はじめに

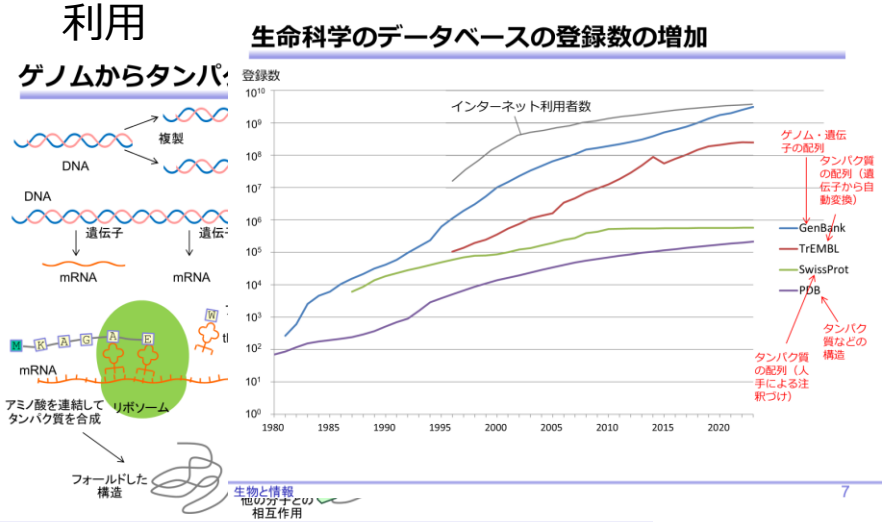
清水謙多郎
shimizuk@fc.jwu.ac.jp

授業の目的

ゲノムやタンパク質などの生物情報のさまざまなデータベース、データ解析手法を紹介し、それらの実習を通して、生命科学、健康・医療、農学の分野で情報科学がどのように利用されているかを体験する。

1. ゲノム、遺伝子、タンパク質

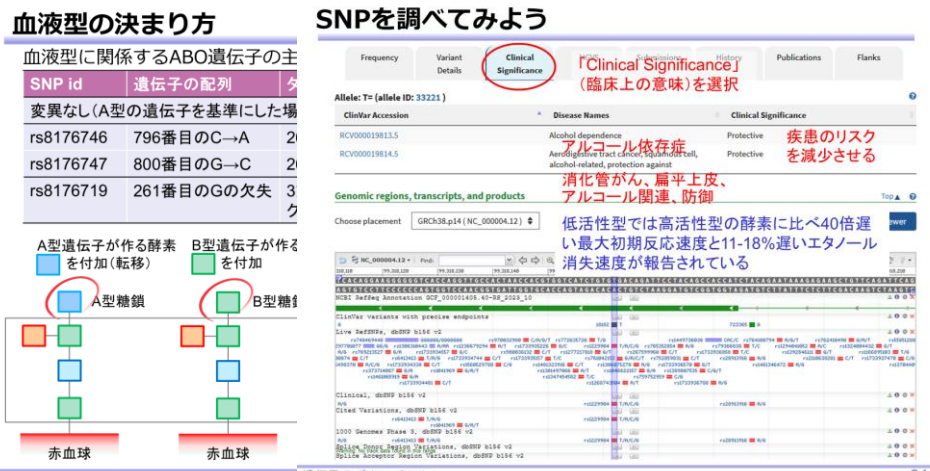
分子生物学の基礎、データサイエンスの利用



2. ゲノムブラウザによるゲノムの閲覧、ヒトと他の生物のゲノム、遺伝子の比較

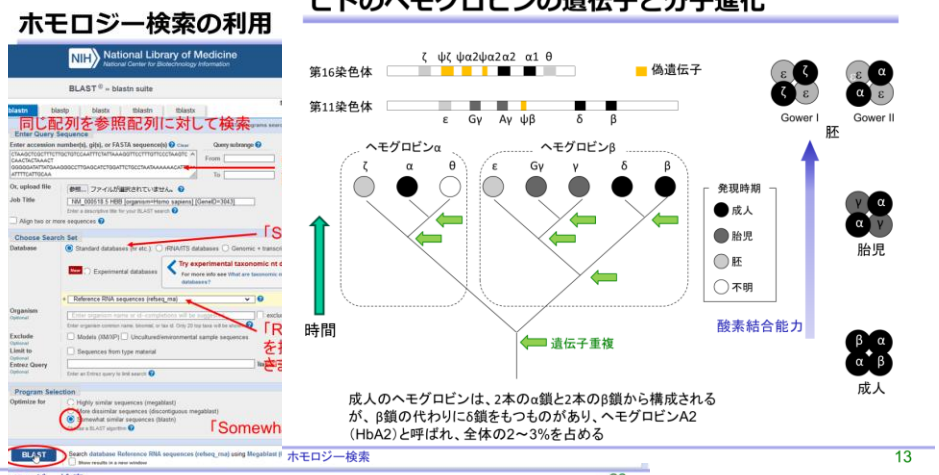
3. 遺伝子のバリアント

お酒が飲める／飲めないに関わる遺伝子、血液型の遺伝子、短距離走／長距離走のどちらが得意かに関わる遺伝子など



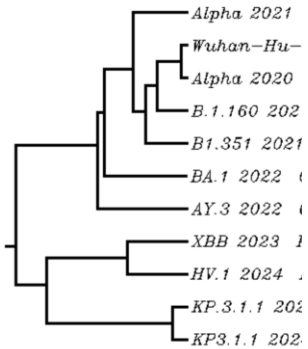
4. ホモロジー検索

アミラーゼや嗅覚遺伝子の進化、マンモスがもつ遺伝子が現存するどの生物の遺伝子に近いか



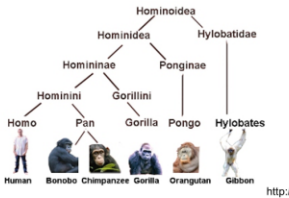
5. 進化の解析

いろいろな遺伝子の系統樹を描く、新型コロナウイルスのゲノムの変異を追う



霊長類のCOX遺伝子の配列（hominoidea-cox1.fasta）について以下の問いに答えよ。

1. Clustal Wのサイトを用いて、分子系統樹を作成せよ。
2. 実際のヒト上科の系統樹は以下の通りであることが知られている。2.で得られた結果と比較し、そうした結果が得られる理由を考えてみよう。



COX: ミトコンドリアのエネルギー生成に関わる酵素

http://aquatic-human-ancestor.org/timeline.html

6. ゲノムやタンパク質の配列の解析

大腸菌のプロモーター領域の解析

大腸菌の10個のプロモーター領域のマルチプルアラインメント

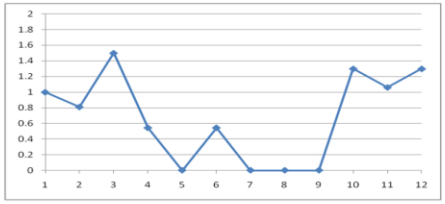
遺伝子	−35領域
araB	GATCCTACCTGACGCTT
araC	CCGTGATTATAGACACT
bioA	CAAAACGTGTTTTTTGT
bioB	ATAATCGACTTGTAAAC
galE	TTATTCCATGTCACACT
lacZ	ACCCCAGGCTTTACACT
lacI	CCATCGAATGGCGCAAA
rrnE1	ATTTTTCTATTGCGGCC
λpR	CCGTGCGTGTTGACTAT
λpL	CTGGCGGTGTTGACATA

転写因子結合部位のロゴ表示

WebLogo表示

位置	A	C	G	T
1	0	0.5	0	0.5
2	0	0.75	0	0.25
3	0.25	0.5	0	0.25
4	0.875	0.125	0	0
5	0	0	0	1
6	0	0	0.125	0.875
7	0	0	1	0
8	0	0	0	1
9	0	0	0	1
10	0	0.625	0.125	0.25
11	0.125	0	0.125	0.75
12	0	0.625	0.25	0.125

エントロピーH

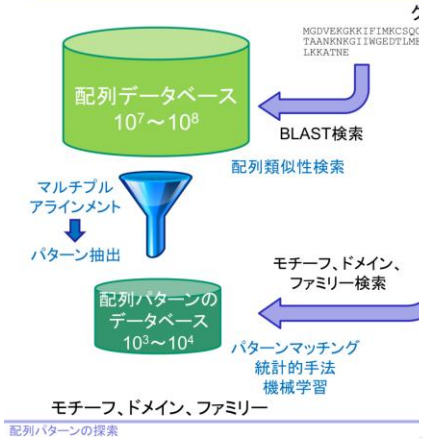


位置	エントロピー
1	1.000
2	0.811
3	1.500
4	0.544
5	0.000
6	0.544
7	0.000
8	0.000
9	0.000
10	1.299
11	1.061
12	1.299

7. 配列パターンの解析

モチーフ、ドメインのようなタンパク質の機能を特徴づけるパターンとその解析方法

モチーフ、ドメイン、ファミリー



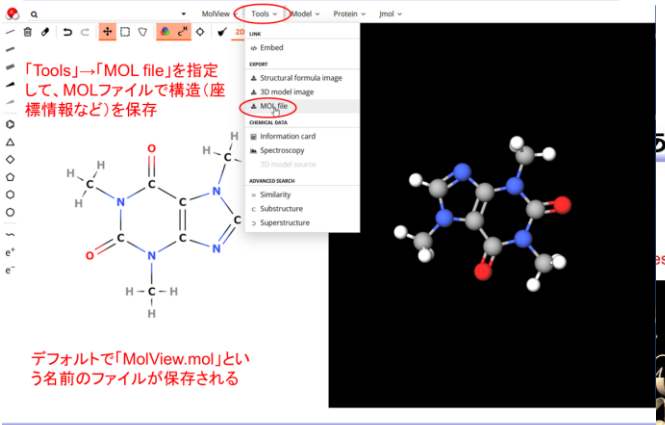
AIの利用

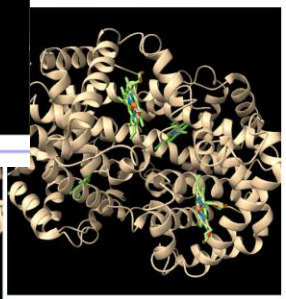
Pythonの実習

```
[ ]  generated_df.to_csv("peptide-protGPT.csv", index=False)

シグナル配列の予測

# 必要なライブラリをインポー
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifi
GradientBoostingClassifier, ExtraTreesClassifier, BaggingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report, roc_curve, auc, ac
precision_score, f1_score
import matplotlib.pyplot as plt

# 外部ライブラリのモデル (XGBoost, LightGBM)
from xgboost import XGBClassifier
# from lightgbm import LGBMClassifier  # ネットワークエラーで未インスト

# データ読み込みと整形
df_pos = pd.read_csv("long.csv")
df_neg = pd.read_csv("long_neg.csv")
df_pos.columns = [col.strip() for col in df_pos.columns]
df_neg.columns = [col.strip() for col in df_neg.columns]
seq_col_pos = 'seq' if 'seq' in df_pos.columns else 'peptide'
seq_col_neg = 'peptide'
df_pos = df_pos.rename(columns={seq_col_pos: 'peptide'})
```


8. 分子グラフィックスの利用

MolViewの利用





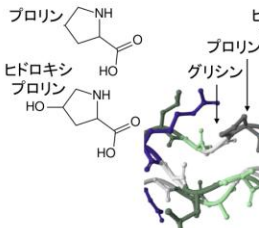
9. タンパク質の構造と機能

さまざまなタンパク質の構造と機能を紹介

コラーゲン

コラーゲン PDB ID:1BKV

皮膚や腱・軟骨などを構成する繊維状タンパク質で、人体のタンパク質全体の30%を占める



Chimeraで水素原子を付加し、水素結合を表示させてみよう

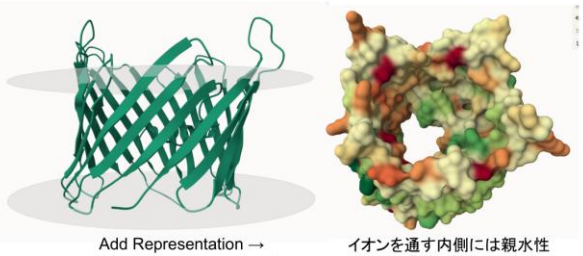
タンパク質の構造

膜タンパク質（イオンチャネルの例）

イオンチャネル PDB ID: 2JK4

膜電位の変化によって開閉が制御されているイオンチャネル

膜タンパク質構造データベースOPMにアクセスし、膜の表面も合わせた構造を取得して、表示してみよう



Add Representation → Membrane Orientationで膜の配向を表示

イオンを通す内側には親水性のアミノ酸が比較的多く存在

タンパク質の構造 59

10.タンパク質の構造の比較と解析

DALIの検索結果の例（2）

Results: 1atnA

Chain: 1atnA

- Matches against PDB25, Correlation matrix
- Matches against PDB50
- Matches against PDB90
- Matches against full PDB
- Download matches against PDB25
- Download matches against PDB50
- Download matches against PDB90
- Download matches against full PDB

Results will be deleted after one week.

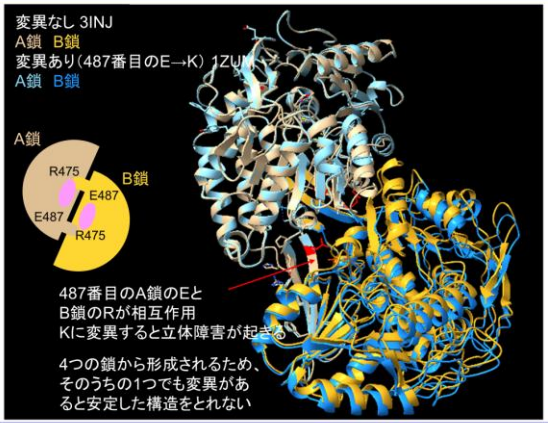
非常に多数のヒットがあるPDBの全チェインが対象
骨格筋ミオシン IIなどの分子が検索結果に表れているが、アクチンを複合体として含んでいるためヒットした

熱ショックプロテインは1000位以内ではヒットしない

Results: 1atnA

タンパク質の構造比較

ALDH2の構造



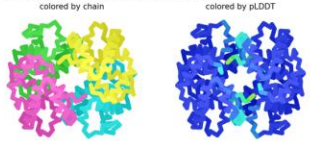
遺伝子のバリエーション 18

11. 分子の構造モデリング、ドッキング

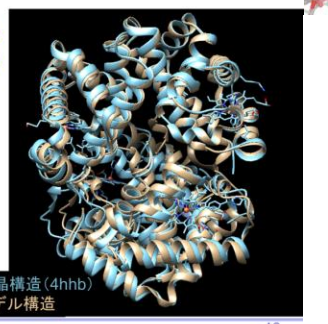
ColabFoldを用いたタンパク質のモデリングを行い、タミフルとウイルスタンパク質とのドッキング

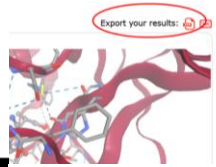
ヘモグロビンの構造モデリング

ヘモグロビン（4つの鎖まとめて）のモデリングの結果（一部）

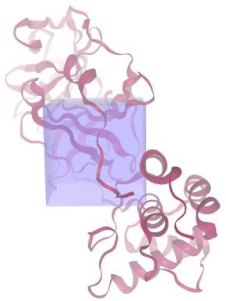


結晶構造（実験によって決定された構造）とモデル構造のとの差は、一致部分は0.437Å（145残基）





結果のダウンロード



タンパク質構造のモデリング 40

12. ネットワーク解析

がんに関わるパスウェイ、カフェイン代謝などの紹介、ネットワーク解析の基礎

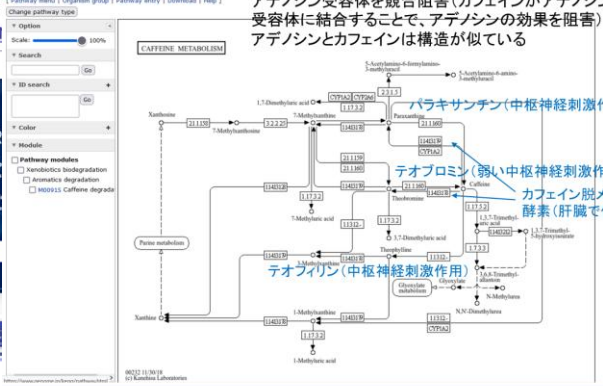
Gene Ontology

- Gene Ontology（GO）
  ノテーションを統一的に
- GO Term: GOの統制語
- http://geneontology.or

21 ゲノムオントロジー

代謝経路の探索（4）

アデノシンは脳内で眠気を誘発するが、カフェインはアデノシン受容体を競合阻害（カフェインがアデノシン受容体に結合することで、アデノシンの効果を阻害）
アデノシンとカフェインは構造が似ている



ネットワーク解析 43

参考書

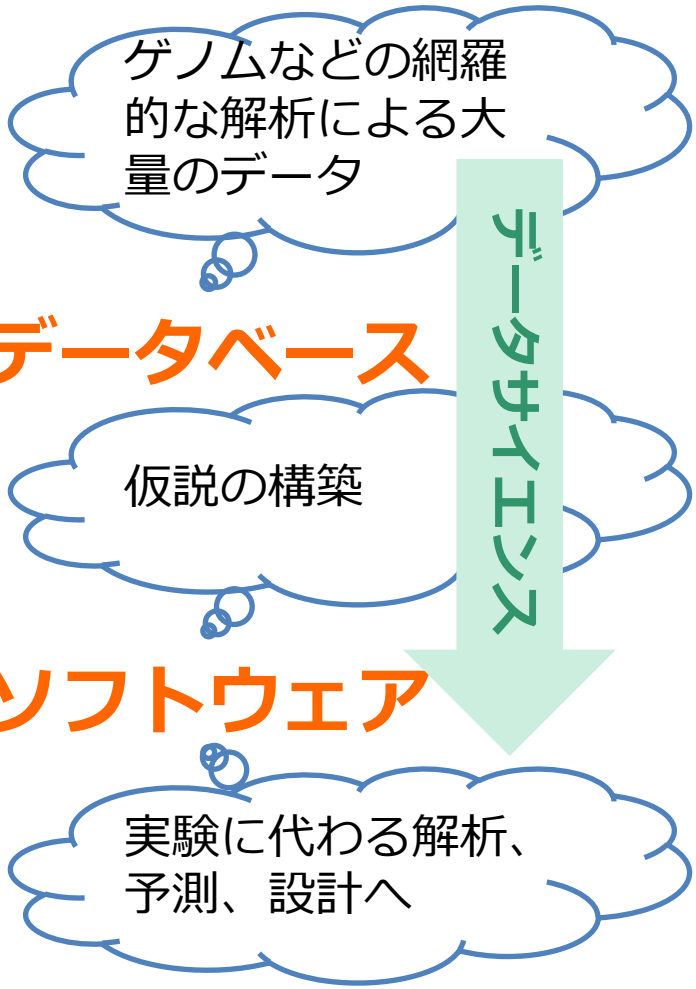
とくに指定しませんが、バイオインフォマティクスの分野、データ解析についてしっかり学びたいという人には以下の書籍を紹介します。

門田幸二, 清水謙多郎, 岸野洋久, 寺田透（編）, Web連携テキスト バイオインフォマティクス 基礎と応用, 培風館, 2022.

基本的な教科書がありますので、あらためて紹介しますが、この講義の内容をカバーするものはありません。

生物と情報

生物という複雑な対象を理解するには、
実験データを蓄積し、



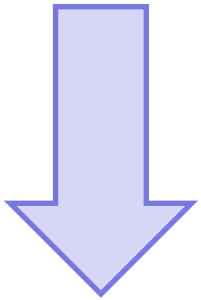
それを解析することにより、
そこに埋もれている情報を探し出して、
新しい知識を得る

生命現象をシミュレートして、
生物の理解、応用に役立てる

理論

生物と情報

- 1980年代後半、ヒトゲノムプロジェクトの発足 → 大量のゲノムデータの処理

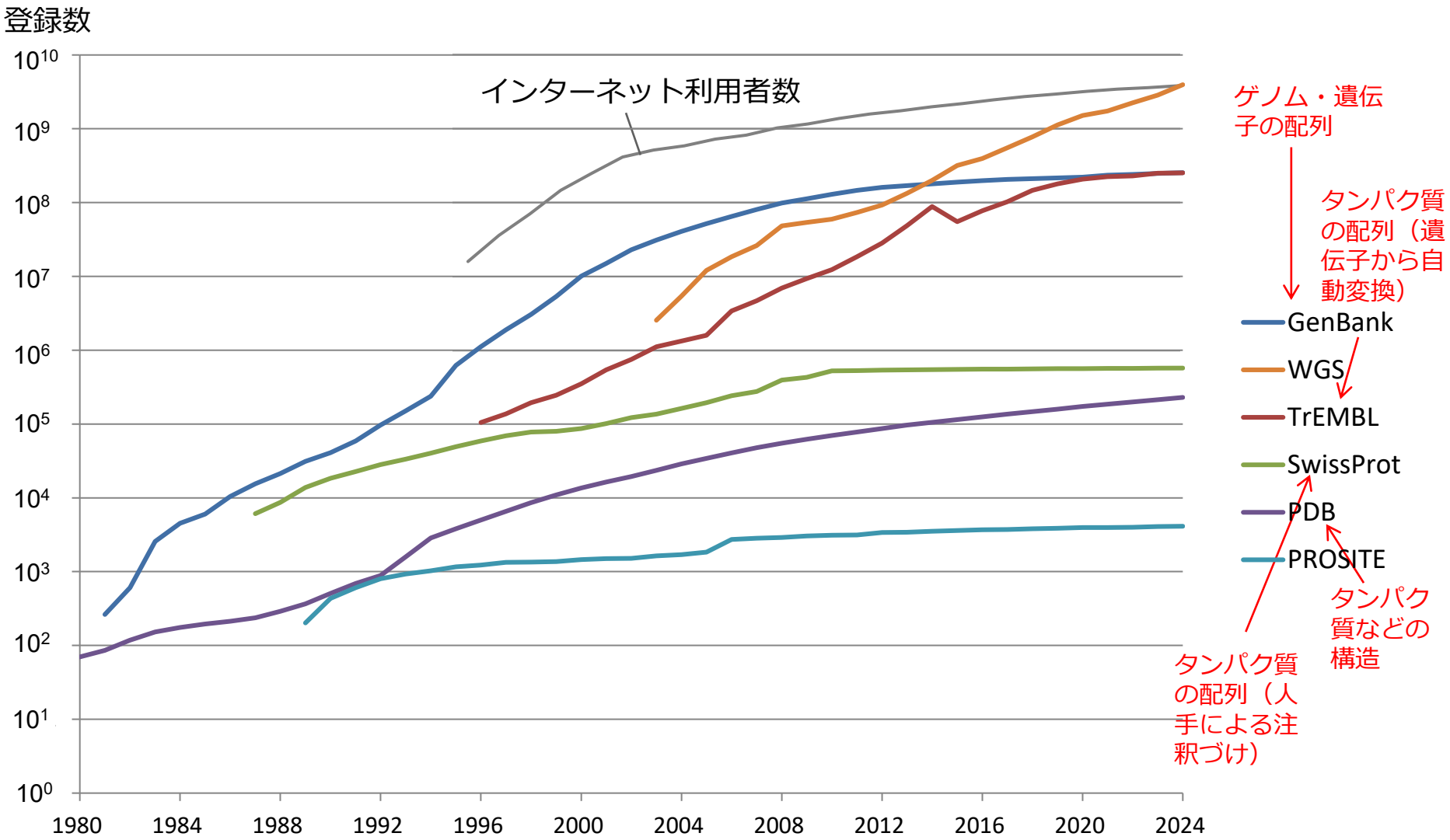


データベース、ソフトウェアの研究は1970年代から

ヒトゲノム解読以降、「ポストゲノム」の時代

- 多数の生物、個体（個人）のゲノム、さらにタンパク質、代謝物質、表現型などを含めた超大量かつ多様なデータ（ビッグデータ）の解析の必要性

生命科学のデータベースの登録数の増加



データベースの話は、あとで改めてします

生命科学のビッグデータ 1

- **ゲノムデータ**
  - ヒトゲノム1個のシーケンスデータは約3.2Gb（ギガベース）
  - 30×深度での全ゲノムシーケンスのrawデータで 80～200GB（ギガバイト）程度／個人
  - 大規模なゲノムプロジェクトでは数十PBの規模に達する（→ バイオバンクデータ）
- **トランスクリプトームデータ**
  - RNAシーケンス（RNA-seq）のデータは、1回のサンプルで10～30Gb
  - 数百～数千サンプルを扱うプロジェクトでは、データ量は TB～数十TB
  - シングルセルRNA-seq（scRNA-seq）では、細胞数やリード数でさらに指数的に増加、サンプルあたりの原データ＋中間ファイルで、1～5TBを超えることもある
- **プロテオームデータ**
  - 質量分析を用いたタンパク質の同定・定量解析では、1回の実験で10GB～100GBのデータが生成される
  - 大規模プロジェクト（例えば人口集団・疾患集団）では 数TB～10PBのデータが蓄積されることがある
- **メタボロームデータ**
  - メタボローム解析では、1回のサンプルで数MBから数GBのデータが生成される
  - 大規模な研究では数百～数千サンプルを解析 → 数十TBのデータ量

生命科学のビッグデータ 2

- **パーソナルヘルスデータ**
  - ウェアラブルデバイスや健康管理アプリから得られるデータは1人あたり数MBから数GB/日程度のデータ量
  - 数百万人規模のコホート研究や国民健康モニタリングプロジェクトでは 数PB 規模のデータが生成・保管されている
- **バイオイメージングデータ**
  - 1回のMRIスキャンで数百MB～数GBのデータが生成される
  - 顕微鏡画像やタイムラプス／3D撮像などでは、1実験で 数十～数百GB
  - 多数の症例や多数の時間点を含む研究で 数十～数百TB に拡大
  - イメージングデータは、分子イメージング、植物のイメージング、微生物の観察、生態系モニタリングなど多岐にわたる
- **バイオバンクデータ**
  - UK Biobank → 約50万人分の健康情報やゲノムデータ、総データ量は約28PB
  - 米国 All of Us → 100万人規模を目標とする国民コホート研究、健康情報・電子カルテ・ゲノムデータを収集、既に数十万人分が登録済み
  - 中国 GSA (Genome Sequence Archive) → 20PB超
  - 東北メガバンク → 地域住民の健康情報やゲノムデータ、総データ量は十数PB

参考

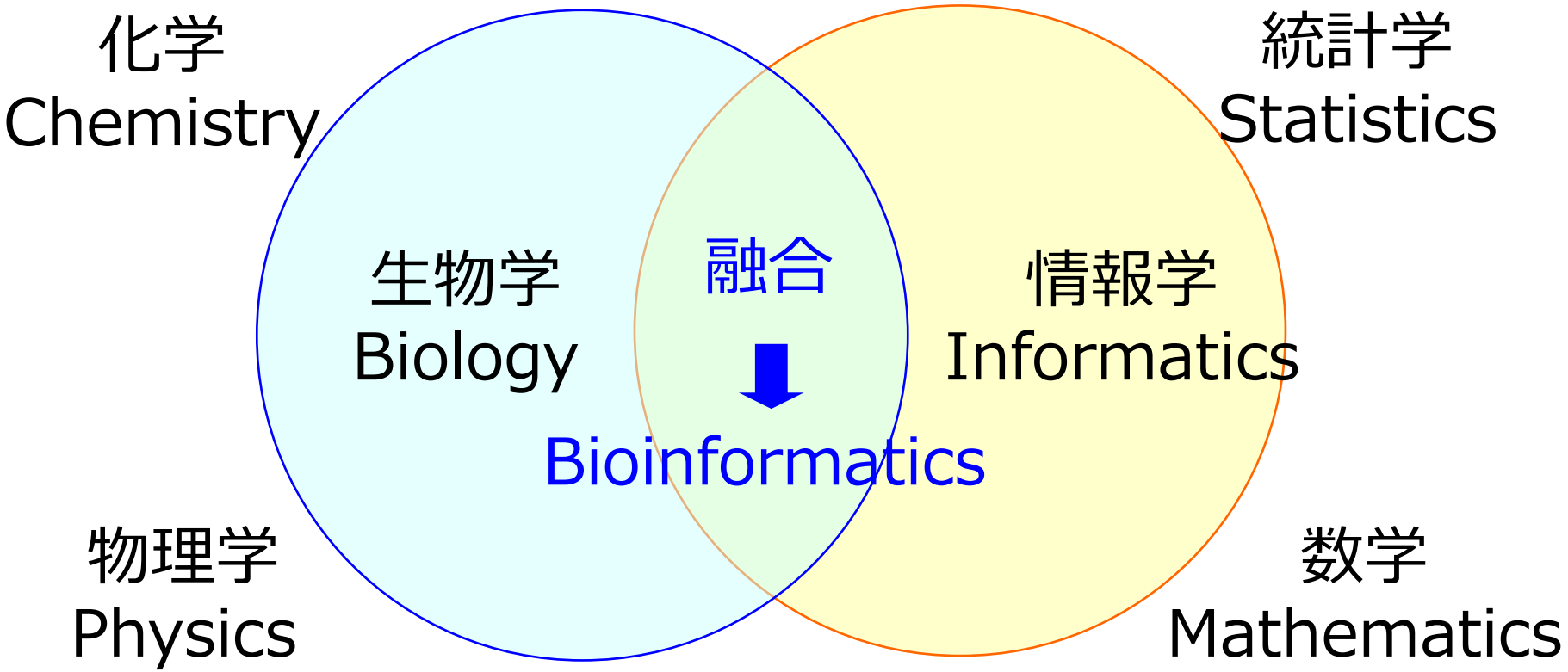
単位について

接頭語	名前	読み	国際単位系(SI)	情報系	命数	語の意味
Y	Yotta	ヨタ	10^{24}	2^{80}=1,208,925,819,614,629,174,706,176	秄	8
Z	Zetta	ゼタ	10^{21}	2^{70}=1,180,591,620,717,411,303,424	10垓	7
E	Exa	エクサ	10^{18}	2^{60}=1,152,921,504,606,846,976	100京	6
P	Peta	ペタ	10^{15}	2^{50}=1,125,899,906,842624	1000兆	5
T	Tera	テラ	10^{12}	2^{40}=1,099,511,627,776	兆	怪物
G	Giga	ギガ	10^{9}	2^{30}=1,073,741,824	10億	巨人
M	Mega	メガ	10^{6}	2^{20}=1,048,576	100万	大量
k	Kilo	キロ	10^{3}	2^{10}=1,024	千	1000
h	hecto	ヘクト	10^{2}		百	100
da	deca	デカ	10^{1}		十	10
d	deci	デシ	10^{-1}			10
c	centi	センチ	10^{-2}			100
m	miri	ミリ	10^{-3}			1000
μ	micro	マイクロ	10^{-6}			微小
n	nano	ナノ	10^{-9}			小人
p	pico	ピコ	10^{-12}			先端
f	famto	ファムト	10^{-15}			15
a	atto	アト	10^{-18}			18
z	zepto	セプト	10^{-21}			7
y	yocto	ヨクト	10^{-24}			8

SIと区別するため、IEC（国際電気標準会議）では、2のべき乗に対して、binaryを付ける単位を提唱している。例えば、2^{10} はkilobinary（略称Kibi, Kiと表記）, 同様に、2^{20} はMibi, Mi, 2^{30} はGibi, Giとなる。

バイオインフォマティクスとは

生命科学の問題を情報学（インフォマティクス）の考え方や手法によって解決しようという学問



in silico コンピュータによる（シリコン内で）

⟺ **in vitro** 試験管内で

in vivo 生体内で

コンピュータ科学と分子生物学の発展

コンピュータ科学		分子生物学
	2025	
生成系AI		AlphaFold
ビッグデータ　機械学習		ゲノム編集　合成生物学
次世代スーパーコンピュータ		メタゲノム
クラウドコンピューティング		次世代シーケンサー
Google検索　Wikipedia	2000	ES細胞、iPS細胞
		ヒトゲノムの解読
		クローン羊ドリー
DVD		DNAマイクロアレイ
World Wide Web		
CD-ROM　携帯電話		PCR法
スーパーコンピュータ		塩基配列データベース
商用パーソナルコンピュータ	1975	
関係データベース　ARPANET		遺伝子工学の特許
UNIX		組換えDNA
		コドン
		形質導入
人工知能、数値計算		DNA構造の発見
シャノン「通信の数学的理論」	1950	
プログラム内蔵型コンピュータ		遺伝子の本体がDNAであることを証明
		一遺伝子一酵素説
チューリングマシン		ウィーバーによる分子生物学の提唱
		合成抗生物質
	1925	ペニシリン

情報学の概念の生物学への浸透

- 核酸はタンパク質を正しい時と場所でつくるための情報を暗号化してもっている
- DNA配列にコードされた情報がどのようにタンパク質に翻訳されるか
- 細胞の死は、内部プログラムによる
- フィードバック制御によって、受容体からのシグナル伝達を負に調節する
- 転写回路によって細胞は論理演算を遂行できる
- それぞれの細胞は特定の組み合わせの細胞外シグナル分子に応答するようにプログラムされている
- 初期の発生段階では、シグナル伝達経路間のクロストークがきわめて強くなる

・・・

細胞の分子生物学 第6版より

プログラミング言語

バイオインフォマティクスでよく使われるプログラミング言語

		バイオ分野のデータ解析に便利な拡張（ライブラリ）
R	統計解析のための言語および開発実行環境、豊富なパッケージが利用可能	多数のバイオインフォマティクスパッケージ
Python	文法がシンプル、書きやすさを重視、豊富なライブラリ、AI、自然言語処理、ソフトウェアツールなど	biopython
Perl	テキスト処理機能が充実、正規表現が組み込まれる、書きやすさ、豊富なライブラリ	BioPerl
Ruby	オブジェクト指向言語として開発される、書きやすさを重視、豊富なライブラリ、Webシステム開発など	BioRuby
Java	オブジェクト指向、機能が豊富、高性能、豊富なライブラリ、システム開発に利用される	BioJava
C++	C言語の拡張、オブジェクト指向、機能が非常に豊富、高性能、システム開発に広く利用される	B10.CPP

PubMedの論文検索ヒット数

PubMed: 生命科学や生物医学に関する文献検索エンジン