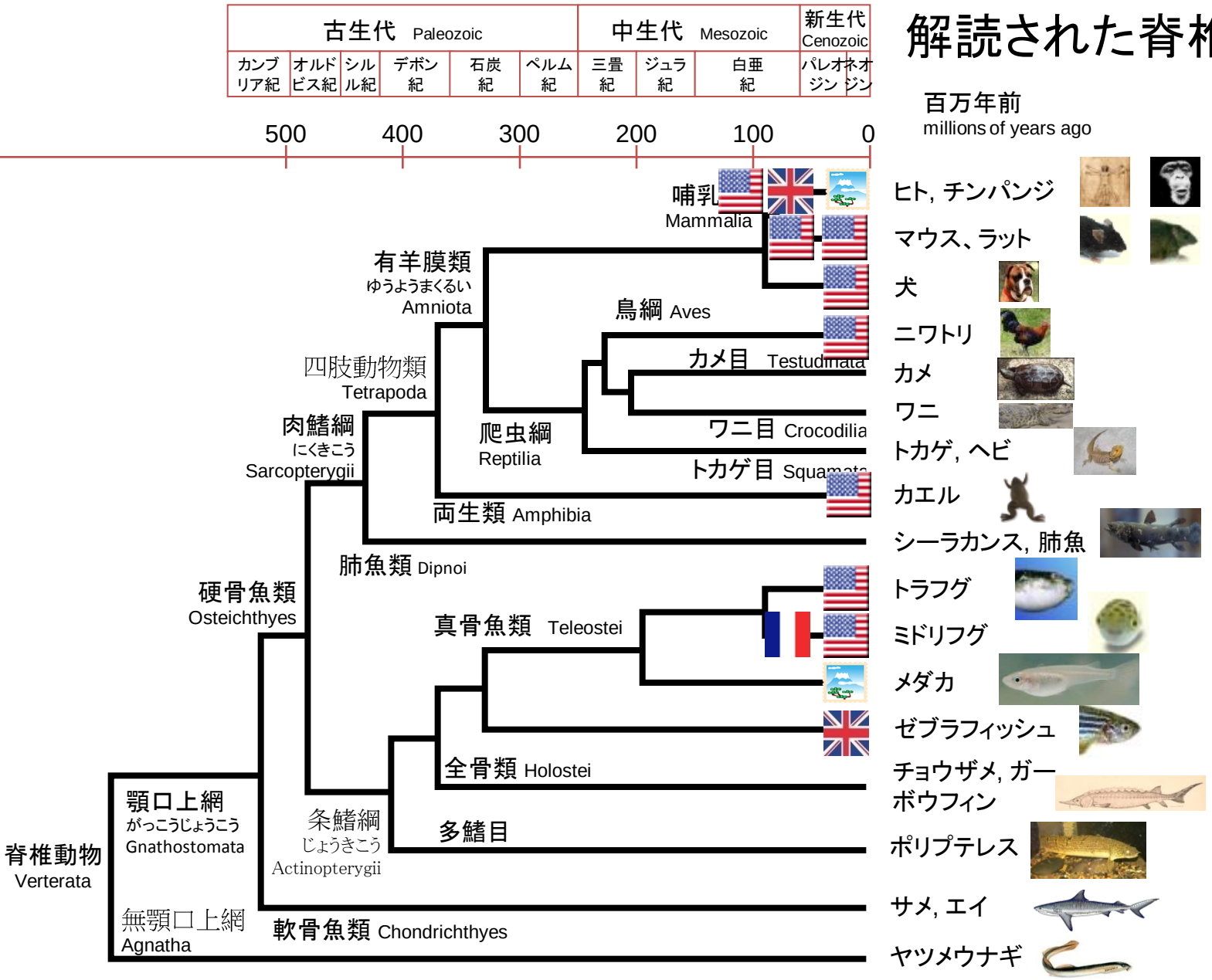
遺伝子コード領域は
ゲノムだけから予測できるか？



コーディングポテンシャル

コドンの使用頻度には生物固有の偏りがある
コード領域には3塩基の周期性がある
6文字塩基（2コドン分）の出現頻度の偏りが標準的に利用
Hidden Markov Model

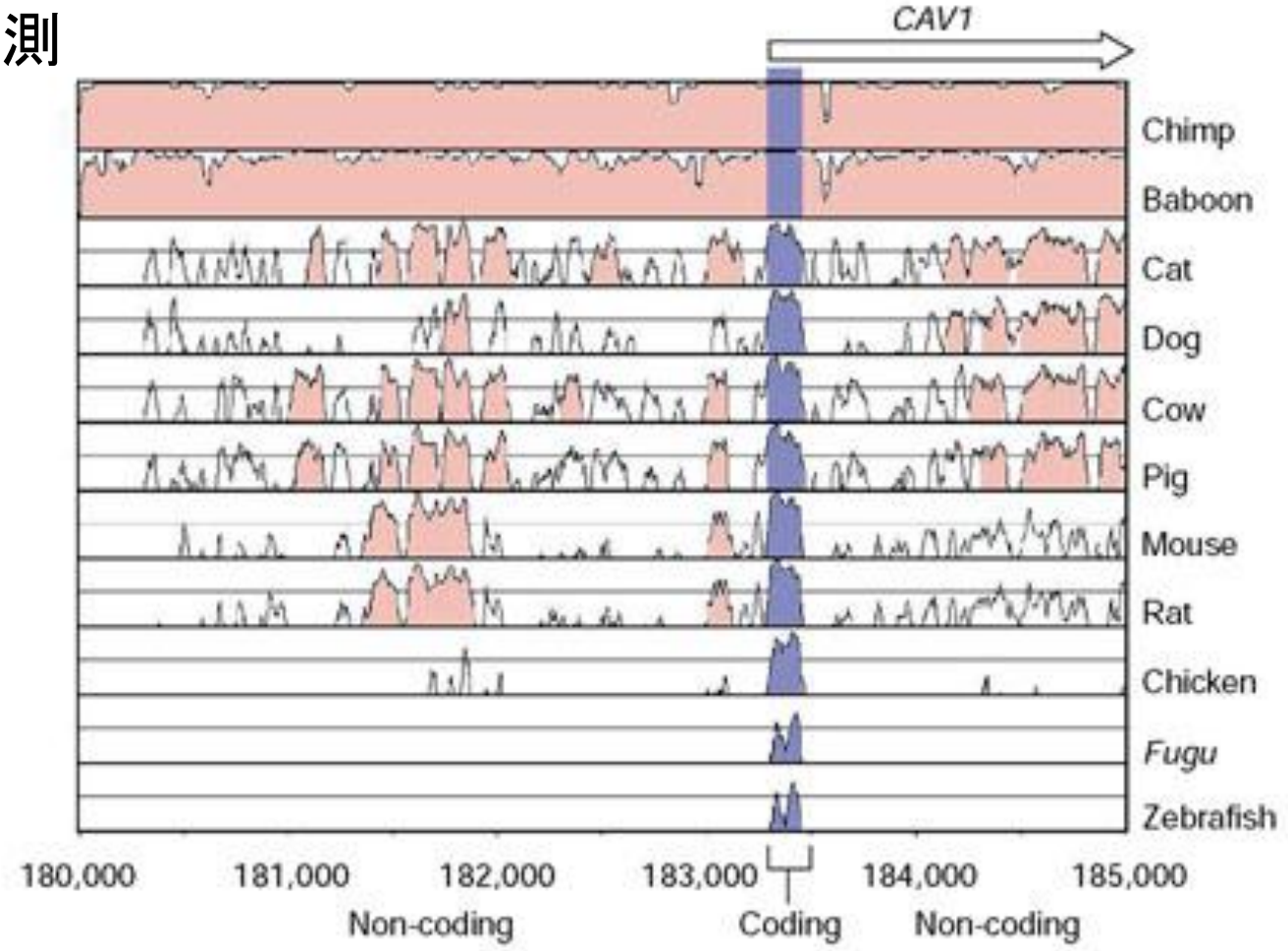
解読された脊椎動物ゲノム



ゲノムを比較し、保存されている領域を見つけ、遺伝子を予測

Nakatani et al., 2007 , Genome Res., 17, 1254-1265

ゲノムを比較し、
保存されている
領域を見つけ、
遺伝子を予測



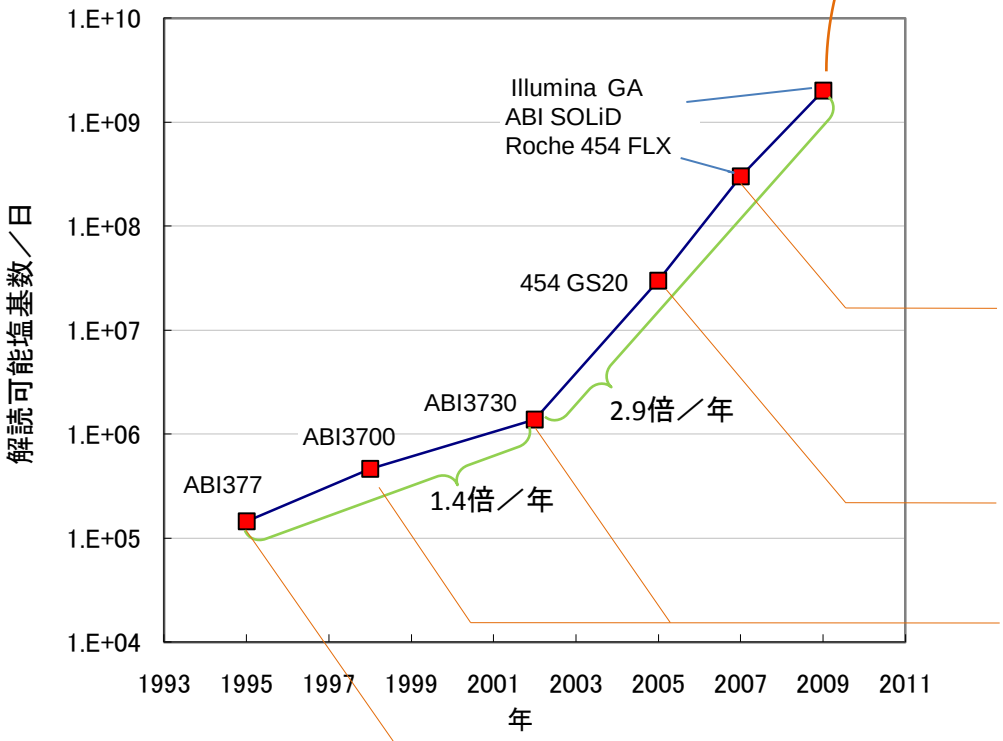
‡ Dubchak and Frazer, 2003 , Genome Biology, 4,122
http://genomebiology.com/2003/4/12/122

遺伝子配列の収集

- mRNA から cDNA を合成
- cDNAはベクターに組込み増殖させ保存
  cDNA ライブラリー
- 我が国が世界的にも強い分野
  菅野純夫 （東大医科研） ヒト　等
  林崎良英 （理研） マウス
- すべての mRNA を見つけるのは困難

ゲノム解読の高速化

	Illumina GAIIx	ABI SOLiD 3	Roche 454FLX Titanium
リード長 (塩基数)	75 x 2 = 150	50	500
リード数 (億)／実験	0.96～1.2	4	0.01
日／実験	9.5	16	0.4 (10時間)
単位時間での塩基数 塩基数 (億)／日	15～19	12.5	12
サンプル量 (μg)	0.1～1	0.01～ 5	3～5



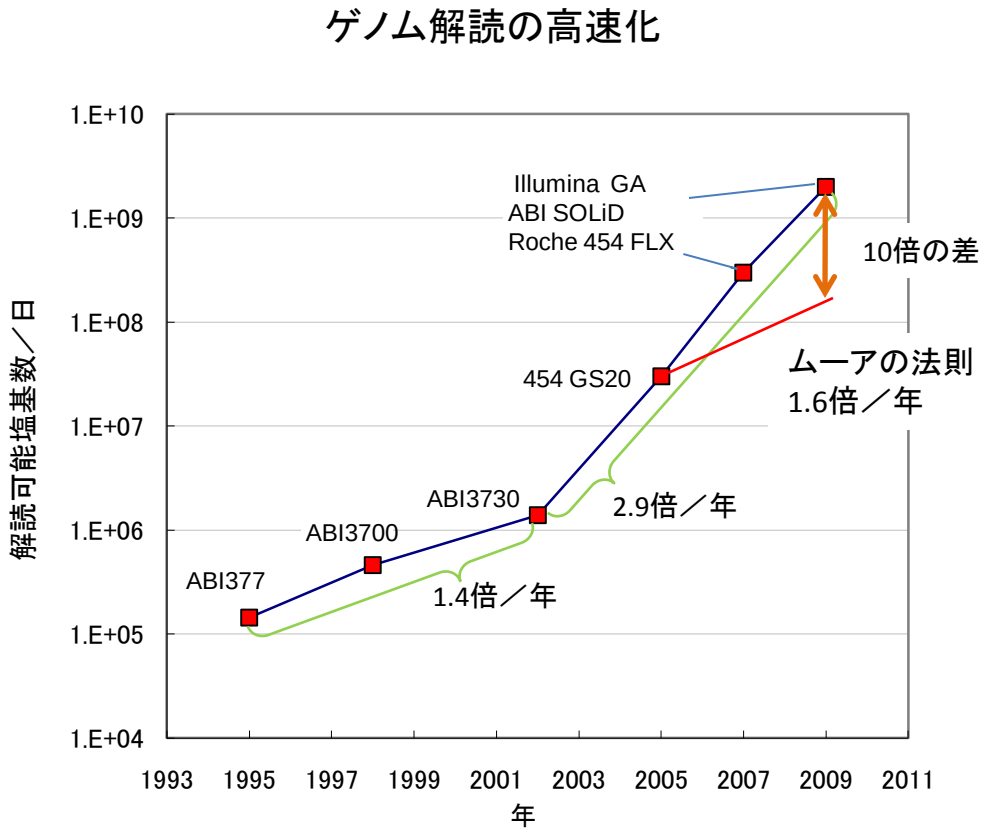
- ゲノムの再解読・転写開始点・クロマチン構造・DNAメチル化・RNA-Seq⇒Illumina GA
- 大規模ゲノムの de novo 解読・全長cDNA解読・選択的スプライシング ⇒ Roche 454
- 1分子計測の実現 ⇒ 発生初期を観察
- ワトソンゲノムの再解読（454: ～250 b）
- アジア人ゲノムの再解読（Illumina: ～35 b）
  変異, 挿入削除, 逆位　等
- DNAメチル化 (Roche 454: 100～250 b, Illumina: 36 b after target capture)
- RNA-Seq (Illumina: 25～35 b)
- 転写開始点の網羅 (Illumina/SOLiD: 25b)
- クロマチン構造 (Illumina/SOLiD: 25 b)

- ネアンデルタール人ゲノムの一部解読
- クロマチン構造（リード長 ～100 b）

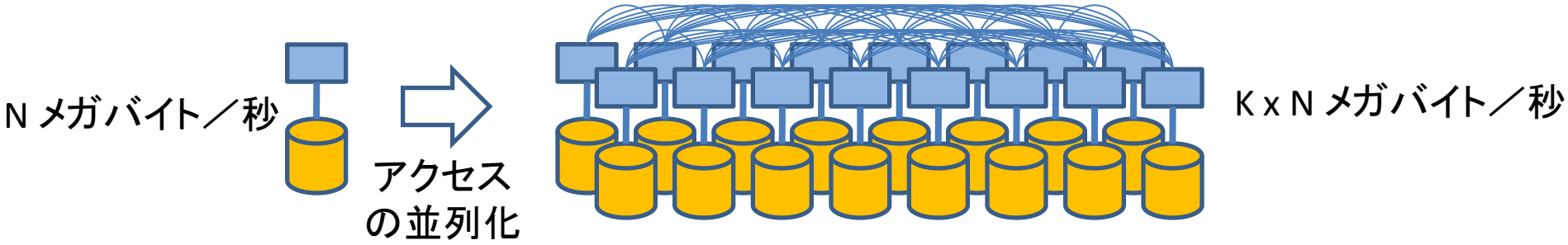
- ヒト等の大規模脊椎動物ゲノムの de novo 解読 ・ 全長cDNAの解読（リード長 500～800 b 塩基）

遺伝子配列の一部を収集

次世代シークエンサーの性能向上に追いつくための計算機資源の並列化



- ムーアの法則
  “CPUの性能(集積回路上のトランジスタ数)は 1.5年で2倍になる”
- ムーアの法則を凌駕する次世代シークエンサーの性能向上
  4年間で約10倍の差
- 約10倍の個数のCPUを並列化して処理速度の維持
- 二次記憶装置へのアクセスが隘路
  並列アクセスによる解決

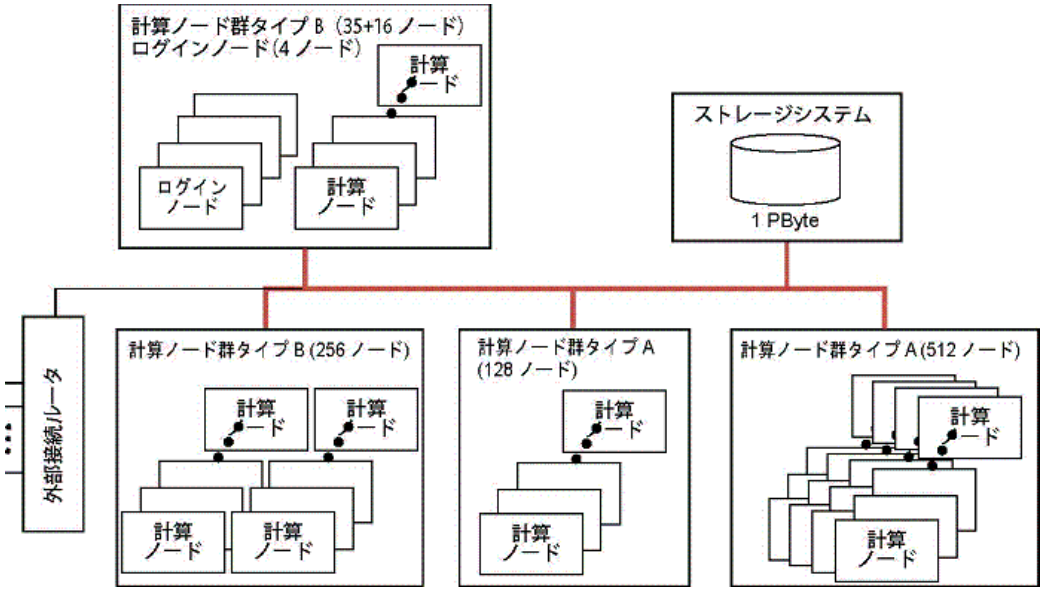


東大情報基盤センター　HA8000クラスタシステム

ノード	理論演算性能	147.2 GFLOPS
	プロセッサ数(コア数)	4(16)
	主記憶容量	32 GB(936ノード) 128 GB(16ノード)
	ローカルディスク容量	250 GB(RAID1 OS領域を含む)
プロセッサ	プロセッサ(周波数)	AMD Opteron プロセッサ 8386(2.3GHz)
	キャッシュメモリ	L2:512 KB/コア L3:2 MB/プロセッサ
	プロセッサコア 理論演算性能	9.2 GFLOPS

国内最大

TOP 500 (世界ランキング)
2008/11　27 位
2008/6　16位



出典
http://www.cc.u-tokyo.ac.jp/ha8000/
‡ 東京大学情報基盤センター

‡

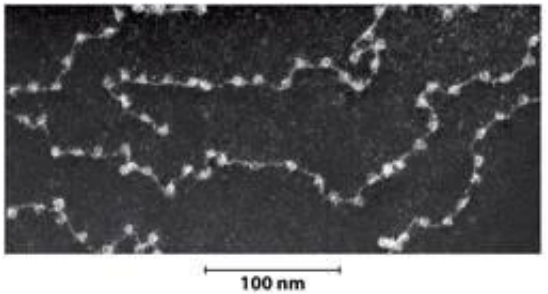
Jun Wang (1976 -)

クロマチン構造の網羅的把握

著作権の都合により、
ここに挿入されていた画像を削除しました。

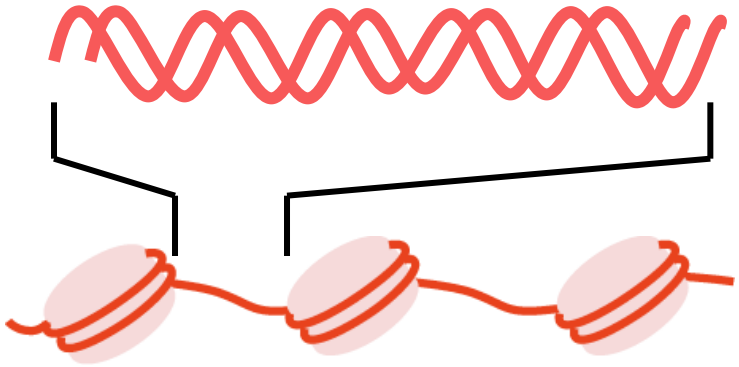
Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 4-72

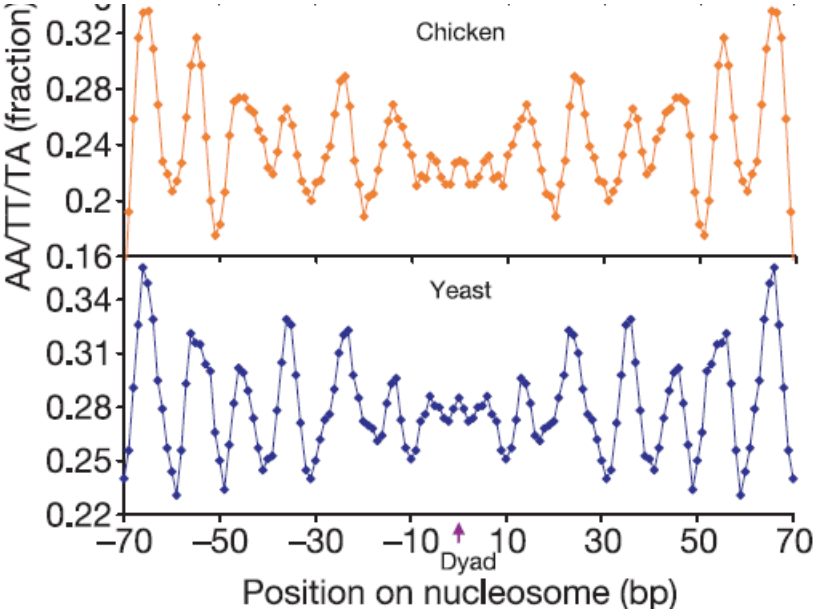
‡

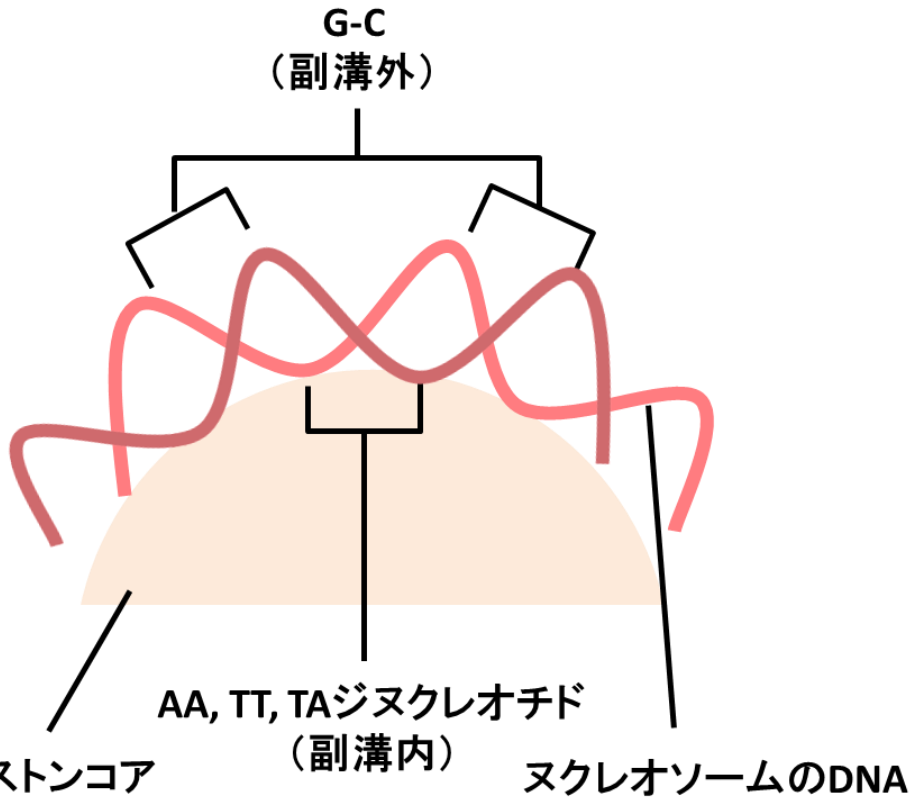


Jeremy M. Berg,
2006,
Biochemistry 6th edition,
W.H. Freeman & Co.

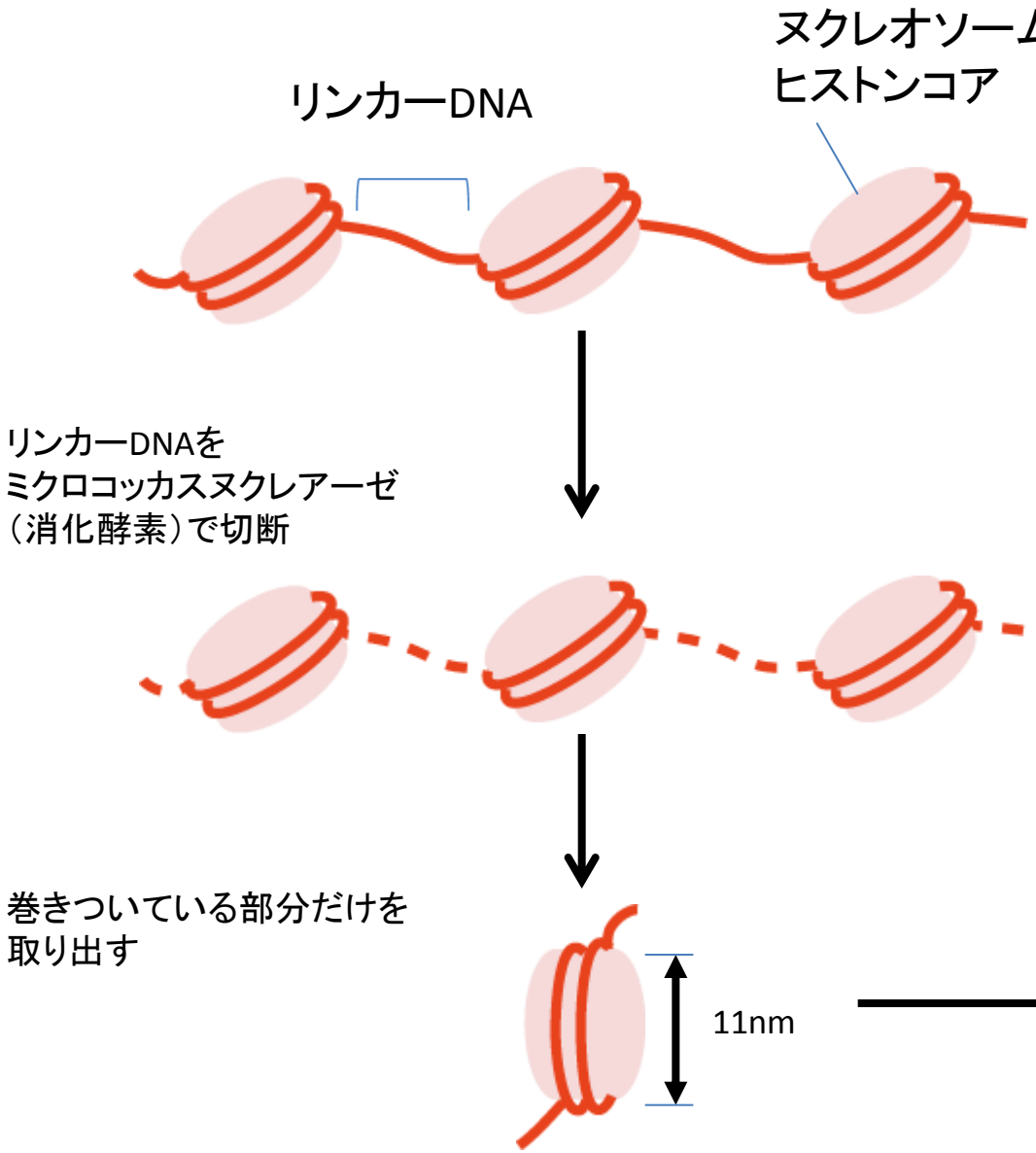
ヌクレオソームコアの位置は
ゲノム配列だけから
予測できるか？







‡Reprinted by permission from Macmillan Publishers Ltd: Segal et al., Nature 442(7104):772-8 , copyright（2006)



- Nucleosome core は
  160～200塩基対
  ごとに存在
- ヒトゲノムの場合
  1500～2000万個の
  nucleosome core
- 2002年 約2000配列/日
  (ABI 3730)
- 2007年 ~1000万配列/日
  (Illumina GA)
  手に届く範囲

In a population of cells, positions of nucleosome cores are unlikely to be stable.

著作権の都合により、
ここに挿入されていた画像を削除しました。

Molecular Biology of the Cell - Fifth Edition
Garland Science (2008)
Figure 4-23 （part2of2）

まとめ

- ゲノムサイズと染色体数は必ずしも生物を特徴づけているわけでない
- ゲノムは多様に利用されている
- 繰返し配列がゲノムの解読を困難にしている
  コンピュータ解析が不可欠
- 遺伝子コード領域の推定には、予測、ゲノム比較、cDNA収集の3通りの方法が併用される
- 近年のゲノム解読装置の能力は革命的に進歩している
- クロマチン構造の把握が可能になってきた