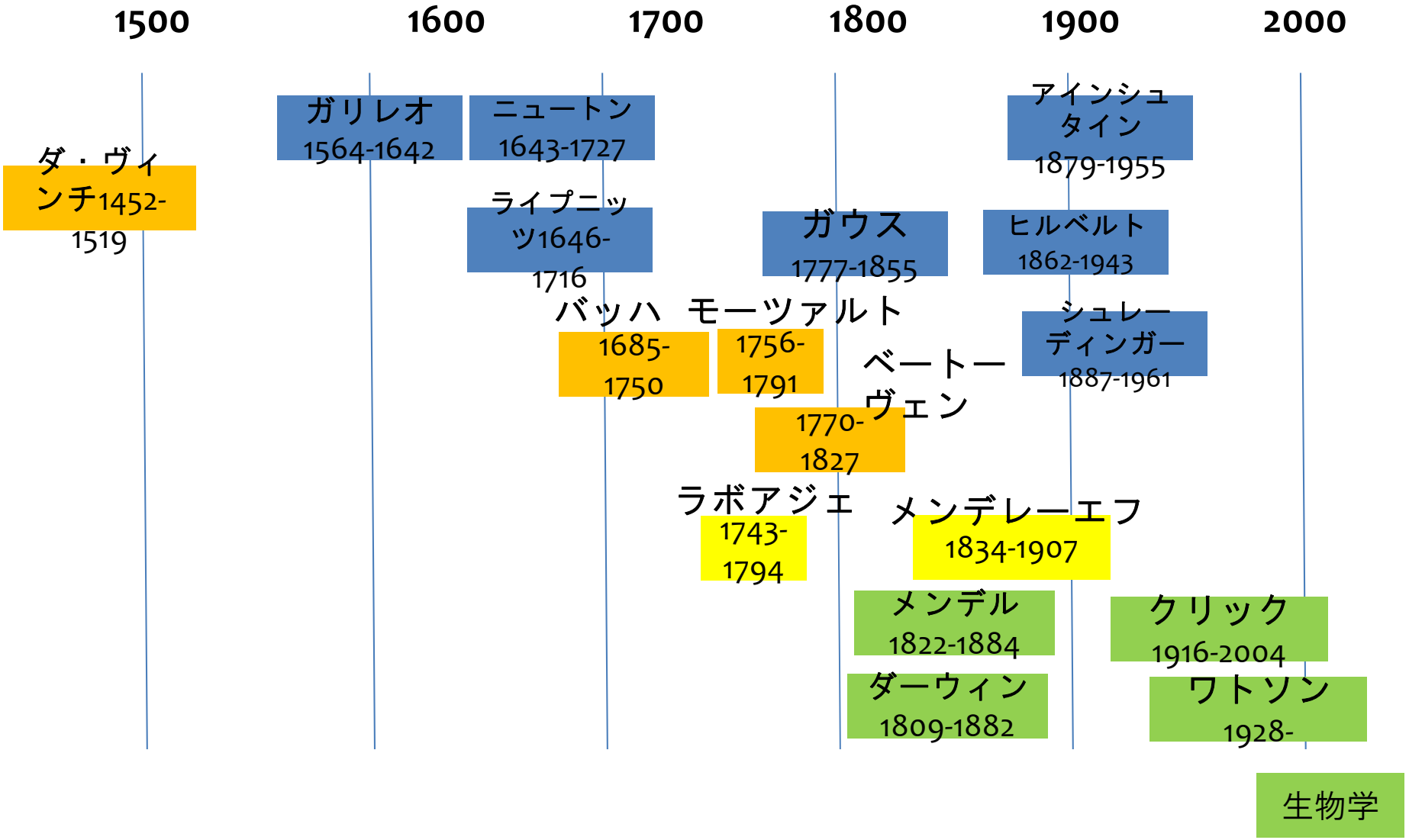
システム生物学概論

有田正規
国立遺伝学研究所
理化学研究所環境資源科学研究センター
arita@nig.ac.jp

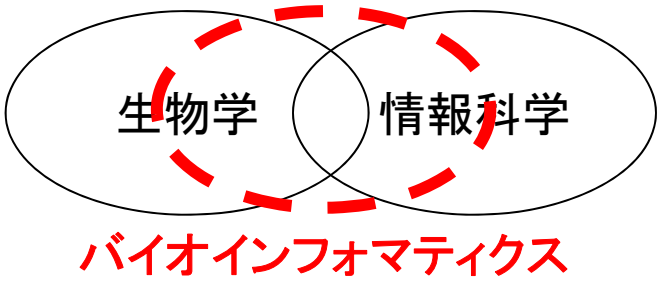
生物学は若い学問


1500
1600
1700
1800
1900
2000
ダ・ヴィンチ1452-1519
ガリレオ1564-1642
ニュートン1643-1727
ライプニッツ1646-1716
バッハ1685-1750
モーツァルト1756-1791
ベートーヴェン1770-1827
ラボアジエ1743-1794
ガウス1777-1855
メンデル1822-1884
ダーウィン1809-1882
メンデレーエフ1834-1907
アインシュタイン1879-1955
ヒルベルト1862-1943
シュレーディンガー1887-1961
クリック1916-2004
ワトソン1928-
生物学

バイオインフォマティクス

計算機科学（インフォマティクス）視点による
生物学（バイオロジー）

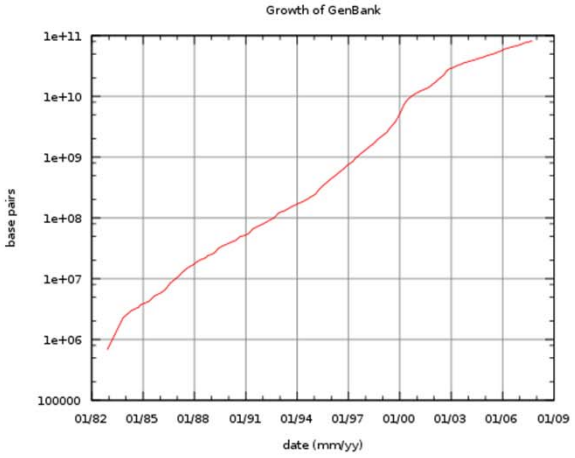
Google : Bioinformatics, Computational biology
日本語は、生物情報、生命情報、など。

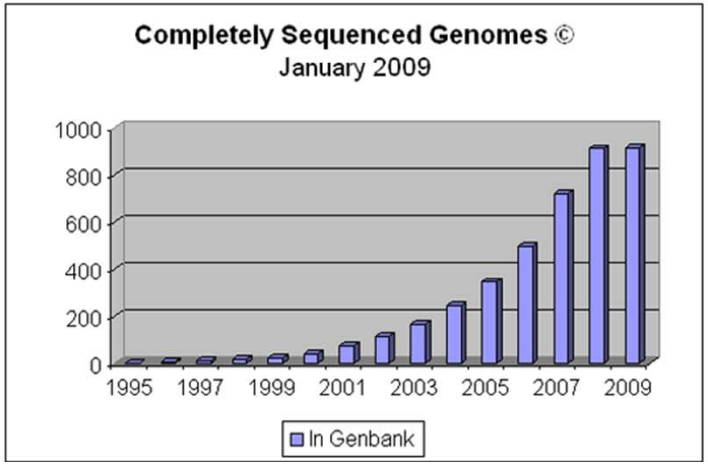


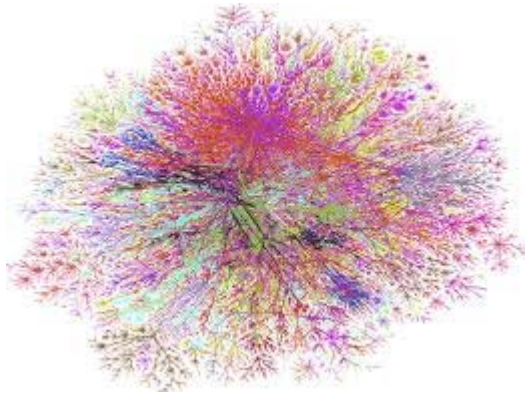
多くの天文学者は望遠鏡で星を見ていない。
生物学も、細胞や動物を見なくなる。

バイオインフォマティクスの歴史

1977	サンガー法の開発
1980	配列比較、相同性検索
1990	遺伝子配列予測、機械学習
2000	DNAマイクロアレイ、ネットワーク
2010	遺伝子多型、統計処理
2015	一細胞計測、遺伝子デザイン

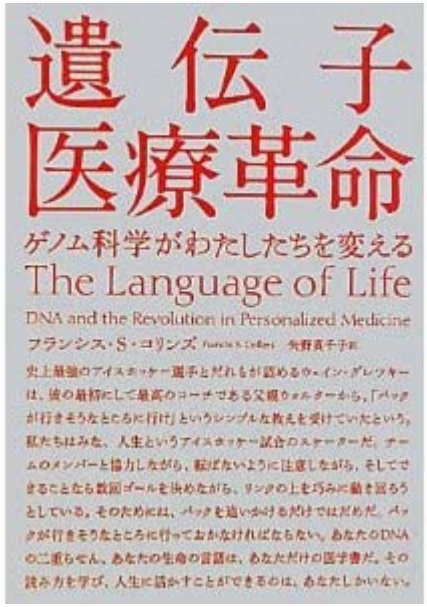


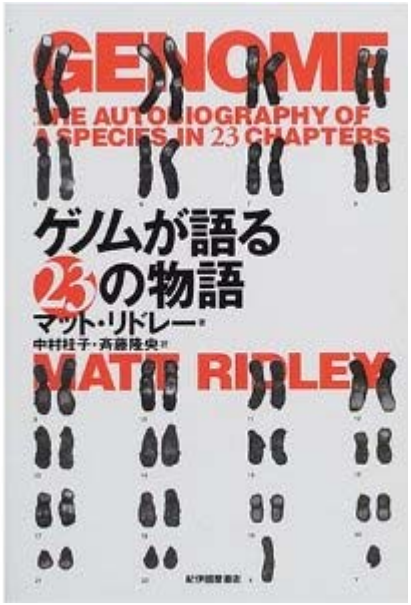


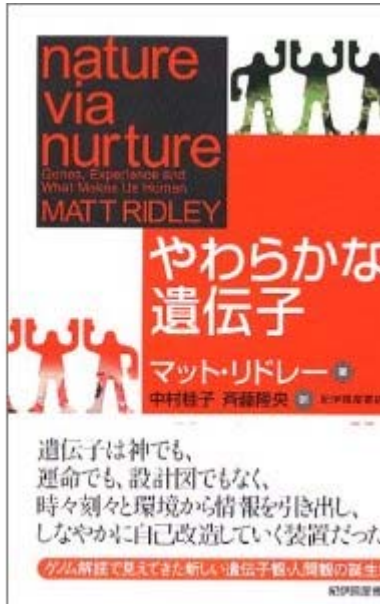


わかりやすい一般書

- Francis Collins (NIHの長官)
- Matt Ridley（もとNatureのエディター）





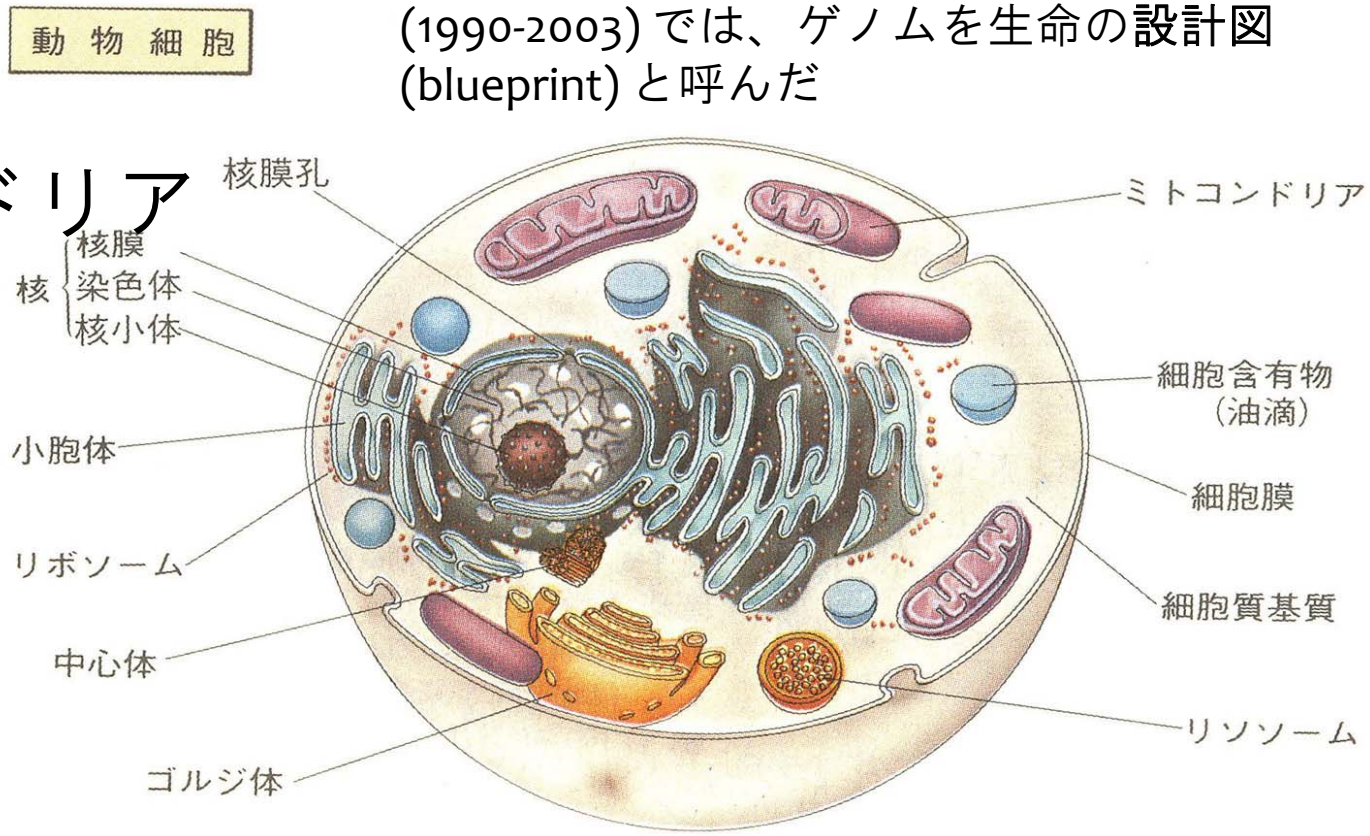


ゲノムって何？

細胞に含まれる、DNA配列全体のこと

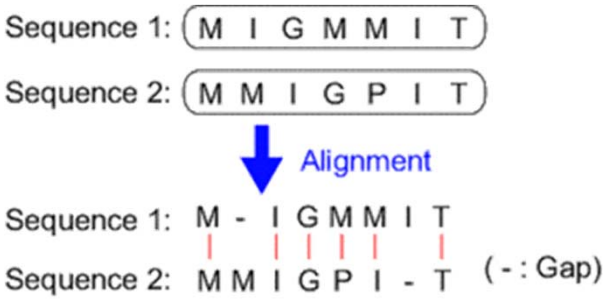
核染色体
ミトコンドリア
（葉緑体）

ワトソンが開始したヒトゲノム計画(1990-2003)では、ゲノムを生命の**設計図**(blueprint)と呼んだ



始まりは遺伝子配列

遺伝子情報の蓄積・比較に、計算機が利用されたのが始まり



1970 Dot-matrix method

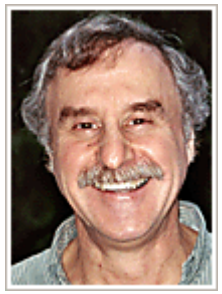
1978 PAM Matrix (M. Dayhoff)

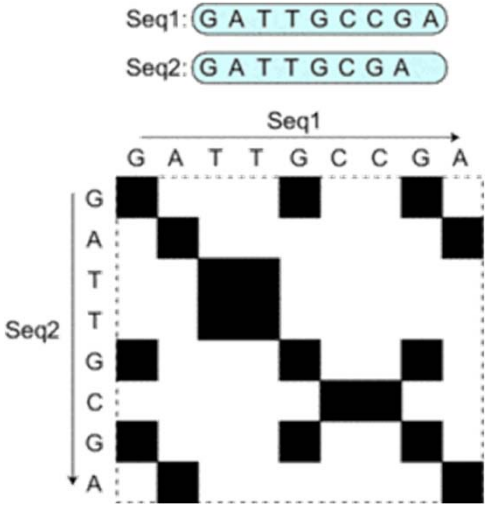
グロビンに基づく置換行列

1982 GenBank Database

1992 BLOSUM Matrix (S. Henikoff)

保存ペプチドに基づく置換行列





配列検索ツールの登場

Sequence Alignment … データベース中に類似配列を探すアルゴリズム

1985 FASTA (Pearson & Lipman)

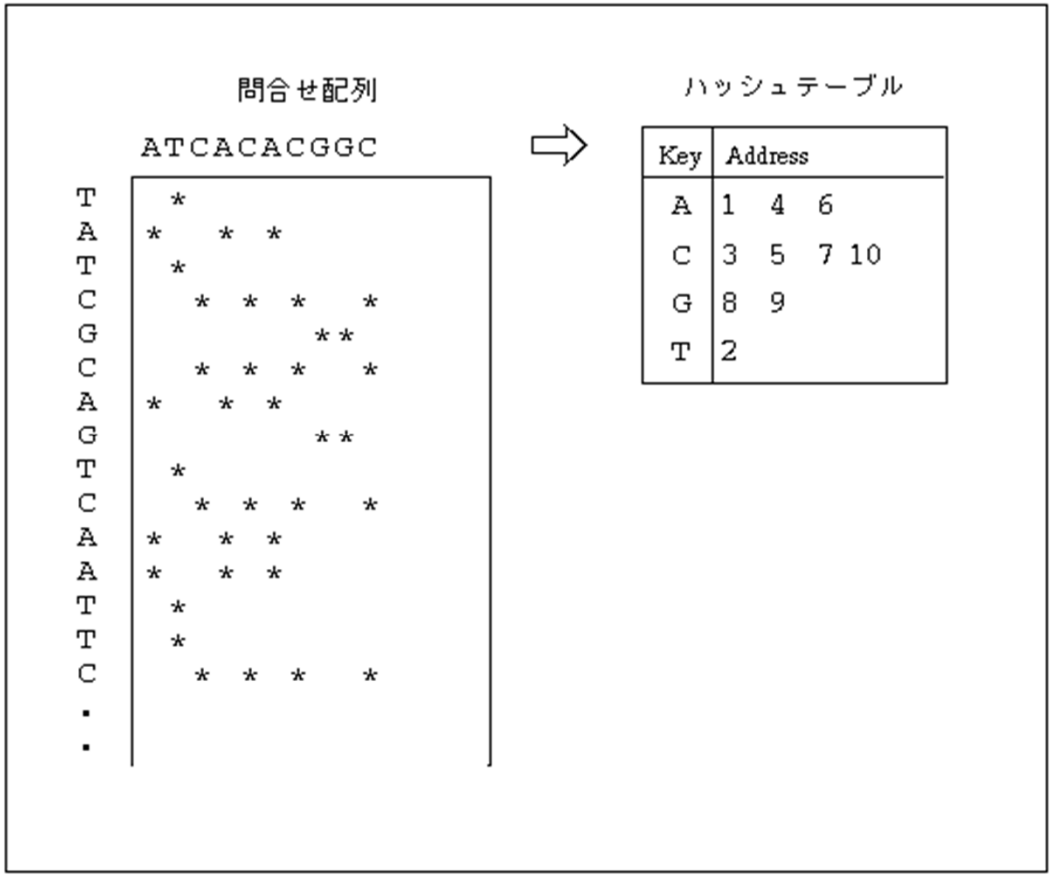
2アミノ酸のハッシュテーブルを用いた動的計画法

1990 BLAST (Altschul & Karlin)

3アミノ酸のシード一致を左右に伸長して類似配列を検索

FASTA (fast A)

動的計画法を用いるとき、問い合わせ配列に関するハッシュテーブルを作成して検索を高速化。

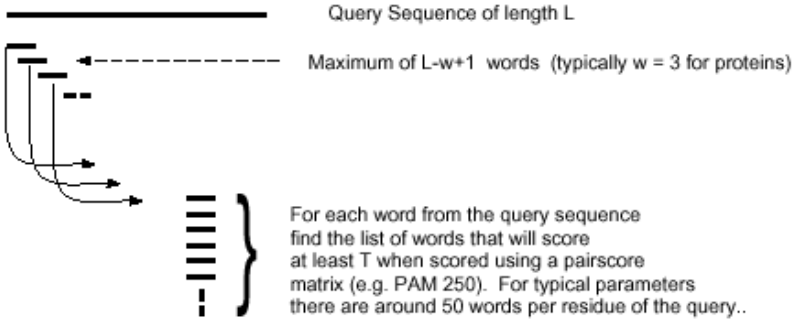


もともとは、問い合わせ配列の長さと、データベースのサイズをそれぞれ m, n とすると m x n に比例する計算時間が必要。

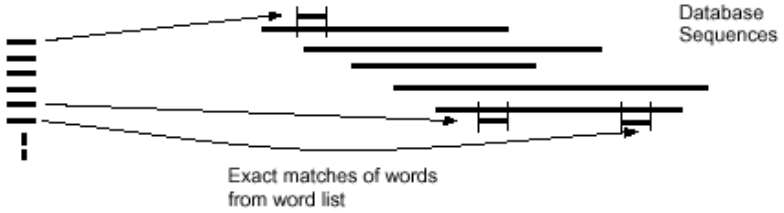
DNAなら6塩基、アミノ酸なら2残基毎のハッシュを作成する。(k-tuple)

BLAST Algorithm

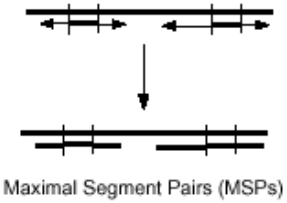
(1) For the query find the list of high scoring words of length w.



(2) Compare the word list to the database and identify exact matches.



(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.



BLAST

文字列検索の世界に、「確率アルゴリズム」を持ち込んだところが革命的

1. 問い合わせ配列からシードを作成
2. シードがデータベース中にでてくる箇所を同定
3. 一致したシードを左右に伸長
4. 伸長した中で一番良いものを選択

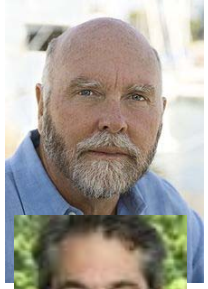
現在までに様々なオプションを開発。

ヒトゲノム計画

当時の研究者は否定的。しかしJames Watsonが強力に推進、国際コンソーシアム化

1990 J. Watsonが提案 30億ドル

1995 *H. influenzae* ゲノム決定 (C. Venter)



1998 Celera社登場 (C. Venter)

1999 *D. melanogaster* ゲノム決定(G. Myer)

2000 ヒトドラフトゲノム発表

2003 ヒトゲノム完成 (99.99%)

2009 HapMap完成 (Phase3)

2012 1000 genome発表

ヒト遺伝子数 2.2万

SNPは 300塩基毎

シーケンサーの進歩

1977	1992	2007	2010	now
Sanger法	キャピラリー	次世代	一分子	小型化
< 1 K / day	< 1 Mb / day	< 1 Gb / day	< 1 Tb / day	40Gb/day
		Illumina HiSeq	PacBio RSII	Nanopore Minion

ヒトゲノム計画

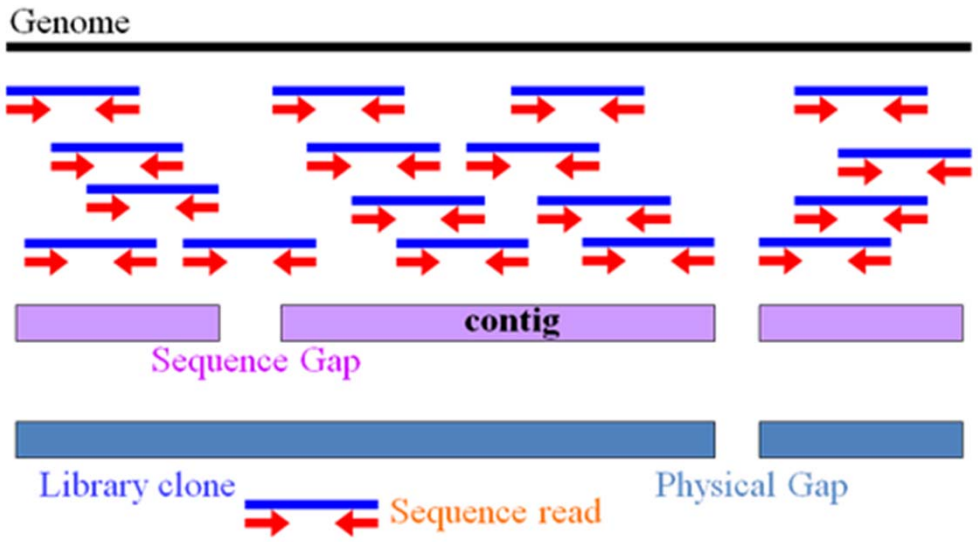
ABI Prism3130

メタゲノム

個人ゲノム

ショットガンシーケンス

ゲノムを細切れにして読み取ってから、計算機によって「再構築」



ドラフト配列を決めるにも、ゲノムサイズ x 10

標準的な次世代機は
30GBase x 8 レーン

DNAチップとの違い

	DNAチップ、アレイ	次世代シーケンサー
原理	DNAの相互作用 を蛍光標識	DNA合成を化学的に 標識、検出
解析対象	既知情報	未知も可
コスト	１チップ数万円	1サンプル10万円

ヒトゲノム計画の発展

パーソナルゲノムの時代に突入
2002 HapMap Project 開始 (現在1000人分)
2003 高精度配列決定

2006 遺伝子検査会社出現 (23andMe, deCODEme)
2010 一千植物ゲノム計画
2011 個人ゲノムの読み取りが50万円
2015 ロングリード長の次世代シーケンサー登場

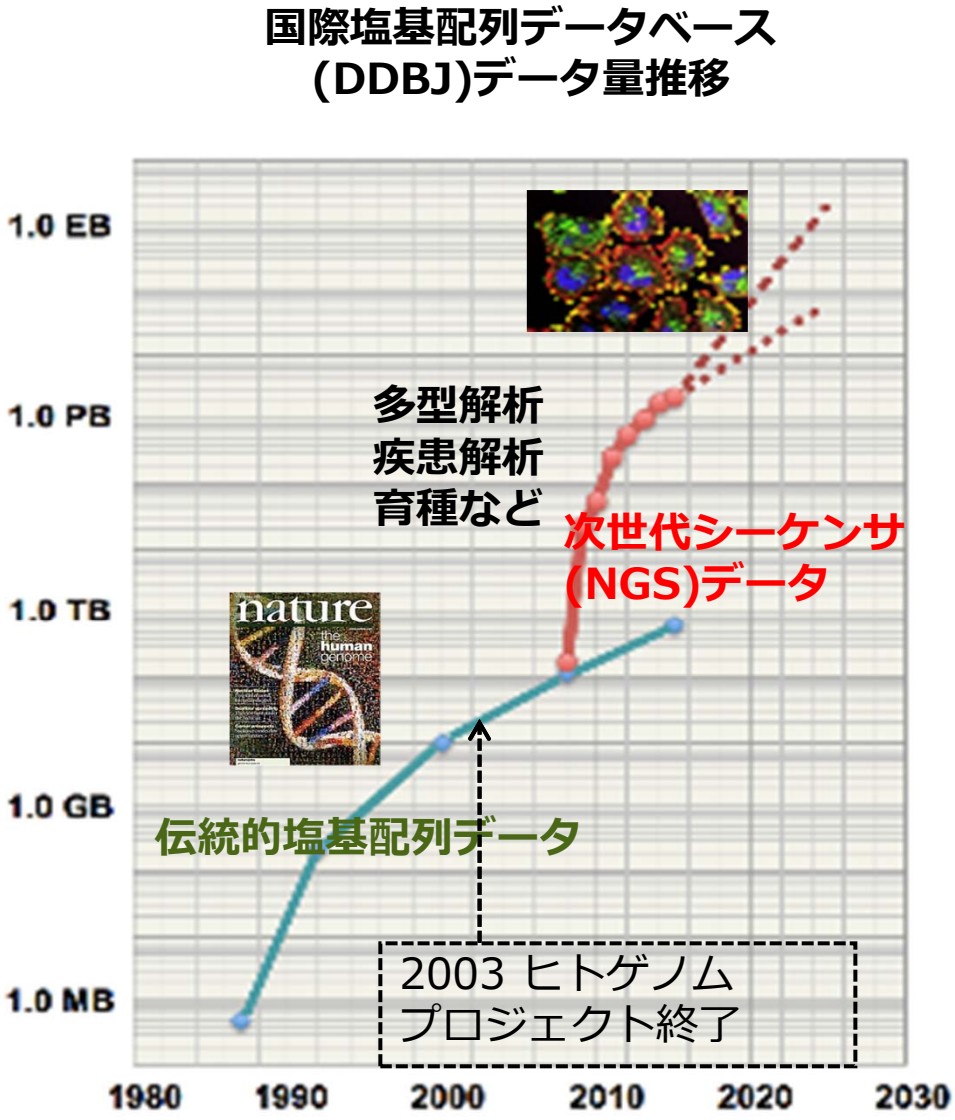
2018 一細胞ゲノム、一細胞RNA

研究者が扱うデータサイズ

2008年を契機に、データサイズがペタへ

個人の解析ではテラ
＝　データの移動が困難
＝　データの近くで解析
＝　スパコンの利用

PCで解析できる時代は、終わりつつある



今のバイオインフォマティクス

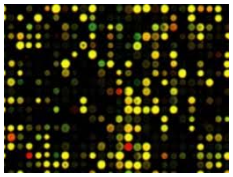
パイプラインの構築か、スパコン利用

- 比較ゲノム解析
  ゲノムアライメント、種レベルの解析
- 統計解析
  クラスタリング、機械学習、SNP解析
- シミュレーション

Omics生物学の台頭

チップやシーケンサーにより、遺伝子発現量を簡単に計測できる
（トランスクリプトーム）

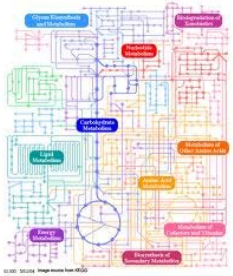




質量分析計の進歩により、発現するタンパク質や量を同定できる
（プロテオーム）



同じく質量分析計で、含まれる代謝物が計測できるようになってきた
（メタボローム）




様々な Ome, Omics

Ome (総体) + ics (学問)

Biome	生態系全体
Genome	遺伝子全体
Transcriptome	遺伝子産物全体
Proteome	タンパク質全体
Metabolome	代謝物全体
Phenome	表現型 (phenotype) 全体

さらなる広がり

- 一人一人の遺伝子情報が手に入る時代
  保険、医療、ライフスタイルに大きな変化が訪れる
  血液検査と同様にエピゲノム検査

- 環境ゲノムが手に入る時代
  生物多様性　＝　遺伝子資源
  自然との共存を目指したデザイン

社会基盤としての生命科学

幅広い分野を学ぶ必要性

- バイオインフォマティクスは、
  極めて早いペースで進展
  - バイオインフォマティクスは生物学を変えた
  - 未来の生物学はインフォマティクスが中心
  - 「環境」＋「生物学」の時代

- 地域に根差した学際的なアプローチが重要
  - ローカルである利点を生かす（生物多様性など）
  - データは膨大で、地理関係はやはり重要
  - スパコンやデータセンターの活用

最近のトピック

- **パーソナルゲノム**

  SNP（一塩基多系）を探すプロジェクトで日本人の「違い」を配列決定。配列のバリエーションを知り、生活習慣病などに応用

- **メタゲノム**

  単離できない微生物を含め、あらゆるゲノムをショットガン法で読みまくる。生物多様性や新しい遺伝子資源の探索

- **エピゲノム**

  DNA配列のメチル化、ヒストン修飾など、全ての後天的変異を解析する。難病の解析、遺伝と環境の関係

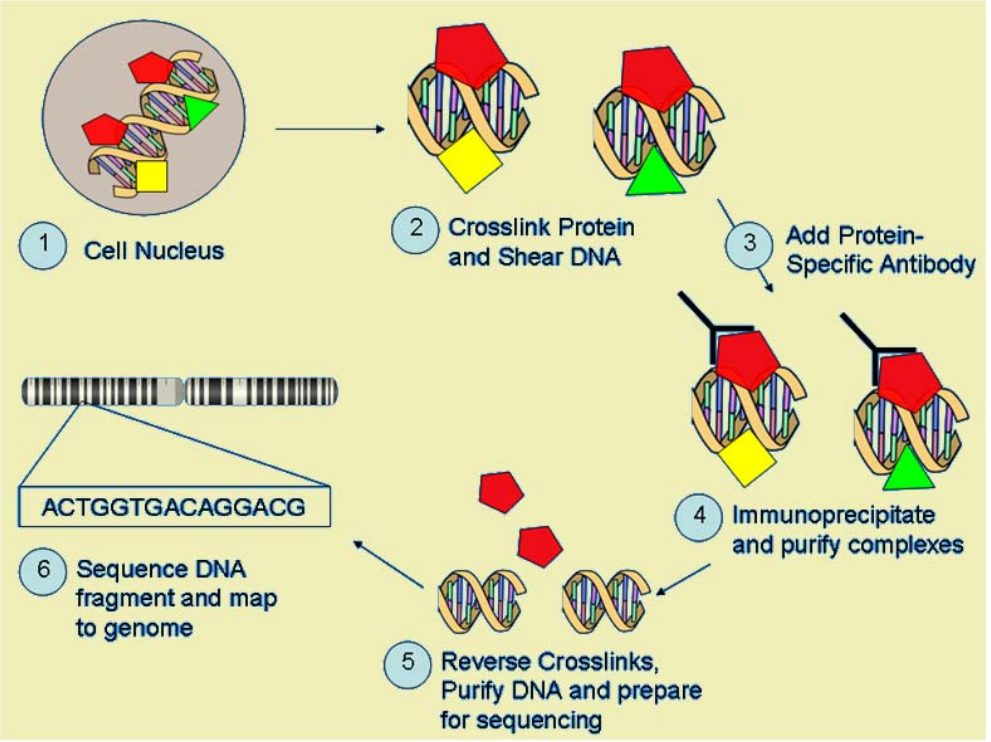
ただし、片方が多発性硬化症の一卵性双子で、３６０万SNP, エピゲノム、マイクロアレイでも差は検出できない。*Nature* 2010, 464(7293):1351-6

DNAメチル化、エピゲノム

ゲノムは単なるATGCではなく、様々な修飾を受ける。
タンパク質も、様々に修飾されてから機能する。
→　配列からは判らない機能が殆んど

- 個人差の解明
- 遺伝するエピゲノム情報

多くのコホート研究
日本では、東北メガバンク



メタゲノム1

腸内メタゲノム

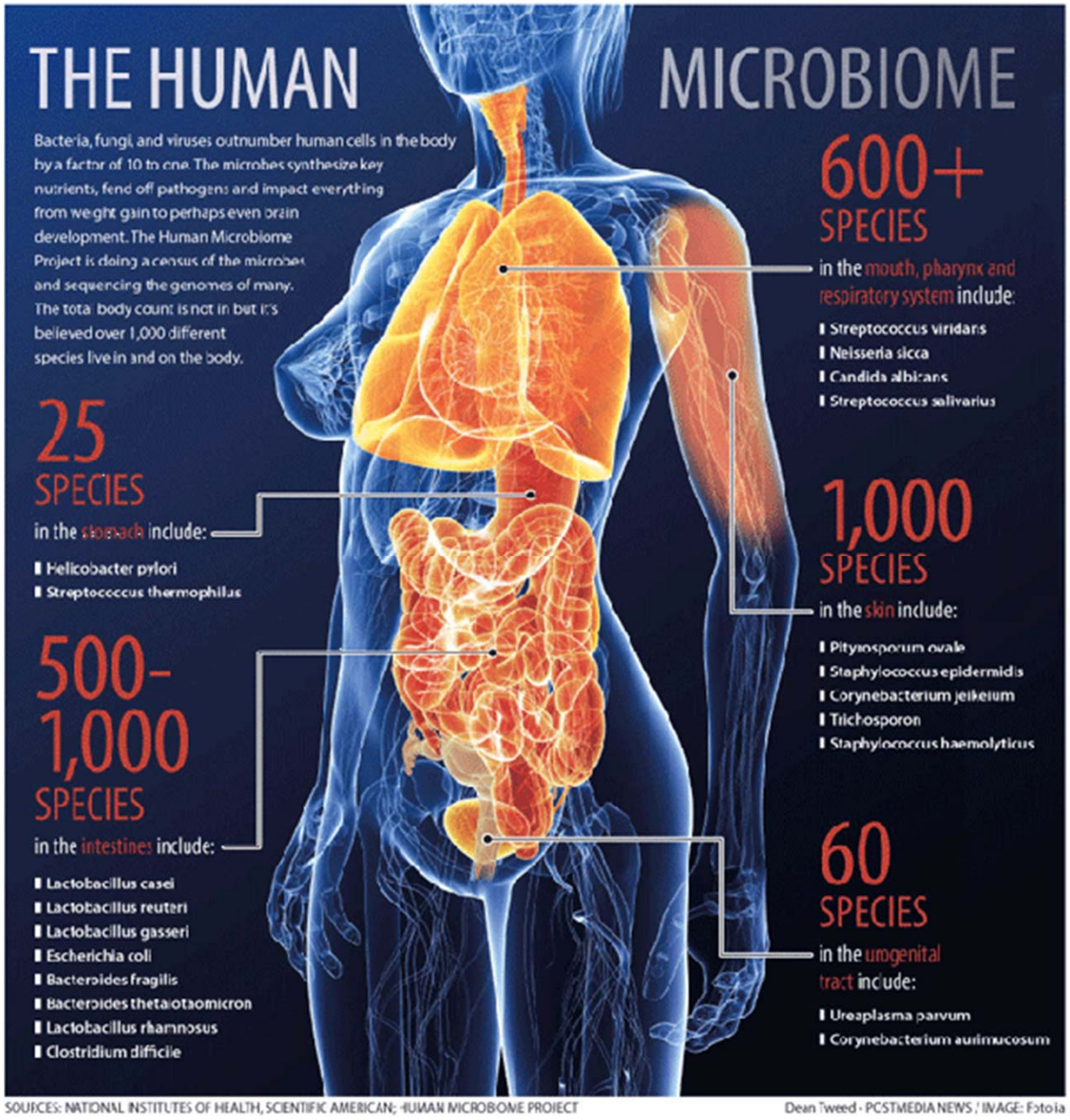
Human Biome

個々人に特有の微生物プロフィール

（親子も違う）

体にはアーキア、病原菌を含む様々な菌が常在

食事とゲノム



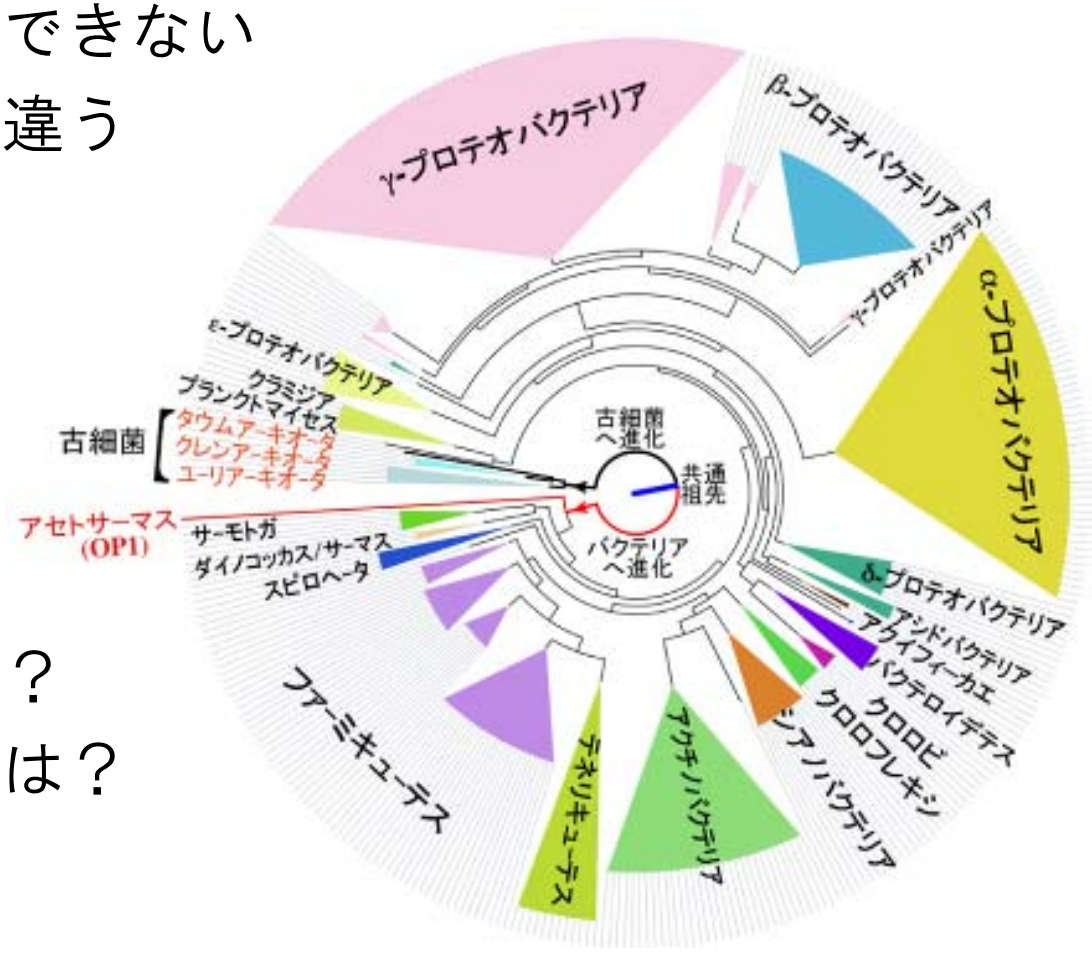
メタゲノム２

環境メタゲノム、バイオリソース

微生物のほとんどは培養できない

土の違いで野菜の育ちは違う

環境とゲノム



名前のある微生物は？

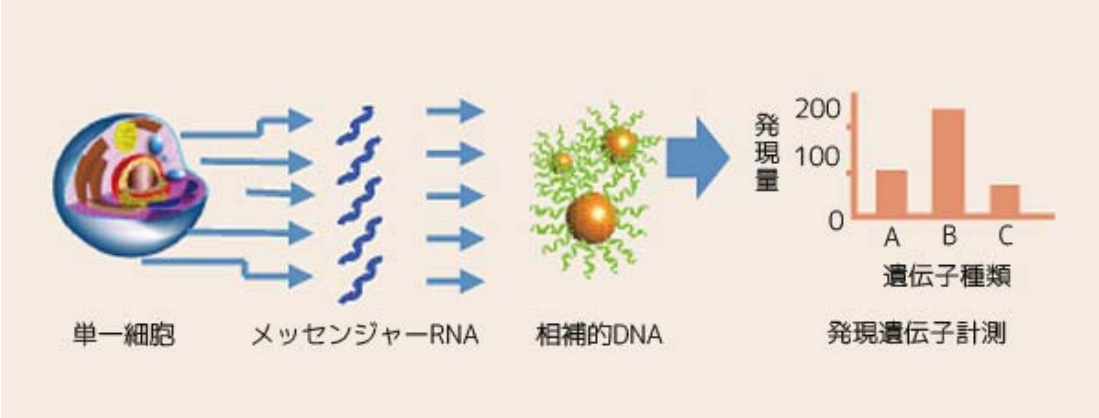
一番炭素量の多い生物は？

真空でも生きられる生物は？

一細胞解析

マイクロデバイス

個々の細胞を選んで
DNAを増幅、
シーケンスする技術



わかったこと：

個々の細胞の突然変異率、
個々の精子のDNA変異は
25程度

公開のヒトゲノムデータ

1000 GENOMES　http://www.internationalgenome.org/

日本人105人を含む2500人分のゲノムデータ (300TB) を公開

ヒトゲノムは驚くほど均質

遺伝的多様性は低く、有効集団サイズ１万。

（トバ・カタストロフ理論によると、７万年前の火山噴火が原因。これを生き延びたのが、ネアンデルタール、デニソワ、ヒト）

有効集団サイズ：

　遺伝学におけるモデルで考えた際の指標

有効集団サイズの観点では、猿やネズミのほうが多様。

例えばネズミは、数十万年前に日本とヨーロッパで分岐。

遺伝子疾患の区分け

単一遺伝子疾患

- 希少疾患の8割は遺伝性
- 5000 以上ある
- 1万人に1人の劣性（潜性）遺伝疾患の場合、保因者は百人に一人

- 自分が無関係の可能性はゼロ

原因が不明なもの

- 統合失調症は、遺伝的要因が7割（発症率1%）
- 生活習慣病も、おそらく　遺伝的要因が高い

数千人規模の研究ではわからない

UK BioBank 500K cohort

Tohoku MB 150K cohort

ハプロタイプ解析

一人ひとりのゲノムは数百～千塩基毎に変異がある。
古い変異は、一定の集団で共有される。
（Single Nucleotide Polymorphism）

領域にわたる SNP のパターンが
「ハプロタイプ」（集団マーカー）

染色体が二本ずつあり、
親から子に染色体が受け継がれる
（減数分裂の）際に、組換えが
おこるため、推定は難しい
（EM法による）

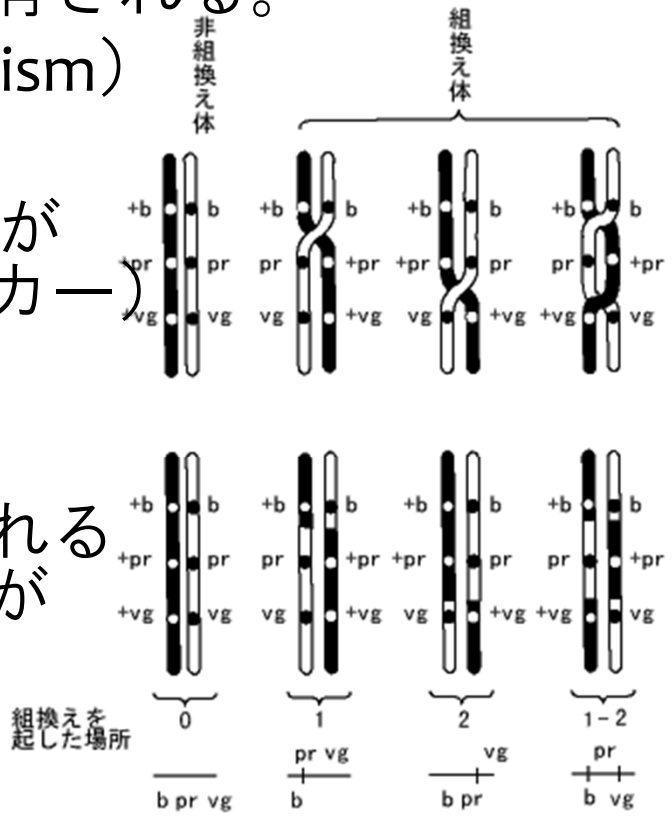


図1･9　相同染色体の組換えによって，三つの遺伝子の配列順序をきめる実験（駒井 編「ショウジョウバエの遺伝と実験」千野 著，1952）

まとめ

- 生命科学は、細分化の一途を辿っている
- ゲノムが生命の設計図のはずだった
- 今はとにかく大量データの取得
  メタゲノム、エピゲノム、１細胞ゲノム

科学の観点から、何を導けるのか。