

ゲノム情報学

Jeffrey Fawcett

RIKEN iTHEMS (jeffrey.fawcett@riken.jp)

June 2018

関西	関東
めばちこ	ものもらい
おとつい	おととい
なおす	しまう
ほかす	捨てる

関西	関東
めばちこ	ものもらい
おとつい	おととい
なおす	しまう
ほかす	捨てる

イギリス	アメリカ
chips	french fries
crisps	chips
colour	color
neighbour	neighbor
analyse	analyze
rubber	eraser

生命の暗号



- 生物の様々な形質が、ある「言語」によって書かれている
- A, T, C, G の4つの「文字」(塩基) から成り、DNA に記されている
- ほぼ全ての生物がこの「言語」で書かれた「本」(ゲノム) を持つ (ヒト: 33億文字)
- 個体や生物によって「本」の中身(「文字」の並び) が異なる

生命の暗号

- 多くの生物の「本」の全ての「文字」がわかってきている
- まだほとんど意味がわかっておらず、解明には程遠い
- この「本」(ゲノム)には何が書かれているの？
- どうやってこの「言語」を読み解くことができるの？

今回の講義の内容

- ゲノムを解明するためのアプローチとは？
- 遺伝学の歴史
- ゲノムとは？
- 遺伝子とは？
- ゲノムにはどんな情報が記されているの？

理論との出会い

- 京都大学農学部（学部・修士）、分子生物学の実験
- 理論家の印南さん（現総研大准教授、当時アメリカ）に出会う
 - ▶ 印南さん「実験嫌いやろ？こっちの水は甘いぞ」
 - ▶ 生物が苦手な同僚「ついていきます」→アメリカへ留学
 - ▶ 印南さん「お前はどうか？実験好きなんか？」
 - ▶ 私「数学やるぐらいなら実験の方が…」
- データへのアプローチの仕方に感銘を受ける
- 下手に実験するより既にあるデータを活用する方法を考える方が効率がいいのでは？
- 博士課程で Bioinformatics の研究室に (Ghent Uni, Belgium)

ゲノムを読み解くための理論とは？

- 実験・観測：ある特定の条件下で行う
- 理論：様々な異なる条件下での結果を予測するための枠組み

ゲノムを読み解くための理論とは？

- 実験・観測：ある特定の条件下で行う
 - ▶ ある個体・種である遺伝子・領域の機能を調べる
- 理論：様々な異なる条件下での結果を予測するための枠組み
 - ▶ 他の個体・種でどんな結果が得られるか
 - ▶ ゲノムの違いにより機能がどう変わるのか
 - ▶ **ゲノムの進化**：現在のデータがどんなプロセスで生じたのか？

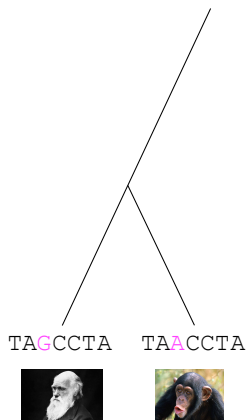
似ている＝同じものに由来する

関西	関東
めばちこ	ものもらい
おとつい	おととい
なおす	しまう

イギリス	アメリカ	オランダ
colour	color	kleur
analyse	analyze	analyseren
Sunday	Sunday	Zondag
Monday	Monday	Maandag

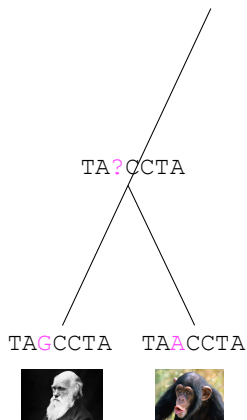
- 元々同じだったのが、片方（もしくは両方）で変化が生じた

似ている＝同じものに由来する



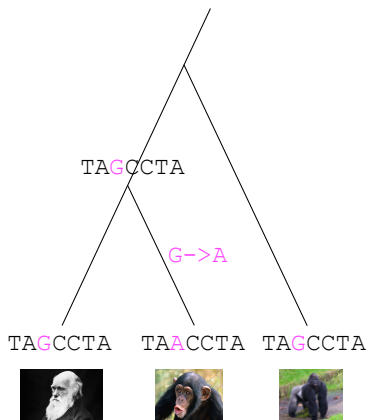
- 過去の状態や、過去に生じた変化が推定できる

似ている＝同じものに由来する



- 過去の状態や、過去に生じた変化が推定できる

似ている＝同じものに由来する



- 過去の状態や、過去に生じた変化が推定できる

似ている度合いと分かれてからの時間が比例

イギリス	アメリカ	オランダ
colour	color	kleur
Sunday	Sunday	Zondag
Monday	Monday	Maandag

- 時間とともに違いを蓄積していく

ヒト同士の違い : 0.1-0.4%, ヒトとチンパンジーの違い : 1-2%

- 似ている度合いからいつ分かれたかを推定できる

変わってもいい部分とまずい部分とがある

イギリス	アメリカ	オランダ
colour	color	kleur
Sunday	Sunday	Zondag
Monday	Monday	Maandag

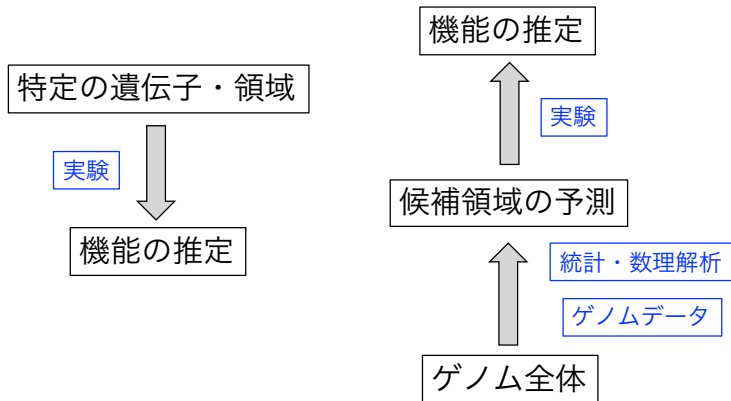
- 変わると重要な機能が失われる領域は、変わっても広まらない（子孫を残せないため）
- このような領域は遠く離れた種同士でも似ている
- 遠い種同士で非常に似ている = 重要な機能を持つのでは？

似ている度合いから歴史がわかる

日本	イギリス	アメリカ	オランダ
コンピューター	computer	computer	computer

- 各言語の成立以前に生じ、ずっと維持されてきた !?
 - ▶ これだけ似ているとその可能性は低い
- 最近急速に（違いがたまる間もなく）広まった
- 最近一気にすごく重要になった言葉 !?
- 子孫を残す確率を高める変化はすぐに広まる

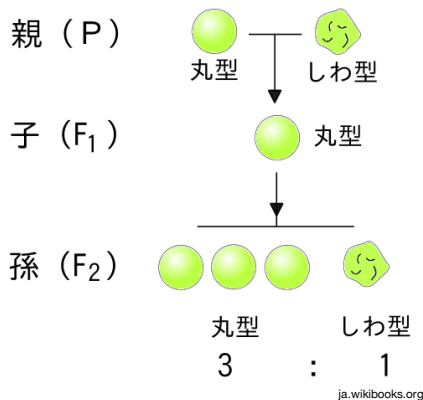
ゲノムを読み解くためのアプローチ



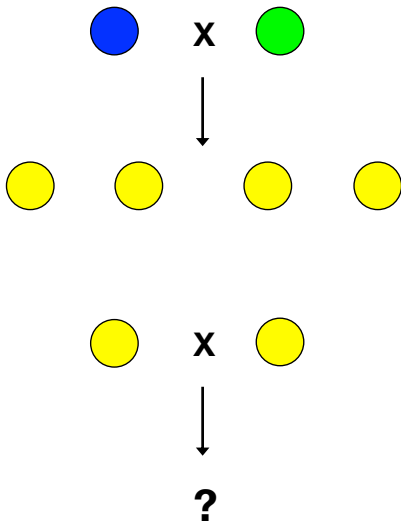
- 理論に基づいた bottom-up なアプローチがスタンダードに

遺伝学の歴史：メンデル遺伝

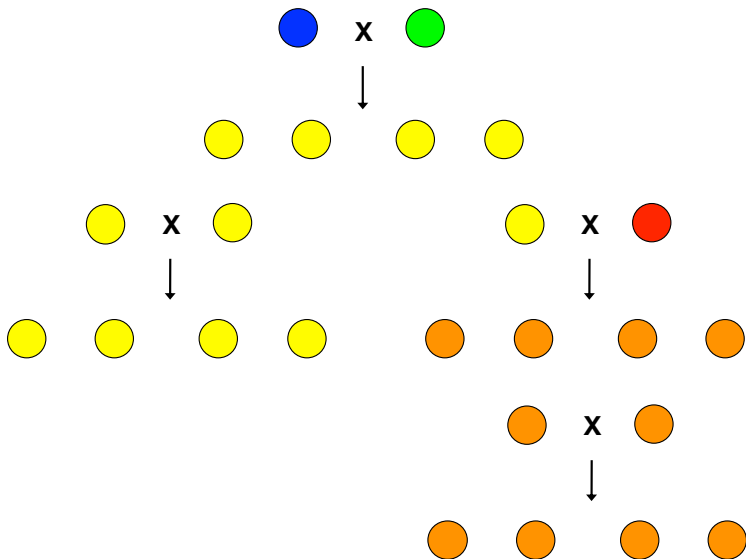
- なぜ親子は似てるの？ 親の形質を決める情報が子に伝わる
- 何がどうやって伝わるの？



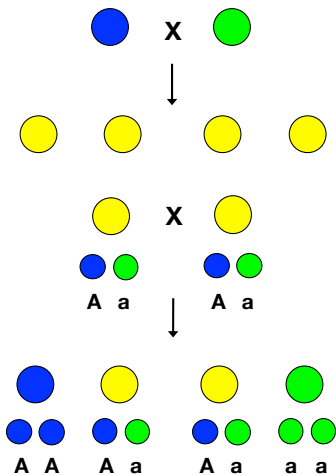
メンデル遺伝



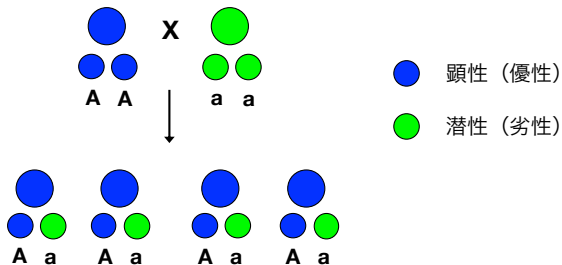
メンデル遺伝：仮説 1



メンデル遺伝：仮説 2



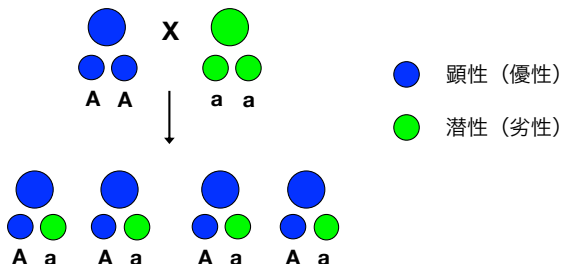
メンデル遺伝



メンデル遺伝の重要性

- 代を経ても変わらない、ずっと追跡できる遺伝物質がある
（「血」や液体のようなものが混じり合うのではない）
- 確率論・統計学をベースとした学問へ (e.g. Fisher, Pearson)

Hardy-Weinberg の法則



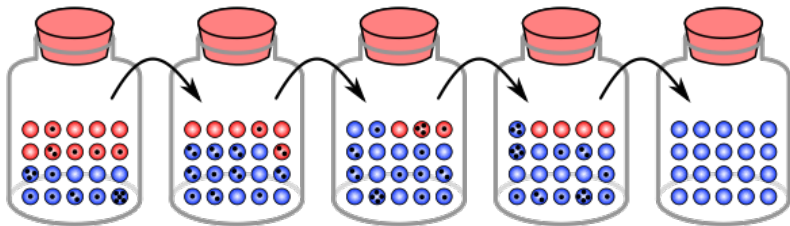
- メンデル遺伝が正しければ、A が顕性、a が潜性の場合、A の頻度が上昇するのでは？
- W. Hardy: いくつかの条件を満たす場合、世代を重ねても A と a の頻度（つまり AA, Aa, aa も）は変わらない（平衡状態）

集団遺伝学

- どのような時に A や a の頻度が変わるのか？
- どのような時に A や a が固定 (全個体が AA か aa になる) するのか？
- 交配がランダムでない (e.g. AA 同士、aa 同士が交配しやすい)
- a が子孫を残す確率を上げる (下げる) 場合、a の頻度が上がる (下がる) (自然選択)
- 無作為抽出の効果で a の頻度が変わる (特に集団サイズが小さい場合) (遺伝的浮動)

分子進化の中立説（木村資生）

- ランダムなプロセス（遺伝的浮動）が進化において非常に重要
- 進化の大部分が遺伝的浮動で説明できる !?

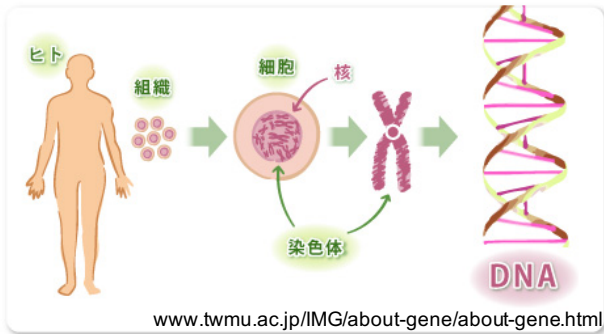


en.wikipedia.org

- 中立進化 vs 自然選択

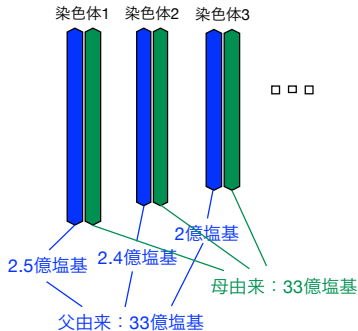
ゲノムとは？

- 遺伝情報は DNA に記されている
- ゲノム：1 個体の全遺伝情報の 1 セット
- 全ての細胞が同じ遺伝情報を持つ



ヒトのゲノム

- 父由来・母由来のゲノムを1セットずつ持つ
- 23対の染色体、常染色体22対と性染色体1対 (XX or XY)
- 各染色体の合計が約33億塩基



生物の「複雑さ」は何が決めているの？

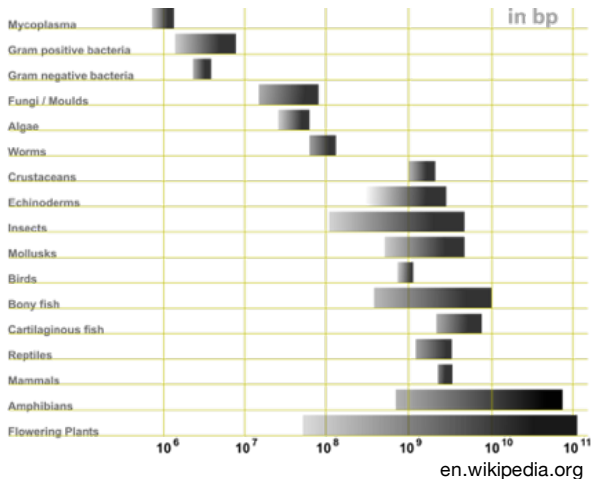
- 「複雑な」生物ほどゲノムの「文字」数が多い？

ゲノムの「文字」数は種によって大きく異なる

種	塩基数
出芽酵母	0.121 億
ショウジョウバエ	1.75 億
フグ	3.9 億
イネ	4.5 億
マウス	27 億
ヒト	33 億
タマネギ	150 億
バッタ	650 億
肺魚	1300 億
キヌガサソウ	1500 億

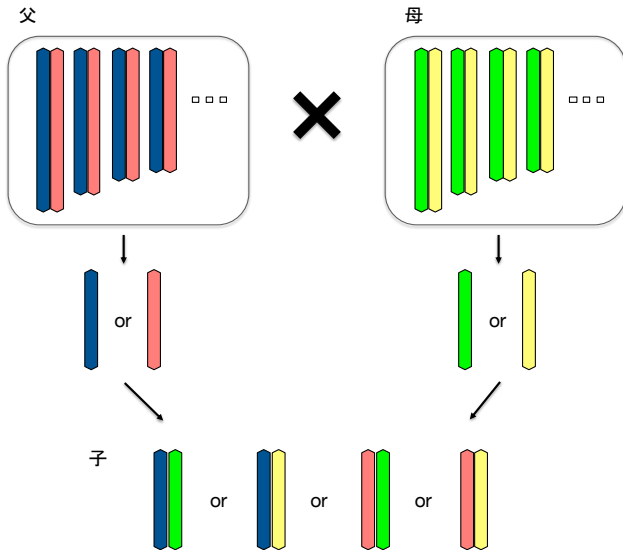
- 塩基数と生物の複雑さはあまり関係がない (C-value paradox)

ゲノムサイズの分布

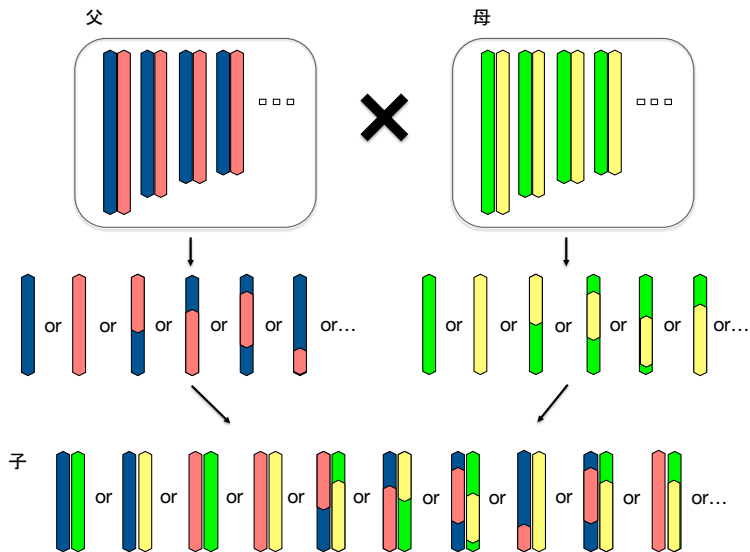


- 哺乳類はゲノムサイズのばらつきが非常に小さい
- 植物はゲノムサイズのばらつきが非常に大きい

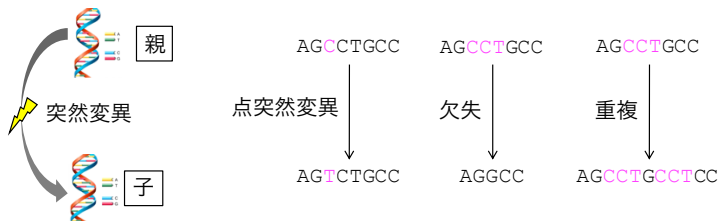
なぜ兄妹は違うの？



なぜ兄妹は違うの？

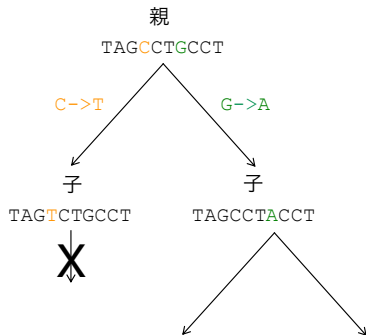


なぜ種や個体によってゲノムが異なるの？



- 親から子に伝わるたびに遺伝情報が少し変わる
(ヒトでは世代あたり 100-200 の突然変異が生じる)

なぜ種や個体によってゲノムが異なるの？

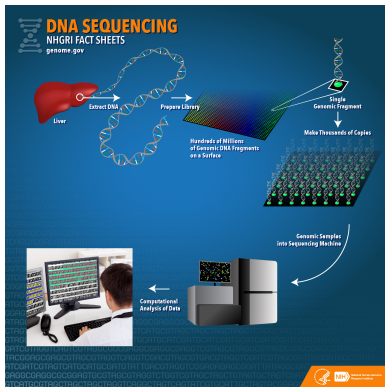


- 突然変異が広まる場合もあれば、広まらない場合もある

環境への適応と進化

- 新しい環境に適した有利な突然変異が起きる？ No!!
- 突然変異はランダムに起きる
- 偶然生じた多様性の中から、有利な突然変異を持つ個体が子孫をより多く残すことによって（あるいは遺伝的浮動によって）、その突然変異の頻度が上昇する
- 環境への適応は個体レベルではなく、集団レベルで起きる

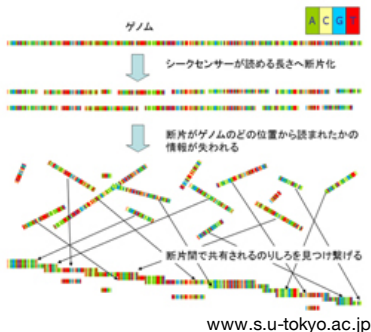
どうやってゲノムの文字の並びを決めるの？



www.genome.gov

- DNA を抽出 → 装置に投入 → コンピュータで解析

ゲノムのアセンブリ



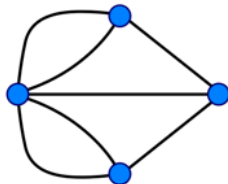
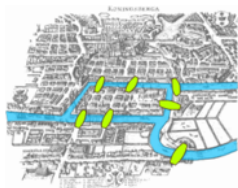
- 問題：一度に読める DNA の長さに限りがある（数百塩基）
- DNA を切断し細かく断片化したものを大量に生成
- 各断片の塩基を決定し、それぞれをつなぎ合わせる

ゲノムのアセンブリの難点

- データが膨大
- ある程度エラーが生じる
- ゲノム中に似たような「反復配列」が大量にある
- 各個体が微妙に異なるゲノムを2セット（以上）持つ

グラフ理論

- 「7つの橋を2度渡らずに、全て渡って、元の所に帰ってくる
ことができるか」



en.wikipedia.org

- 「一筆書きが可能か」
- オイラー (1736) 「無理」
- グラフ上の全ての辺を一度ずつ通る経路: Eulerian path

ゲノムアセンブリとグラフ理論

- 当初のアルゴリズム

- ▶ グラフ上の全ての頂点を一度ずつ通る経路を探す (Hamiltonian path)
- ▶ 非常に難解な問題

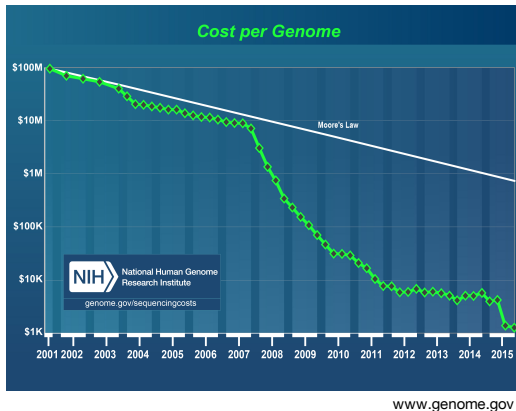
- アルゴリズムの改良

- ▶ 各断片をさらに断片化 (解析上)
- ▶ グラフ上の全ての辺を一度ずつ通る経路を探す (Eulerian path)
- ▶ ほとんどの解読されたゲノムが不完全で断片化されている
- ▶ まだまだ改良が必要

ヒトゲノムプロジェクト

- 1990 年：国際コンソーシアム発足、多額の公的資金投入
- 1998 年：Craig Venter, Celera Genomics 社の参入
 - ▶ ショットガン・シーケンシング法の開発
 - ▶ 生物学の問題ではなく情報処理の問題と認識
 - ▶ コンピュータ等の設備、情報科学の研究者に投資
- 2001 年：双方が全ゲノム配列決定の論文を発表

配列決定技術の革新的な進歩



- 「次世代シーケンシング」、コストの大幅減

配列決定技術の革新的な進歩

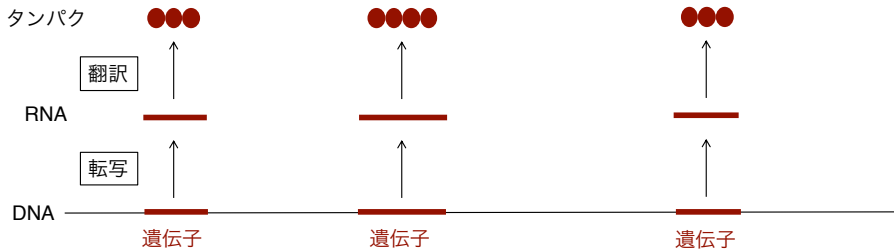
- 限られたモデル生物、1種1個体 → 数千～数万種、各種数百～数千個体の全ゲノム情報
- モデル生物中心の理解から生物の多様性を組み込んだ理解へ
- 「ビッグデータ」時代、情報科学がより重要に
- 新しい理論の構築が可能・必要
- 実験・データ・理論が相互作用しながら進歩

ゲノムを決定した後どうするの？

- 配列を決定しただけではあまり意味がない（ただのデータ）
- データに意味のある情報を「注釈付け」（アノテーション）する必要がある
- 一番欲しい情報： **どこにどんな遺伝子があるか**

遺伝子とは

- 機能を持つ分子（タンパク or RNA）を生成するゲノム領域



DNA からタンパクへ

- セントラルドグマ : DNA → RNA → タンパク (アミノ酸)
- DNA: A,G,T,C, RNA: A,U,T,C, アミノ酸 : 20 種類
- DNA/RNA の 3 塩基の組み合わせ (64 通り) が一つのアミノ酸に対応 (冗長性)

		Second Letter				
		T	C	A	G	
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } Ser TCC } TCA } TCG }	TAT } Tyr TAC } TAA Stop TAG Stop	TGT } Cys TGC } TGA Stop TGG Trp	T C A G
	C	CTT } Leu CTC } CTA } CTG }	CCT } Pro CCC } CCA } CCG }	CAT } His CAC } CAA } Gln CAG }	CGT } Arg CGC } CGA } CGG }	T C A G
	A	ATT } Ile ATC } ATA } ATG Met	ACT } Thr ACC } ACA } ACG }	AAT } Asn AAC } AAA } Lys AAG }	AGT } Ser AGC } AGA } Arg AGG }	T C A G
	G	GTT } Val GTC } GTA } GTG }	GCT } Ala GCC } GCA } GCG }	GAT } Asp GAC } GAA } Glu GAG }	GGT } Gly GGC } GGA } GGG }	T C A G

genetics-notes.wikispaces.com

生物の「複雑さ」は何が決めているの？

- 「複雑な」生物ほどゲノムの塩基数が多い？ ×
- 「複雑な」生物ほど遺伝子数が多い？

遺伝子の数

- ヒトは遺伝子がものすごくたくさんあるに違いない！
- ヒトゲノムプロジェクトの開始当初の予想は >10 万
- どんどん下方修正、現時点では約 21,000

遺伝子の数の比較

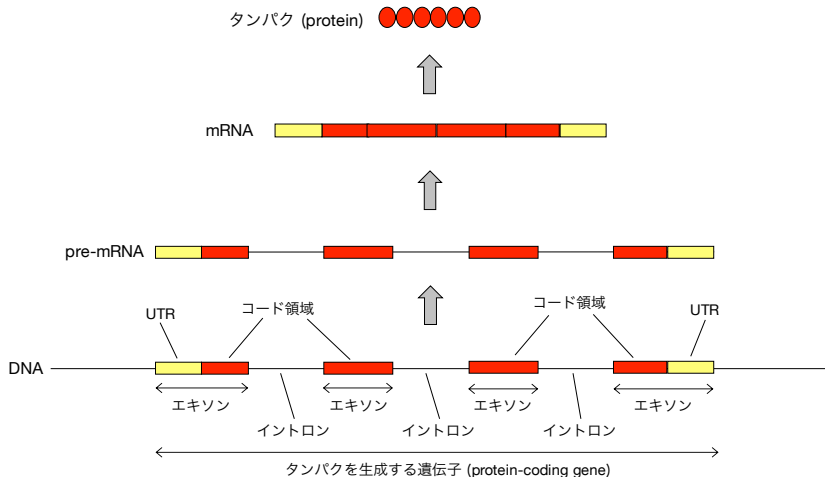
種	塩基数	遺伝子数
出芽酵母	0.121 億	6,000
ショウジョウバエ	1.75 億	17,000
フグ	3.9 億	28,000
イネ	4.5 億	40,000
マウス	27 億	23,000
ヒト	33 億	21,000
タマネギ	150 億	?
バッタ	650 億	?
肺魚	1300 億	?
キヌガサソウ	1500 億	?

- 遺伝子数も複雑さとあまりリンクしない
- 遺伝子数とゲノムサイズはあまり相関がない

生物の「複雑さ」は何が決めているの？

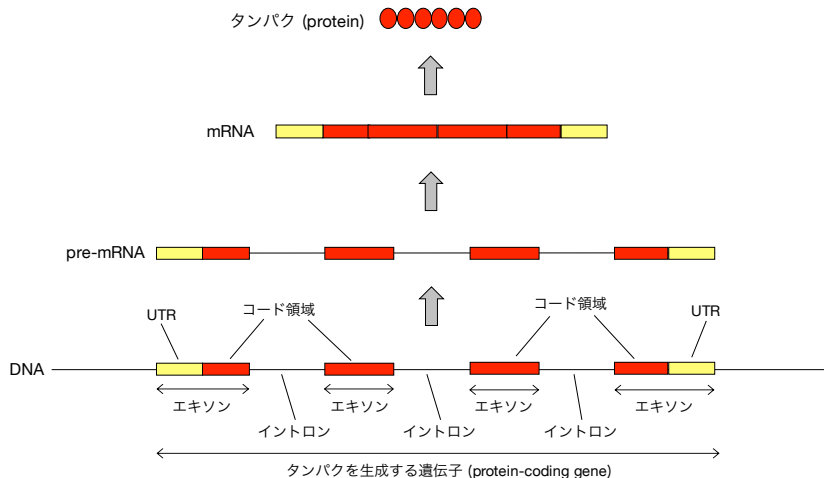
- 「複雑な」生物ほどゲノムの塩基数が多い？ ×
- 「複雑な」生物ほど遺伝子数が多い？ ×

遺伝子の構造



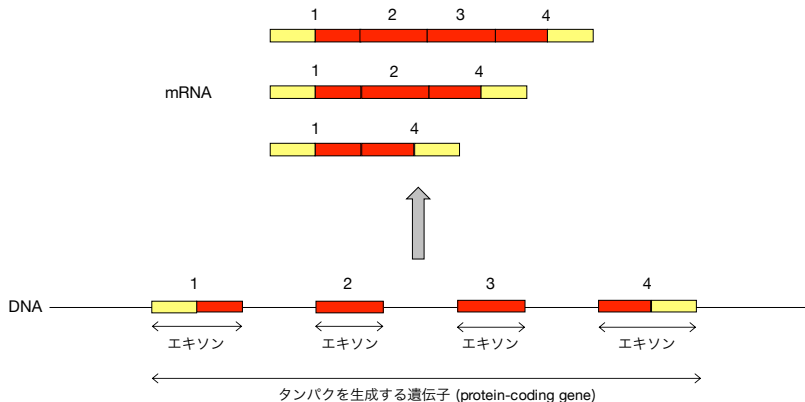
- 転写の際にイントロンが切り出される（スプライシング）

遺伝子の構造



- コード領域：ヒトゲノムの <2%

選択的スプライシング (Alternative splicing)

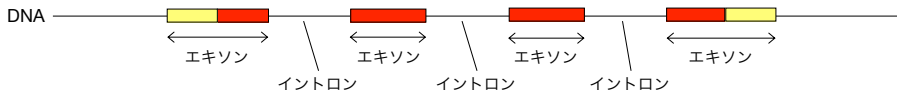


- 一つの遺伝子から複数の mRNA・タンパクが生成される
- ヒト・哺乳類で特に顕著

生物の「複雑さ」は何が決めているの？

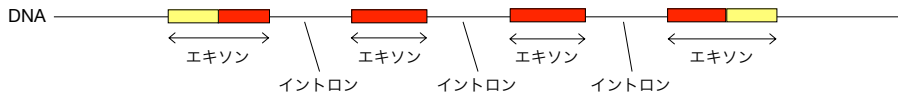
- 「複雑な」生物ほどゲノムの塩基数が多い？ ×
- 「複雑な」生物ほど遺伝子数が多い？ ×
- 「複雑な」生物ほど生成されるタンパクの数が多い？ △

イントロン



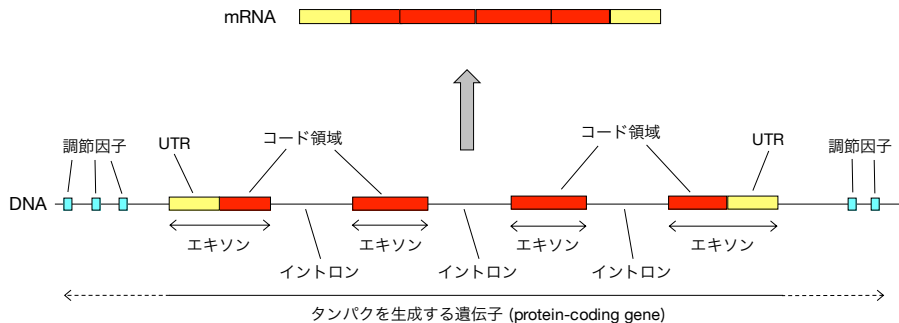
- ない遺伝子もたくさんある
- ほとんどない種もある
- 転写の制御や転写効率に関与している場合も
- なくても構わない（機能を持たない）ものが多い

イントロンの進化



- ほとんどの真核生物の共通祖先がイントロンを大量に保持
- 生物の複雑さとイントロンの多さはあまり関係しない
(ヒトよりイントロンがはるかに多い微生物も多数存在)
- 大昔になぜか一気に増幅した
 - ▶ 個体数が少なくて有利じゃない変異が広まりやすかった？
 - ▶ 自己増殖の機構を備えていた？

転写の制御



- ヒト・哺乳類で特に複雑 !?

種の形質の違いはゲノムのどんな違いに由来するのか

- 遺伝子の違い !? タンパクを生成する配列の違い !?
 - ▶ 種間で大きく違う
- 遺伝子の発現パターン（いつどこでどれだけタンパクが生成されるか）の違い？
 - ▶ ヒトとチンパンジーではこの関与が大きいのでは

生物の「複雑さ」は何が決めているの？

- 「複雑な」生物ほどゲノムの塩基数が多い？ ×
- 「複雑な」生物ほど遺伝子数が多い？ ×
- 「複雑な」生物ほど生成されるタンパクの数が多い？ △
- 「複雑な」生物ほど転写の制御（ネットワーク）が複雑 !?

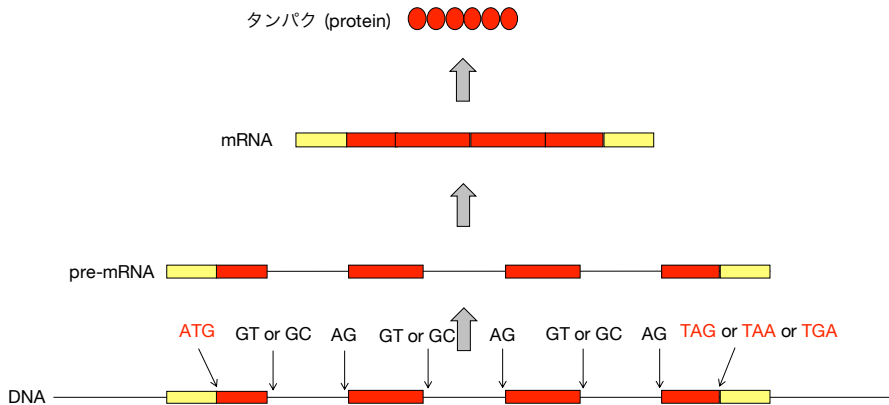
ゲノムを決定した後どうするの？

- 配列を決定しただけではあまり意味がない（ただのデータ）
- データに意味のある情報を「注釈付け」（アノテーション）する必要がある
- 一番欲しい情報： **どこにどんな遺伝子があるか**
- どうやってどこにどんな遺伝子があるかわかるの？

遺伝子予測

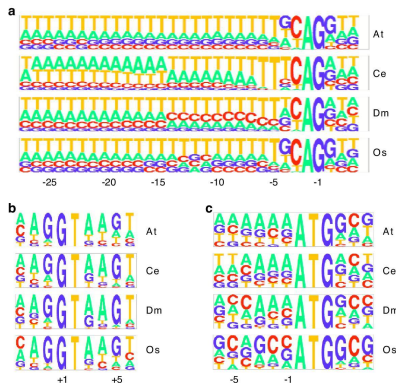
- 機械学習などを用い、様々な情報を活用して予測
 - ▶ 遺伝子に特徴的なパターン
 - ▶ 実験により得られた情報
 - ▶ 他の種の遺伝子

遺伝子に特徴的なパターン



- 翻訳 : ATG から開始、3 文字ずつアミノ酸に変換、TAA/TAG/TGA で終了

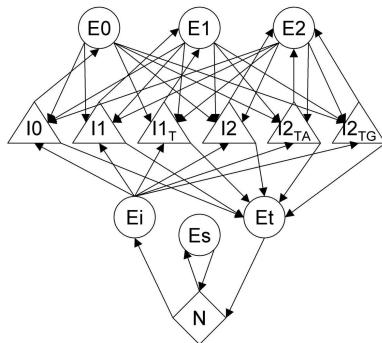
遺伝子に特徴的なパターン



Korf (2004) BMC Bioinformatics

- 実験的にスプライスサイトと分かっている GT,AG を集める
- ゲノム中の全ての GT と AG につき、機械学習によりスプライスサイトかどうかを判定

隠れマルコフモデル (HMM)



Korf (2004) BMC Bioinformatics

- エキソンとイントロンの間の移行確率などの HMM を構築

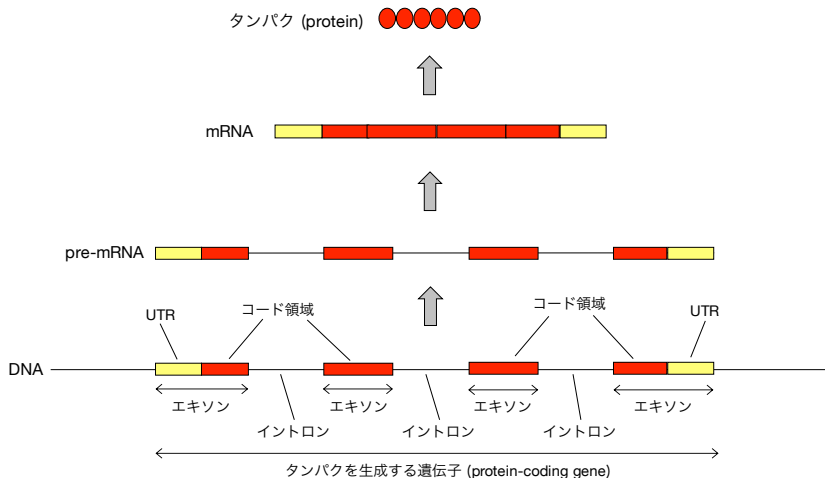
遺伝子予測

- 実験により得られた情報
 - ▶ RNA を抽出し、配列を決定
 - ▶ 各 RNA に対応する DNA 領域を探索
- 他の種の遺伝子
 - ▶ 遺伝子のアミノ酸配列は種間で非常に似ていることが多い
 - ▶ 他の種で同定されている遺伝子に対応する（似ている）領域を探索

相同性検索

- 1つの配列 vs 全ゲノム、全遺伝子
 - ▶ 相同性検索 (BLAST) blast.ncbi.nlm.nih.gov
- 2つの「相同な」配列を並べる（アラインメント）
- 1つのゲノム vs 1つのゲノム

遺伝子の構造



- コード領域：ヒトゲノムの <2%

非コード領域にはどんな情報があるの？

- タンパクをいつ、どこで、どれくらい生成するかという情報
- ほとんど (~90%) が役に立たない情報 !? 「ジャンク DNA」
- **トランスポゾン (転移因子)**
 - ▶ ゲノムに寄生するウイルスのようなもの
 - ▶ ゲノム中でどんどん増殖していく
 - ▶ ヒトゲノムの ~60%
 - ▶ ゲノムの大きさの違いは主にトランスポゾンの違い
 - ▶ 悪影響を及ぼすことが多いが、重要な役割を持つものもある

ゲノムのほとんどがジャンク DNA?

- 役に立たないものは進化の過程で除かれるのでは？
 - ▶ 遺伝的浮動
 - ▶ 「宿主」とトランスポゾンや、トランスポゾン間の「軍拡競争」
 - ▶ ヒトのゲノムのほとんどが何らかの機能を持っているとすると、じゃあタマネギやバッタはどうなるの？
- 一方、未知の機能領域がたくさんあるのも事実