

はじめに

清水謙多郎

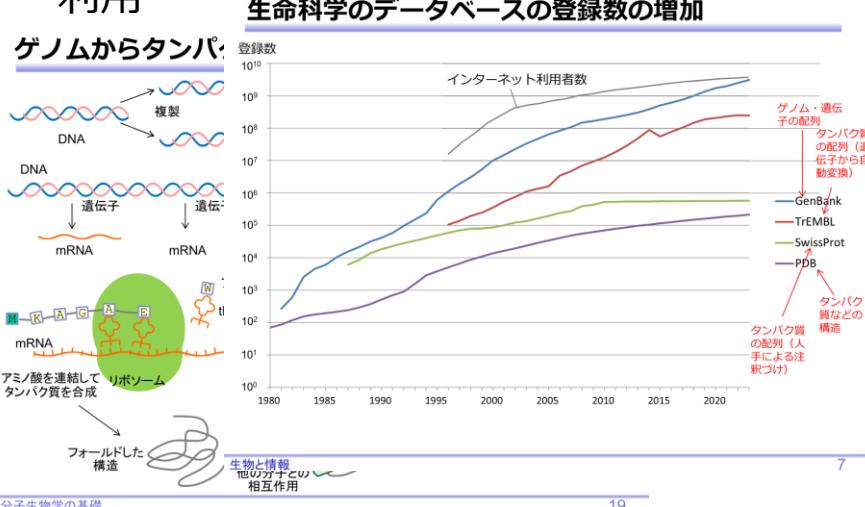
shimizuk@fc.jwu.ac.jp

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

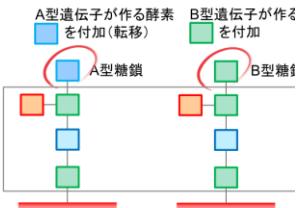


3. 遺伝子のバリアント

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

血液型の決まり方

血液型に関するABO遺伝子の主SNP id 遺伝子の配列 変



SNPを調べてみよう

Frequency Variant Clinical Significance Publications Flanks

Allele T= (allele ID: 33221)

ClinVar Accession Disease Names Clinical Significance

RCV000019815.5 Alcohol-Dependence Protective 疾患のリスクを減少させる

RCV000019814.5 Aerodigestive tract cancer, alcohol-related, protection against

Genomic regions, transcripts, and products

Choose placement GRCh38.p14 (NC_000004.12) ▾

消化管がん, 扁平上皮, アルコール関連、防御

低活性型では高活性型の酵素に比べ40倍遅い最大初期反応速度と11-18%遅いエタノール消失速度が報告されている

NC_000004.12 - Fwd 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000 10100 10200 10300 10400 10500 10600 10700 10800 10900 11000 11100 11200 11300 11400 11500 11600 11700 11800 11900 12000 12100 12200 12300 12400 12500 12600 12700 12800 12900 13000 13100 13200 13300 13400 13500 13600 13700 13800 13900 14000 14100 14200 14300 14400 14500 14600 14700 14800 14900 15000 15100 15200 15300 15400 15500 15600 15700 15800 15900 16000 16100 16200 16300 16400 16500 16600 16700 16800 16900 17000 17100 17200 17300 17400 17500 17600 17700 17800 17900 18000 18100 18200 18300 18400 18500 18600 18700 18800 18900 19000 19100 19200 19300 19400 19500 19600 19700 19800 19900 20000 20100 20200 20300 20400 20500 20600 20700 20800 20900 21000 21100 21200 21300 21400 21500 21600 21700 21800 21900 22000 22100 22200 22300 22400 22500 22600 22700 22800 22900 23000 23100 23200 23300 23400 23500 23600 23700 23800 23900 24000 24100 24200 24300 24400 24500 24600 24700 24800 24900 25000 25100 25200 25300 25400 25500 25600 25700 25800 25900 26000 26100 26200 26300 26400 26500 26600 26700 26800 26900 27000 27100 27200 27300 27400 27500 27600 27700 27800 27900 28000 28100 28200 28300 28400 28500 28600 28700 28800 28900 29000 29100 29200 29300 29400 29500 29600 29700 29800 29900 29999 30000 30001 30002 30003 30004 30005 30006 30007 30008 30009 30010 30011 30012 30013 30014 30015 30016 30017 30018 30019 30020 30021 30022 30023 30024 30025 30026 30027 30028 30029 30030 30031 30032 30033 30034 30035 30036 30037 30038 30039 30040 30041 30042 30043 30044 30045 30046 30047 30048 30049 30050 30051 30052 30053 30054 30055 30056 30057 30058 30059 30060 30061 30062 30063 30064 30065 30066 30067 30068 30069 30070 30071 30072 30073 30074 30075 30076 30077 30078 30079 30080 30081 30082 30083 30084 30085 30086 30087 30088 30089 30090 30091 30092 30093 30094 30095 30096 30097 30098 30099 300999 301000 301001 301002 301003 301004 301005 301006 301007 301008 301009 301010 301011 301012 301013 301014 301015 301016 301017 301018 301019 301020 301021 301022 301023 301024 301025 301026 301027 301028 301029 301030 301031 301032 301033 301034 301035 301036 301037 301038 301039 3010399 3010400 3010401 3010402 3010403 3010404 3010405 3010406 3010407 3010408 3010409 3010410 3010411 3010412 3010413 3010414 3010415 3010416 3010417 3010418 3010419 3010420 3010421 3010422 3010423 3010424 3010425 3010426 3010427 3010428 3010429 3010430 3010431 3010432 3010433 3010434 3010435 3010436 3010437 3010438 3010439 30104399 30104400 30104401 30104402 30104403 30104404 30104405 30104406 30104407 30104408 30104409 30104410 30104411 30104412 30104413 30104414 30104415 30104416 30104417 30104418 30104419 30104420 30104421 30104422 30104423 30104424 30104425 30104426 30104427 30104428 30104429 301044299 301044300 301044301 301044302 301044303 301044304 301044305 301044306 301044307 301044308 301044309 301044310 301044311 301044312 301044313 301044314 301044315 301044316 301044317 301044318 301044319 301044320 301044321 301044322 301044323 301044324 301044325 301044326 301044327 301044328 301044329 3010443299 3010443300 3010443301 3010443302 3010443303 3010443304 3010443305 3010443306 3010443307 3010443308 3010443309 3010443310 3010443311 3010443312 3010443313 3010443314 3010443315 3010443316 3010443317 3010443318 3010443319 3010443320 3010443321 3010443322 3010443323 3010443324 3010443325 3010443326 3010443327 3010443328 3010443329 30104433299 30104433300 30104433301 30104433302 30104433303 30104433304 30104433305 30104433306 30104433307 30104433308 30104433309 30104433310 30104433311 30104433312 30104433313 30104433314 30104433315 30104433316 30104433317 30104433318 30104433319 30104433320 30104433321 30104433322 30104433323 30104433324 30104433325 30104433326 30104433327 30104433328 30104433329 301044333299 301044333300 301044333301 301044333302 301044333303 301044333304 301044333305 301044333306 301044333307 301044333308 301044333309 301044333310 301044333311 301044333312 301044333313 301044333314 301044333315 301044333316 301044333317 301044333318 301044333319 301044333320 301044333321 301044333322 301044333323 301044333324 301044333325 301044333326 301044333327 301044333328 301044333329 3010443333299 3010443333300 3010443333301 3010443333302 3010443333303 3010443333304 3010443333305 3010443333306 3010443333307 3010443333308 3010443333309 3010443333310 3010443333311 3010443333312 3010443333313 3010443333314 3010443333315 3010443333316 3010443333317 3010443333318 3010443333319 3010443333320 3010443333321 3010443333322 3010443333323 3010443333324 3010443333325 3010443333326 3010443333327 3010443333328 3010443333329 30104433333299 30104433333300 30104433333301 30104433333302 30104433333303 30104433333304 30104433333305 30104433333306 30104433333307 30104433333308 30104433333309 30104433333310 30104433333311 30104433333312 30104433333313 30104433333314 30104433333315 30104433333316 30104433333317 30104433333318 30104433333319 30104433333320 30104433333321 30104433333322 30104433333323 30104433333324 30104433333325 30104433333326 30104433333327 30104433333328 30104433333329 301044333333299 301044333333300 301044333333301 301044333333302 301044333333303 301044333333304 301044333333305 301044333333306 301044333333307 301044333333308 301044333333309 301044333333310 301044333333311 301044333333312 301044333333313 301044333333314 301044333333315 301044333333316 301044333333317 301044333333318 301044333333319 301044333333320 301044333333321 301044333333322 301044333333323 301044333333324 301044333333325 301044333333326 301044333333327 301044333333328 301044333333329 3010443333333299 3010443333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333315 3010443333333316 3010443333333317 3010443333333318 3010443333333319 3010443333333320 3010443333333321 3010443333333322 3010443333333323 3010443333333324 3010443333333325 3010443333333326 3010443333333327 3010443333333328 3010443333333329 30104433333333299 30104433333333300 3010443333333301 3010443333333302 3010443333333303 3010443333333304 3010443333333305 3010443333333306 3010443333333307 3010443333333308 3010443333333309 3010443333333310 3010443333333311 3010443333333312 3010443333333313 3010443333333314 3010443333333

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

ALDH2遺伝子の検索

生物種('Home Sapiens'(ヒト)がデフォルト)

いろいろな情報を使って検索できる

十をクリックすると、さらにたくさん

の生物種が表示される

Genome Reference Consortium (GRC)が

管理するヒカゲムのバージョン38、バッチ14の参照配列

The screenshot shows the NCBI Genome Data Viewer interface. On the left, a phylogenetic tree highlights the relationship between various species, including humans, chimpanzees, mice, rats, and many others. A red box highlights the search bar at the top, which contains the query "Homo sapiens (human)". Another red box highlights the assembly dropdown menu set to "GRCh38.p14". The main panel displays the "Homo sapiens (human)" genome page, showing assembly details like "Name: GRCh38.p14", "Release date: RS_2024_08", and "Release date: Aug 26, 2024". Below this, a genomic track viewer shows the 3'UTR region of the EK13 gene, spanning coordinates 111,809,543 to 111,817,532. The track includes multiple colored tracks representing different genomic features, with labels for "遺伝子", "エクソン13", "3'-UTR", and "ポリアデニル化シグナルなど". A red arrow points from the text "Genome Reference Consortium (GRC)が管理するヒカゲムのバージョン38、バッチ14の参照配列" to the assembly name "GRCh38.p14" in the assembly details section.

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

ホモロジー検索の利用

National Library of Medicine
National Center for Biotechnology Information

BLAST® - blastn suite

blastn **blastp** **blastx** **blastn** **blastx**

同じ配列を参照配列に対して検索

Enter query sequence...
 Enter accession number(s), gff or FASTA sequence(s)

Query exchange

From:
 To:

Or, specify file:

Job title:

Job ID:

Align two or more sequences

Choose Search Set

Databases

Standard databases (in NCBI BLAST databases) Genomic & environmental databases Experimental databases By experimental taxonomy (at NCBI BLAST databases)

Reference RNA sequences (RefSeq_rna) Reference protein sequences (RefSeq_protein) Reference nucleic acid sequences (RefSeq_nucleotide) Reference environmental sample sequences Reference transcript sequences (RefSeq_transcript) Reference genome sequences (RefSeq_genome) Reference metagenomic sequences (RefSeq_megabase) Reference similar sequences (blastn) Reference similar sequences (blastx) Reference similar sequences (blastp) Reference similar sequences (blastx2)

Organism:

Exclude: Models (MPN) Uncultured/environmental sample sequences

Limit to: Reference from type material Reference to type material

Entered Query:

Program Selection:

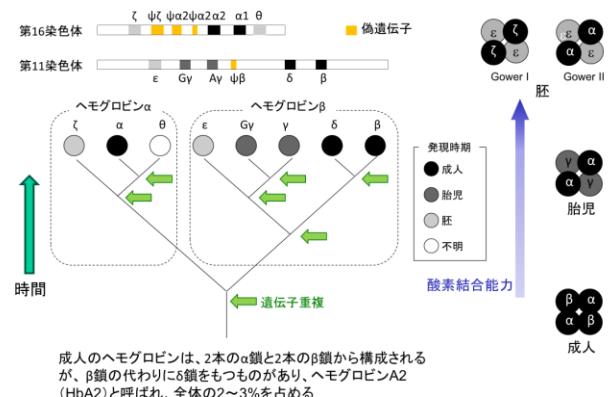
Optimize for: Highly similar sequences (megablast) Highly similar sequences (megablast) Somewhat similar sequences (blastn) Somewhat similar sequences (blastx) Somewhat similar sequences (blastp) BLAST+ (blastn)

Search database Reference RNA sequences (RefSeq_rna) using Megablast Search database Reference protein sequences (RefSeq_protein) using Megablast Search database Reference nucleic acid sequences (RefSeq_nucleotide) using Megablast Search database Reference environmental sample sequences (RefSeq_transcript) using Megablast Search database Reference genome sequences (RefSeq_genome) using Megablast Search database Reference metagenomic sequences (RefSeq_megabase) using Megablast

blastn

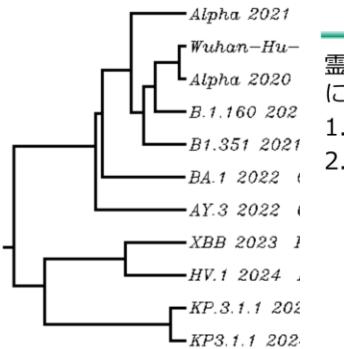
Search database Reference RNA sequences (RefSeq_rna) using Megablast

ヒトのヘムoglobinの遺伝子と分子進化



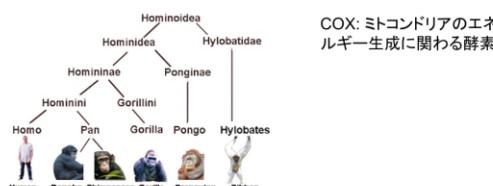
5. 進化の解析

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合わせよ。

- Clustal Wのサイトを用いて、分子系統樹を作成せよ。
- 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。

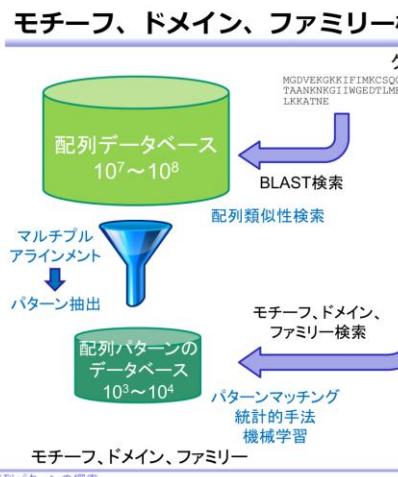


遺伝子のバリエーション

30

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



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

シグナル配列の予測 AIの利用 Pythonの実習

```
# 必要なライブラリをインポート
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, AdaBoostClassifier, 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, accuracy_score, f1_score
import matplotlib.pyplot as plt

# 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'

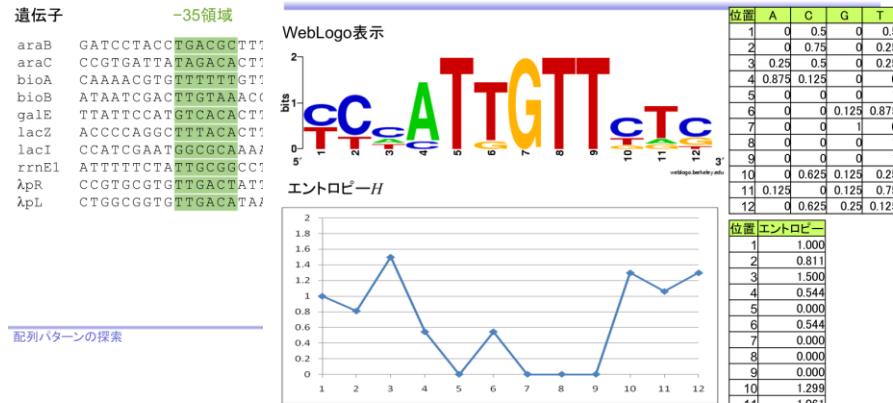
df_pos = df_pos.rename(columns=[seq_col_pos: 'peptide'])
```

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レアライメント

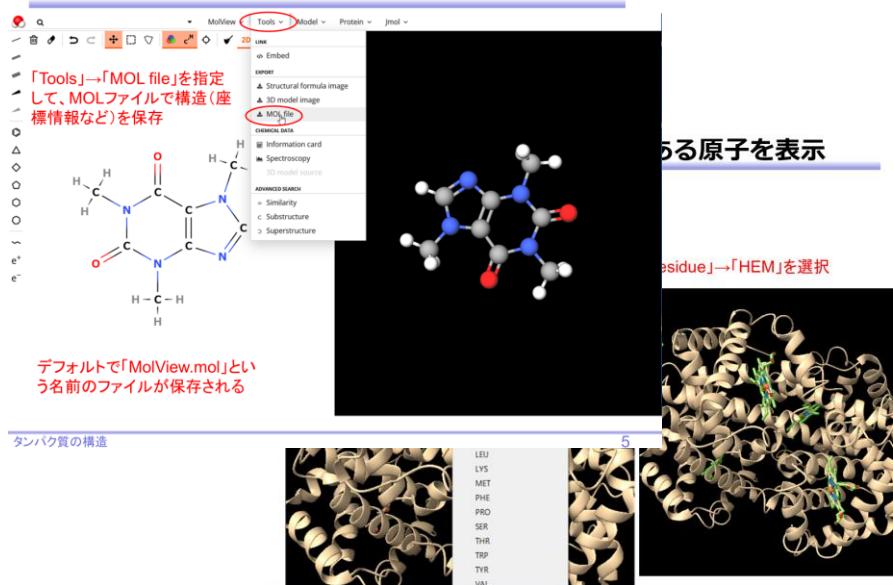
転写因子結合部位のロゴ表示



11

8. 分子グラフィックスの利用

MolViewの利用



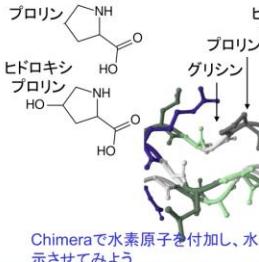
12

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
タンパク質で、人体のタンパク質全体
占める



タンパク質の構造

膜タンパク質（イオンチャネルの例）

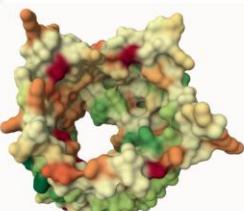
イオンチャネル PDB ID: 2JK4

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



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

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



59

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

ヘムoglobinの構造モデリング

ヘムoglobin(4つの鎖まとめ)のモデリングの結果(一部)

① 2022-03-03 04:06:02_149_Solving was successful, exec_time=0.000000
② 2022-03-03 04:06:02_149_Solving was successful, exec_time=0.000000
③ 2022-03-03 04:06:02_149_Solving was successful, exec_time=0.000000
④ 2022-03-03 04:06:02_149_Solving was successful, exec_time=0.000000

colored by chain

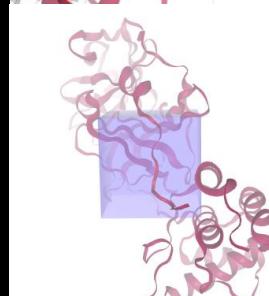


結晶構造(実験によって決定された構造)
とモデル構造との差は、一致部分は
0.437Å(145残基)

水色: 結晶構造(4hhb)
茶色: モデル構造



結果のダウンロード



40

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 PDB80
- **Matches against full PDB** (highlighted)
- Download matches against PDB25
- Download matches against PDB50
- Download matches against PDB80
- Download matches against full PDB

Results will be deleted after one week.

Results: 1atnA

Q1

MOI

Self

Exc

Stn

Stu

ALDH2の構造

変異なし 3INJ

A鎖 B鎖

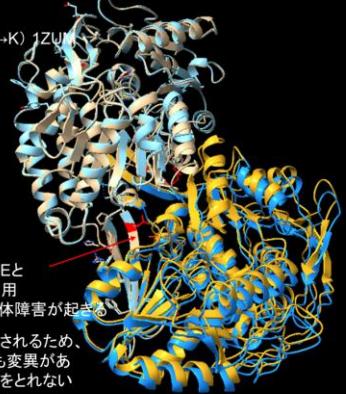
変異あり(487番目のE→K) 1ZU1

A鎖

R475 E487 R475

E487

R475



タンパク質の構造比較

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)

ノーテーションを統一的に

• GO Term: GOの統制語

• <http://geneontology.org>

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, computable knowledge base ranging from the molecular to the organism level, across the multiplicity of species.

The Gene Ontology (GO) knowledgebase is the world's largest source of information known to be both human-readable and machine-readable, and is a resource for scaling molecular biology and genetics experiments in biomedical research.

search GO term or Gene Product in AsciGO...

Any • Ontology • Gene Product

Search

ID search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

代謝経路の探索 (4)

Caffeine metabolism - Reference pathway

Pathway menu | Organism group | Pathway entry | Download | Help]

Change pathway type

Option Scale: 100%

Search

ID search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

Aromatic acid degradation

1009125 Caffeine degradation

Pathway

Search

Color

Pathway modules

Xanthine oxidase degradation

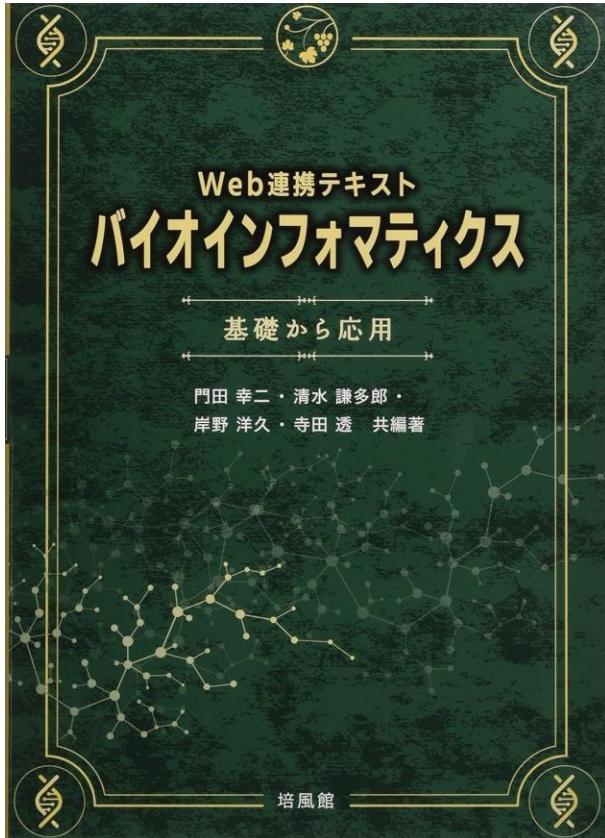
Aromatic acid degradation

1009125 Caffeine degradation

Pathway

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

データベース

仮説の構築

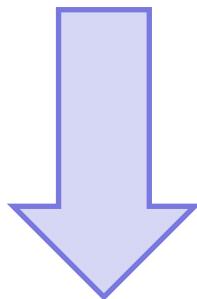
ソフトウェア

実験に代わる解析、
予測、設計へ

データサイエンス

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



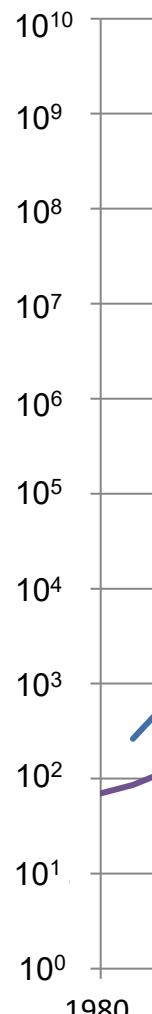
データベース、ソフトウェア
の研究は1970年代から

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

登録数



インターネット利用者数

ゲノム・遺伝子の配列

タンパク質の配列（遺伝子から自動変換）

GenBank

WGS

TrEMBL

SwissProt

PDB

PROSITE

タンパク質などの構造

タンパク質の配列（人手による注釈づけ）

データベースの話は、あとで改めてしまいます

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

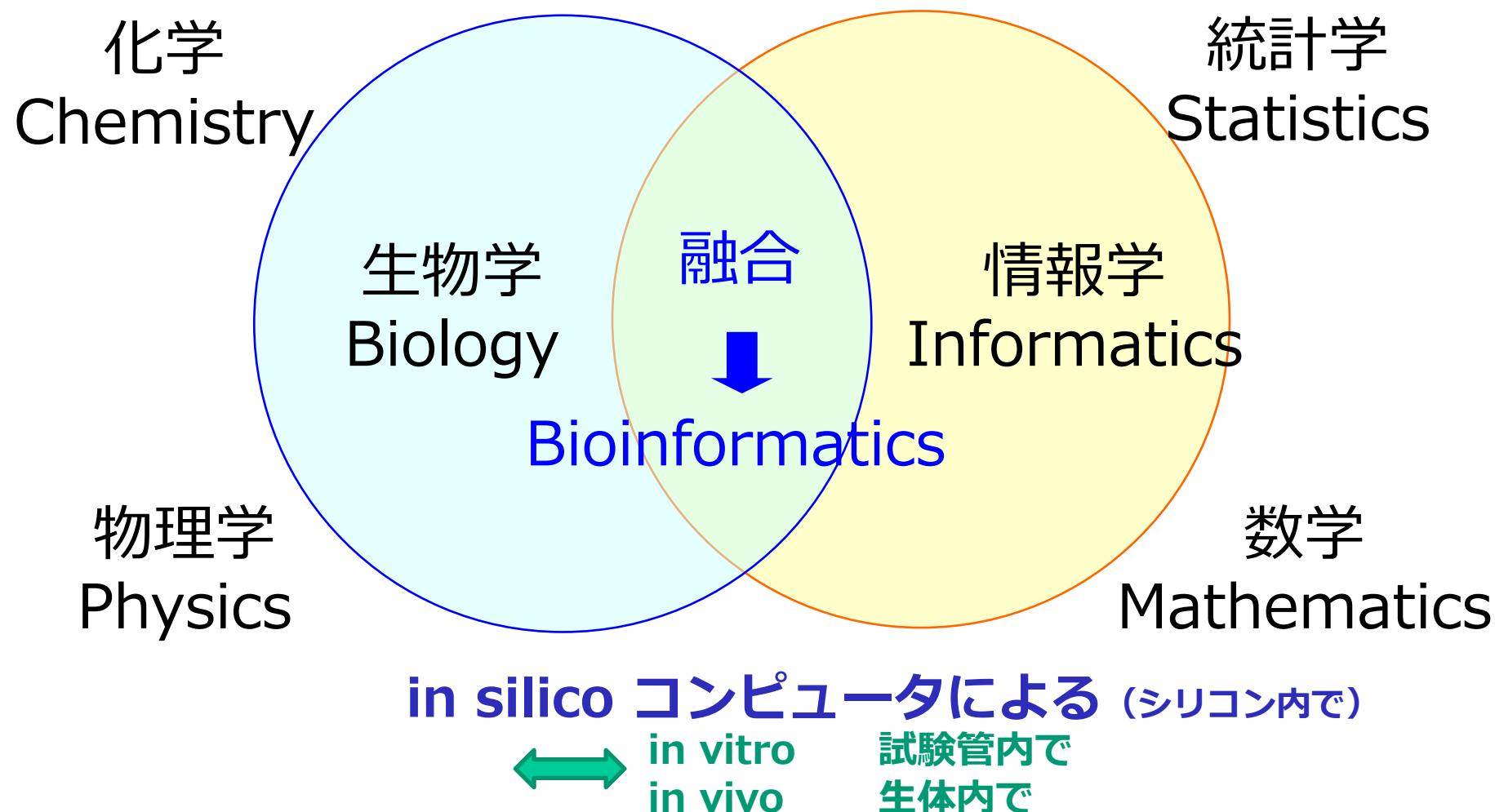
- パーソナルヘルスデータ
 - ウェアラブルデバイスや健康管理アプリから得られるデータは1人あたり数MBから数GB/日程度のデータ量
 - 数百万人規模のコホート研究や国民健康モニタリングプロジェクトでは 数PB 規模のデータが生成・保管されている
- バイオイメージングデータ
 - 1回のMRIスキャンで数百MB～数GBのデータが生成される
 - 顕微鏡画像やタイムラプス／3D撮像などでは、1実験で 数十～数百GB
 - 多数の症例や多数の時間点を含む研究で 数十～数百TB に拡大
 - イメージングデータは、分子イメージング、植物のイメージング、微生物の観察、生態系モニタリングなど多岐にわたる
- 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

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}	$2^{80}=1,208,925,819,614,629,174,706,176$	秭	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}	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。		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

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



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

生成系AI
ビッグデータ 機械学習
次世代スーパーコンピュータ
クラウドコンピューティング
Google検索 Wikipedia

DVD
World Wide Web
CD-ROM 携帯電話
スーパーコンピュータ
商用パーソナルコンピュータ
関係データベース ARPANET
UNIX

人工知能、数値計算
シャノン「通信の数学的理論」

プログラム内蔵型コンピュータ
チューリングマシン

2025
2000
1975
1950
1925

分子生物学

AlphaFold
ゲノム編集 合成生物学
メタゲノム
次世代シーケンサー
ES細胞、iPS細胞
ヒトゲノムの解読
クローン羊ドリー
DNAマイクロアレイ

PCR法
塩基配列データベース

遺伝子工学の特許
組換えDNA
コドン
形質導入
DNA構造の発見

遺伝子の本体がDNAであることを証明
一遺伝子一酵素説
ウィーバーによる分子生物学の提唱
合成抗生物質
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らっている
- ・ 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	B1O.CPP

PubMedの論文検索ヒット数

件数

PubMed: 生命科学や生物医学に関する文献検索エンジン

