#1
Introduction

Topics:

- What is Bioinformatics?
- Life is a huge hierarchical system
- DNA sequence, Chromosome, and Genome
- Sequence analysis

この授業の狙い

バイオインフォマティクス（生命情報科学）

① 情報科学の諸手法が、様々な応用分野でどう活かされるのかについて、活発に研究が進む当分野を「一つの実例」として学ぶ。

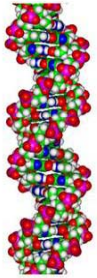
② バイオインフォマティクスの基礎を学ぶ。

③ ２１世紀の必修科目であるバイオについて、情報工学者として知るべき程度に学ぶ。

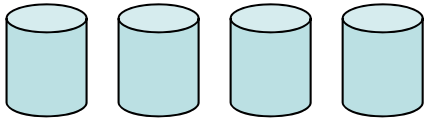
様々な理由により （大学院GP, グローバルCOEなどの指定科目になったことなど）
学部向けに実施していた「生命情報解析（旧：生命知識論第一）」の内容に近づけます。
学部時代に履修済みの方には物足りないかも知れません。
また上記の理由により、英語ではなく今年は日本語で説明する予定です。

バイオインフォマティクス（生命情報科学, *Bioinformatics*）

生物学実験の結果



データベースへの格納・整理

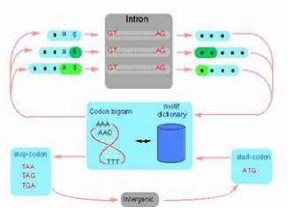


ルール抽出、クラスタ化、相同性

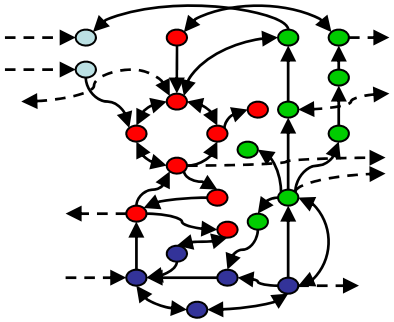
“計算機で扱い易いモデル”

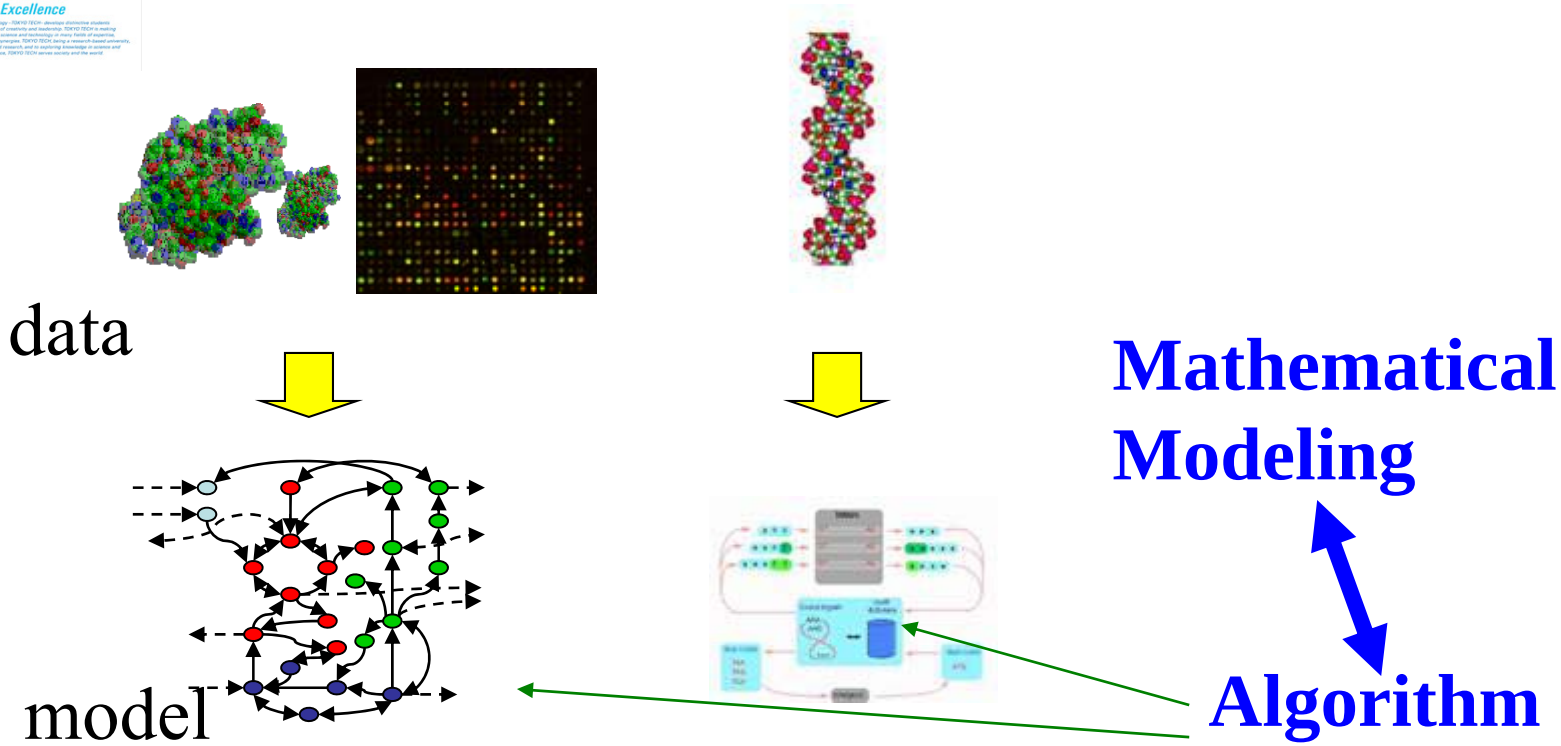
シミュレーション、量的予測
いわゆる *in silico* 生物学

実験生物学のIT化



バイオインフォマティクス





単なるデータ処理ではなく、データを数理モデルに変換する過程が重要。変換の方法を見いだすことがBIの中心課題。
簡潔なモデルほど、アルゴリズムは汎用的かつ効率的だが、生物の実態から離れるほど単純化しては何も意味がない。

情報科学の知識と、生物科学の知識の融合が必要

バイオインフォマティクスへの期待

クリントン−ブレア： ヒトゲノム配列読取終了宣言

『人類が作った最も重要な地図』（2000年6月26日）

→ 遺伝子の**意味解釈**の国際競争が激化。

⇩

バイオインフォマティクス（生命情報科学）

- **期待1（高速性）**

  膨大情報からいち早く産業的価値を発見、特許化。

- **期待2（経済性、安全性、倫理性）**

  コンピュータで候補絞り込み、試行錯誤の実験減らす。

バイオインフォマティクスの学際性

バイオインフォマティクス

- 目的意識は: 生物学
- 方法論は: 情報科学、数学、物理学、化学、システム工学、機械工学、など。

目的からは、バイオ分野に分類されるべきだが、
方法論は情報科学を中心とした学際的手法である。

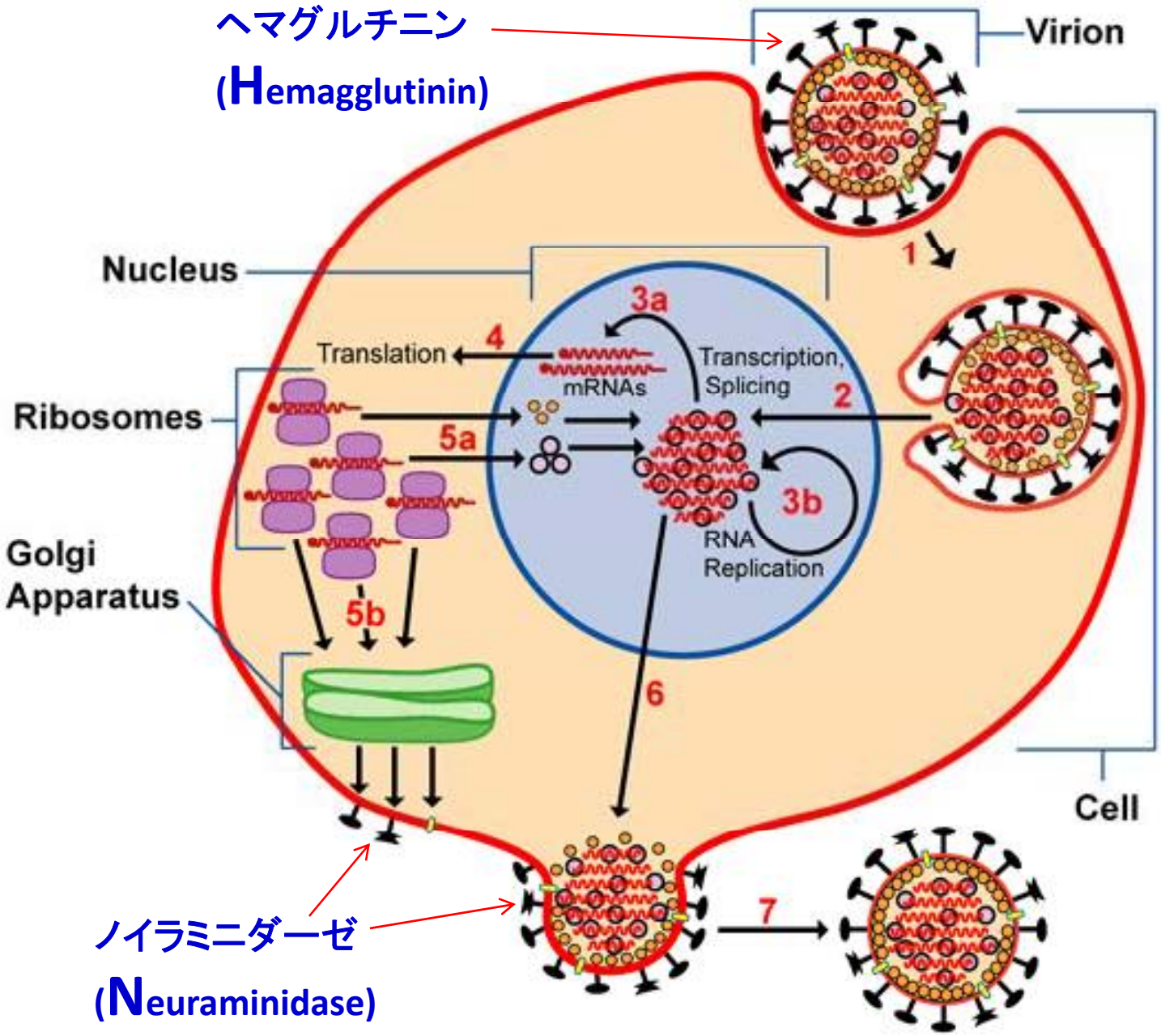
バイオインフォマティクス　世界の常識

1. 学際的な人材を結集すべし
2. 既存の実験生物学とは完全に独立した組織に

イントロダクション

（蓄積されるバイオ情報）

ザナミビル （商品名 リレンザ）



From NCBI "Entrez Genome" WWW page (Viral Genomes)

Entrez cross-database search - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/sites/gquery
NIHの統合データベース検索
Entrez （フランス語で玄関の意）
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
GO
Clear
Help
influenza virus neuraminidase
PubMed: biomedical literature citations and abstr
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
OMIA: online Mendelian Inheritance in Animals
Nucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
dbGaP: genotype and phenotype
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data

NIHの統合データベース検索

Entrez cross-database search - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/gquery

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Google Entrez 検索 ブックマーク

Entrez cross-database search

NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases influenza virus neuraminidase GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

3259	**PubMed:** biomedical literature citations and abstracts	37	**Books:** online books
2355	**PubMed Central:** free, full text journal articles	none	**OMIM:** online Mendelian Inheritance in Man
6	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals

1127	**Nucleotide:** Core subset of nucleotide sequence records	3	**dbGaP:** genotype and phenotype
none	**EST:** Expressed Sequence Tag records	none	**UniGene:** gene-oriented clusters of transcript sequences
none	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
1424	**Protein:** sequence database	509	**3D Domains:** domains from Entrez Structure
2	**Genome:** whole genome sequences	none	**UniSTS:** markers and mapping data
82	**Structure:** three-dimensional macromolecular structures	10	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	3	**GEO Profiles:** expression and molecular

http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=search&term=influenza%20virus%20neuraminic

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

ゲノム情報　データベース

influenza virus neuraminidase - Genome Resu...

NCBI

Entrez Genome

My NCBI [Sign In] [Reg

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Boo

Search Genome for influenza virus neuraminidase Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Send to

All: 2

Items 1 - 2 of 2

One p

About Entrez

Entrez Genome
Help

Submiting
Genome Project
Genome sequence

Microbial
Genome Projects
PDB neighbors

Genomic BLAST
Microbial
Eukaryotic

Archaea
Chromosome
Plasmid
DraftAssembly

Bacteria
Chromosome
Plasmid

1: NC_002209
Influenza B virus RNA 6, complete sequence
ssRNA; linear; Length: 1,557 nt
Replicon Type: viral segment
Replicon Name: RNA 6
Created: 1993/08/02

2: NC_002018
Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 6, complete sequence
ssRNA; linear; Length: 1,413 nt
Replicon Type: viral segment
Replicon Name: segment 6
Created: 1993/08/02

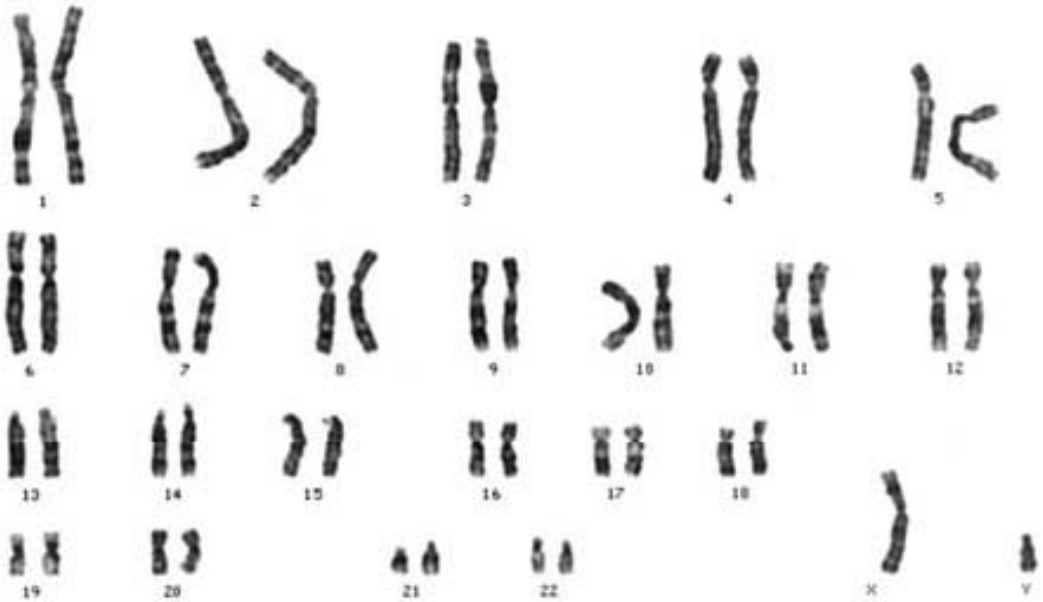
インターネット

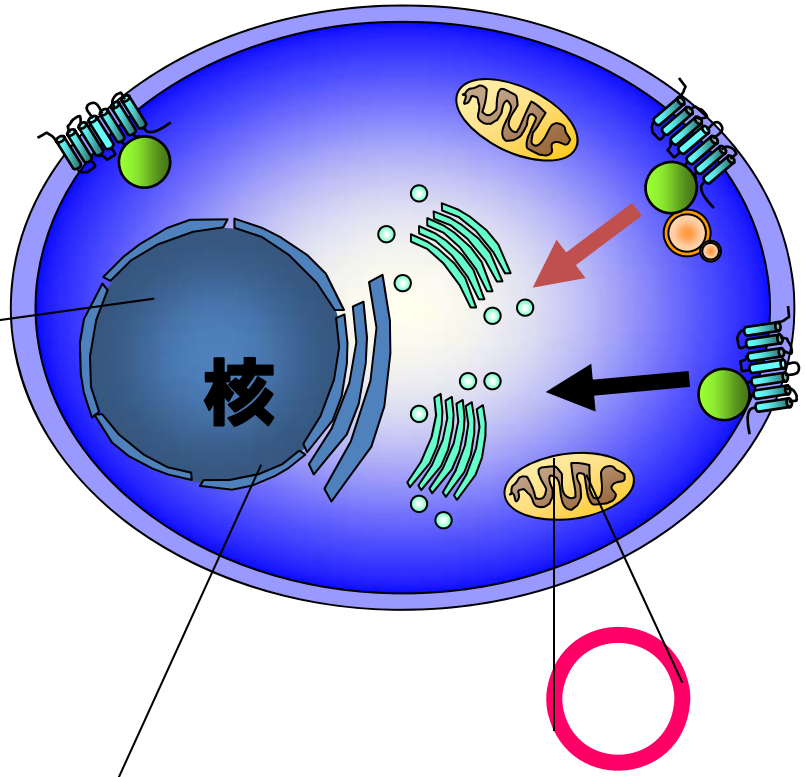
ゲノムとは

genome = （gene） + (~ome)
遺伝子　　総体
その個体における遺伝情報の総体

細胞

核 ゲノム



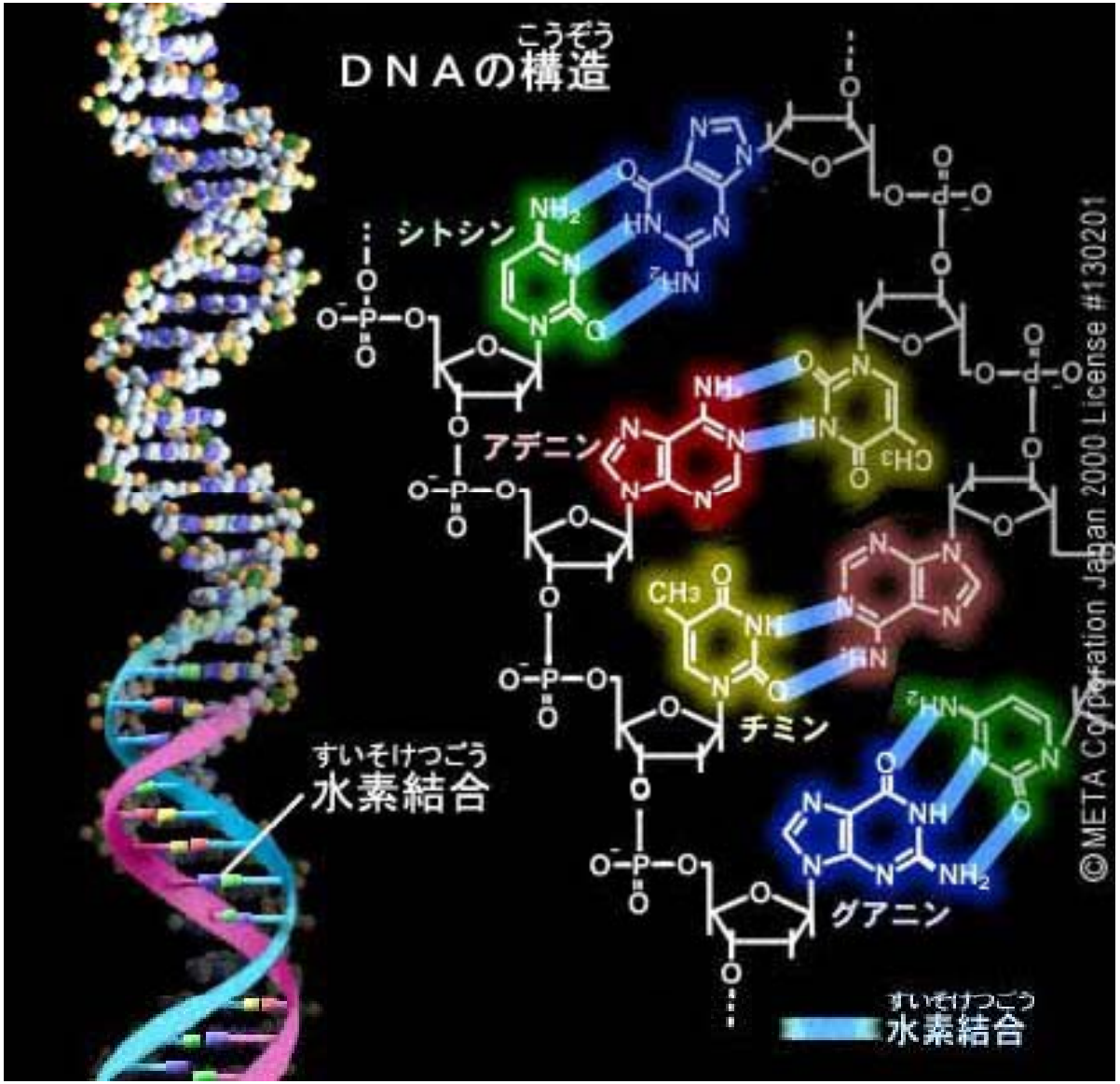


ミトコンドリアゲノム

母方からのみ遺伝

ヒトの場合、 23対の染色体
父母から、1対ずつ遺伝

遺伝情報の格納庫： DNA分子



4種類の塩基

C: Cytosine（シトシン）

A: Adenine（アデニン）

T: Thymine（チミン）

G: Guanine（グアニン）

相補的な塩基
（水素結合）

A・・・T

G・・・C

出典：IPA「教育用画像素材集サイト」http://www2.edu.ipa.go.jp/gz/

相補的塩基と水素結合（まとめ）

A ＝ T

水素結合が2本

G ≡ C

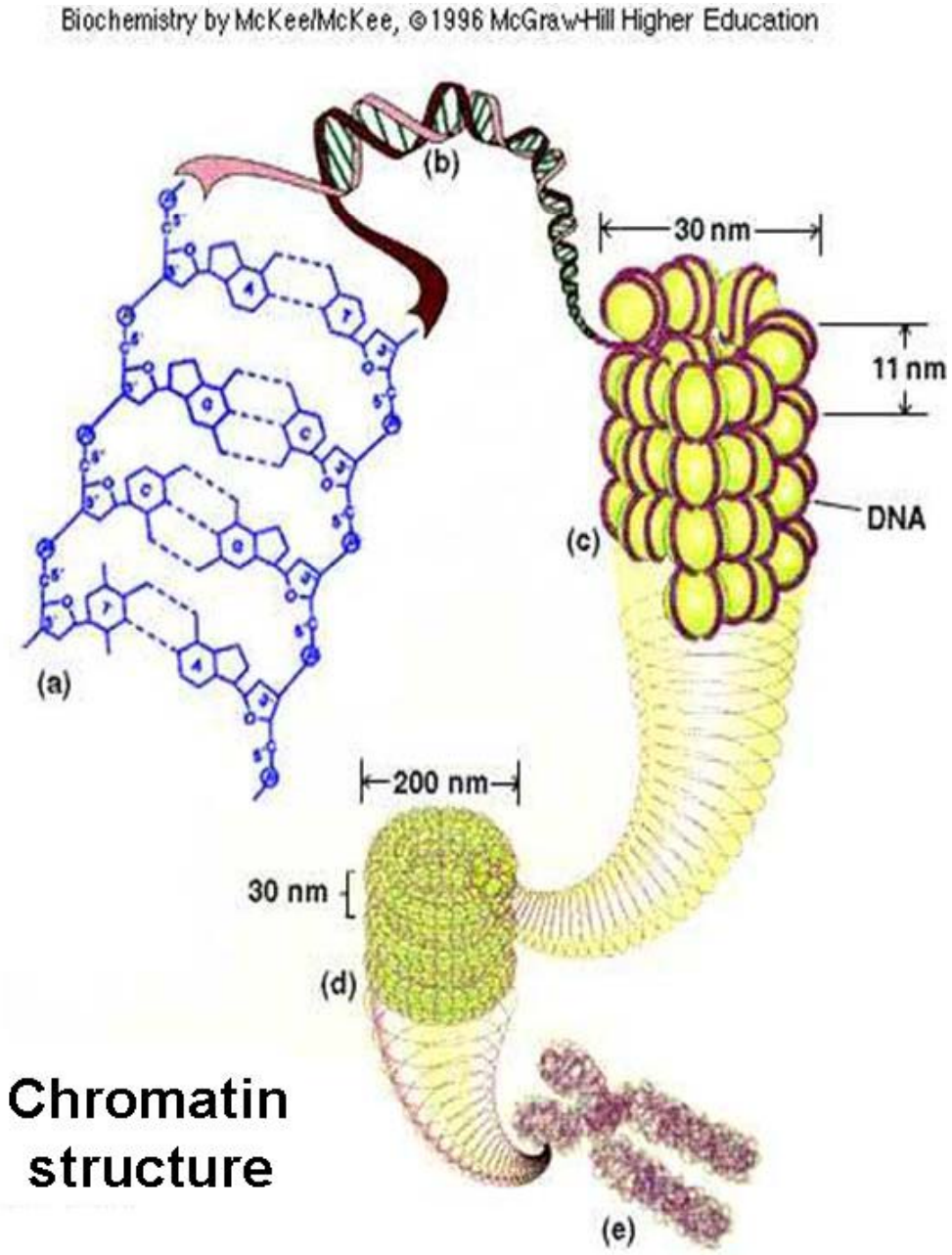
水素結合が3本
（より強固な結合）

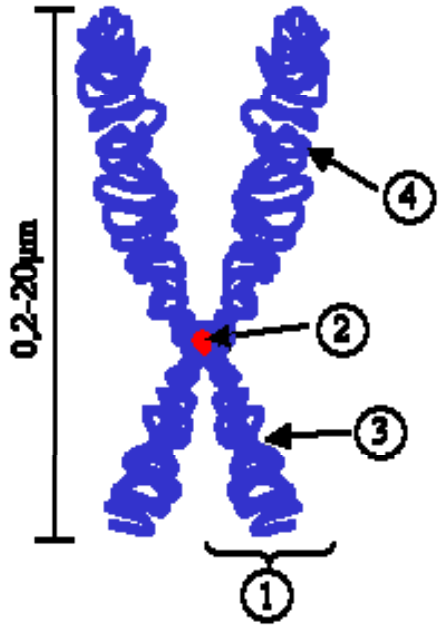
プリン塩基
（少し大きい）

ピリミジン塩基
（少し小さい）

物理構造の制約に起因する、
配列情報上の「弱い周期性」
は存在するのか？

クロマチン構造





Genome Home – Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Google Entrez genome 検索 ブックマーク

ページ(P) ツール(O)

ゲノム情報　データベース

Entrez　genome

NCBI

Entrez Genome

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Genome
Help

Submiting
Genome Project
Genome sequence

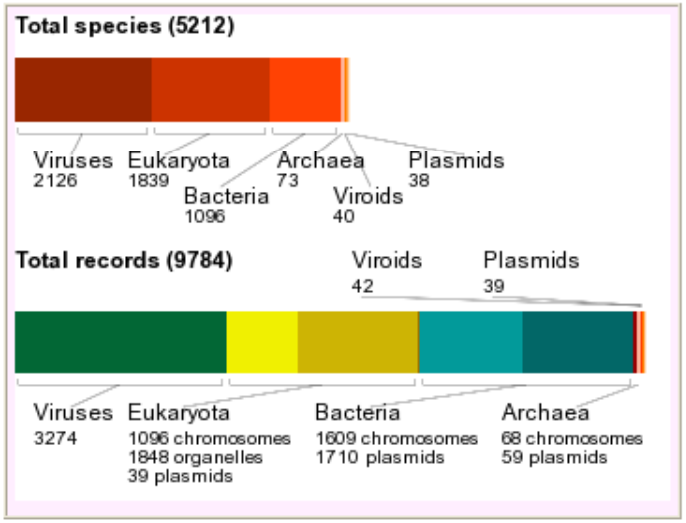
Microbial
Genome Projects
PDB neighbors

Genomic BLAST
Microbial
Eukaryotic

Archaea
Chromosome
Plasmid
DraftAssembly

Bacteria
Chromosome
Plasmid

The Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps. The database is organized in six major organism groups: Archaea, Bacteria, Eukaryotae, Viruses, Viroids, and Plasmids and includes complete chromosomes, organelles and plasmids as well as draft genome assemblies.



Related resources

Entrez Genome Project
complete and incomplete large-scale sequencing projects

Entrez Protein Clusters
a collection of related protein sequences

Eukaryotic
genome projects and sequences

Genomes of Bacillus anthracis
reference genome and related sequences

Influenza Virus Resource
sequence database and analyses

Microbial Genomes
reference sequences and resources

Organelle
reference sequences and

インターネット 100%

スタート | 3 Internet Explorer | 音響学会発表 [互... | Influenza | 9:11

Genomes - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239&opt=Virus&sort=genome

ゲノム情報　データベース

NCBI

Entrez Genome

HOME | SEARCH | SITE MAP | Viral Genomes Home | Taxonomy groups | All viruses | All viroids | Help | Contact us

Viruses - 2193 complete genomes

Sequence Info

- *Deltavirus* [1]
- *Retro-transcribing viruses* [98]
- *Satellites* [124]
- *dsDNA viruses, no RNA stage* [698]
- *dsRNA viruses* [113]
- *ssDNA viruses* [353]
- *ssRNA viruses* [754]
- *unclassified archaeal viruses* [4]
- *unclassified phages* [36]
- *unclassified virophages* [1]
- *unclassified viruses* [11]

Genome	Accession	Source information	Segm	Length	Protein	Date	Nb
⊞ Abaca bunchy top virus		isolate:Q767	6	6422 nt	5	02/06/2008	
Abalone shriveling syndrome-associated virus	NC_011646		-	34952 nt	31	12/09/2008	
Abelson murine leukemia virus	NC_001499		-	5894 nt	3	01/21/1998	
⊞ Abutilon mosaic virus			2	5217 nt	6	11/12/1990	
Acanthamoeba polyphaga mimivirus	NC_006450	isolate:Rowbotham-Bradford	-	1181404 nt	911	11/18/2004	
Acanthocystis turfacea Chlorella virus 1	NC_008724	strain:ATCV-1	-	288047 nt	860	12/27/2006	
Acheta domesticus densovirus	NC_004290		-	5234 nt	5	02/01/2002	
Acholeplasma phage L2	NC_001447		-	11965 nt	14	06/11/1993	
Acholeplasma phage MV-L1	NC_001341	strain:JA-1	-	4491 nt	4	07/02/1991	
Acidianus bottle-shaped virus	NC_009452		-	23814 nt	57	05/11/2007	
Acidianus filamentous virus 1	NC_005830		-	20869 nt	40	03/23/2004	
Acidianus filamentous virus 2	NC_009884		-	31787 nt	52	04/01/2005	
Acidianus filamentous virus 3	NC_010155		-	40449 nt	68	09/16/2007	
Acidianus filamentous virus 6	NC_010152		-	39577 nt	66	09/16/2007	
Acidianus filamentous virus 7	NC_010153		-	36895 nt	57	09/16/2007	
Acidianus filamentous virus 8	NC_010154		-	38179 nt	61	09/16/2007	

Genomes – Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=5&name=Viruses

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Google Entrez genome 検索

ゲノム情報　データベース

Genomes

(H3N2))								
Influenza A virus (A/New York/392/2004 (H3N2))	segment 8	NC 007370	890 nt	2	0	2	Aug 30 2005	Mar 30 2006
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 1	NC 002023	2341 nt	1	0	1	Nov 3 1982	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 2	NC 002021	2341 nt	2	0	2	Aug 2 1993	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 3	NC 002022	2233 nt	1	0	1	Nov 3 1982	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 4	NC 002017	1778 nt	1	0	1	Jul 13 1983	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 5	NC 002019	1565 nt	1	0	1	Aug 2 1993	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 6	NC 002018	1413 nt	1	0	1	Aug 2 1993	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 7	NC 002016	1027 nt	2	0	2	Jun 9 1982	Jul 10 2008
Influenza A virus (A/Puerto Rico/8/34 (H1N1))	segment 8	NC 002020	890 nt	2	0	2	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 1	NC 002204	2368 nt	1	0	1	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 2	NC 002205	2313 nt	1	0	1	Nov 10 1999	Jul 10 2008
Influenza B virus	RNA 3	NC 002206	2204 nt	1	0	1	Nov 10 1999	Jul 10 2008
Influenza B virus	RNA 4	NC 002207	1882 nt	1	0	1	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 5	NC 002208	1841 nt	1	0	1	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 6	NC 002209	1557 nt	2	0	2	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 7	NC 002210	1191 nt	2	0	2	Aug 2 1993	Jul 10 2008
Influenza B virus	RNA 8	NC 002211	1096 nt	2	0	2	Aug 2 1993	Jul 10 2008
Influenza C virus	7	NC 006306	935 nt	2	0	2	Sep 24 2004	Nov 26 2008
Influenza C virus	segment 1	NC 006307	2325 nt	1	0	1	Sep 24 2004	Nov 26 2008
Influenza C virus	segment 2	NC 006308	2265 nt	1	0	1	Sep 24 2004	Nov 26 2008
Influenza C virus	segment 3	NC 006309	2130 nt	1	0	1	Sep 24 2004	Nov 26 2008
Influenza C virus	segment 4	NC 006310	1968 nt	1	0	1	Sep 24 2004	Nov 26 2008
Influenza C virus	segment 5	NC 006311	1807 nt	1	0	1	Sep 24 2004	Jul 10 2008
Influenza C virus	segment 6	NC 006312	1125 nt	2	0	2	Sep 24 2004	Nov 26 2008
Invertebrate iridescent virus 3		NC 008187	191100 nt	126	0	126	Jun 20 2006	Oct 3 2007
Invertebrate iridescent virus 6		NC 003038	212482 nt	468	0	463	Jul 31 2001	Mar 30 2006

検索: 次の結果 前へ すべてハイライト表示

NC_002209.1: Influenza B virus RNA 6, complete sequence – Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/nuccore/8486156?report=graph&log$=seqview

NCBI Reference Sequence: NC_002209.1

Influenza B virus RNA 6, complete sequence

NC_002209.1 (1557 bases)

Sequence | Set Origin | Views & Tools | Markers | Search...

1 100 200 300 400 500 600 700 800 900 1 K 1,100 1,200 1,300 1,400 1,557

1 - 1558 (1557 bases shown, positive strand)

Sequence | Flip Strands | Tools | Markers | Details | Options

1 100 200 300 400 500 600 700 800 900 1 K 1,100 1,200 1,300 1,400 1,557

NC_002209.1: Influenza B virus RNA 6, complete sequence

Gene Models:

NA

NP_056663.1: neuraminidase

NB

NP_056662.1

All Other Features:

Shaw,M.W. (1983) Proc. Natl. Acad. Sci. ...

RNA6のほぼ全長がneuraminidase

NCBI Sequence Viewer v2.0 - Windows Internet Explorer

C:¥Documents and Settings¥Administrator¥デスクトップ¥NC002209.mht

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

核酸（DNA/RNA）配列データベース

NCBI Sequence Viewer v2.0

```
LOCUS       NC_002209               1557 bp ss-RNA     linear   VRL 10-JUL-2008
DEFINITION  Influenza B virus RNA 6, complete sequence.
ACCESSION   NC_002209
VERSION     NC_002209.1  GI:8486156
PROJECT     GenomeProject:14656
DBLINK      Project:14656
KEYWORDS    .
SOURCE      Influenza B virus
  ORGANISM  Influenza B virus
            Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
            Influenzavirus B.
REFERENCE   1  (bases 1 to 358)
  AUTHORS   Shaw,M.W., Choppin,P.W. and Lamb,R.A.
  TITLE     A previously unrecognized influenza B virus glycoprotein from a
            bicistronic mRNA that also encodes the viral neuraminidase
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 80 (16), 4879-4883 (1983)
   PUBMED   6308656
REFERENCE   2  (bases 1 to 1557)
  AUTHORS   Shaw,M.W., Lamb,R.A., Erickson,B.W., Briedis,D.J. and Choppin,P.W.
  TITLE     Complete nucleotide sequence of the neuraminidase gene of influenza
            B virus
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 79 (22), 6817-6821 (1982)
   PUBMED   6294654
REFERENCE   3  (bases 1 to 1557)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (12-JUN-2000) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from J02095.
            Sequence flanked on 5' end by cellular 'cap transfer' DNA
            (AGACACACGC). Neuraminidase cds determined by consideration of open
            reading frames, and mw and composition of the protein product. The
            second open reading frame codes for a protein (nb) found only in
```

Protein – neuraminidase [Influenza B virus] – Windows Internet Explorer

C:¥Documents and Settings¥Administrator¥デスクトップ¥NP056663.mht

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

タンパク質配列データベース

Protein BLAST: search p... | influenza virus neuramini... | Protein – neuraminida...

NCBI Reference Sequence: NP_056663.1

neuraminidase [Influenza B virus]

Comment Features Sequence

```
LOCUS       NP_056663                466 aa            linear   VRL 10-JUL-2008
DEFINITION  neuraminidase [Influenza B virus].
ACCESSION   NP_056663
VERSION     NP_056663.1  GI:8486158
DBSOURCE    REFSEQ: accession NC_002209.1
KEYWORDS    .
SOURCE      Influenza B virus
  ORGANISM  Influenza B virus
            Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
            Influenzavirus B.
REFERENCE   1  (residues 1 to 466)
  AUTHORS   Shaw,M.W., Lamb,R.A., Erickson,B.W., Briedis,D.J. and Choppin,P.W.
  TITLE     Complete nucleotide sequence of the neuraminidase gene of influenza
            B virus
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 79 (22), 6817-6821 (1982)
   PUBMED   6294654
REFERENCE   2  (residues 1 to 466)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (12-JUN-2000) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAA43749.
            Sequence flanked on 5' end by cellular 'cap transfer' DNA
            (AGACACACGC). Neuraminidase cds determined by consideration of open
            reading frames, and mw and composition of the protein product. The
            second open reading frame codes for a protein (nb) found only in
```

Change Region Shown

Customize View

Articles about NA

- A previously unrecognized influenza B vi glycoprot [Proc Natl Acad Sci U S A. 19
- Complete nucleotide sequence of the neuramin [Proc Natl Acad Sci U S A. 19

» See a

More about the gene NA

Also Known As: FLUBVsRNA-6gp2

Recent Activity

Turn Off Clea

Chain A, The 2.2 Angstroms Resolution Crystal Structure Of

Chain A, Influenza B Virus Neuraminidase Complexed With 1-

gi|8486158 (466 letters)

gi|8486158 (466 letters)

A previously unrecognized influenza B virus glycoprotein from a

otein - neuraminidase [Influenza B virus] - Windows Internet Explorer

mhtml:file://C:¥Documents%20and%20Settings¥Administrator¥デスクトップ¥NP056663.mht

ル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

定期更新されている関連情報リンク

Protein - neuraminidase [Influenza B virus]

(AGACACACGC). Neuraminidase cds determined by consideration of open reading frames, and mw and composition of the protein product. The second open reading frame codes for a protein (nb) found only in type b influenzas Its functional significance is not yet known. Method: conceptual translation.

```
ATURES             Location/Qualifiers
    source         1..466
                   /organism="Influenza B virus"
                   /strain="B/Lee/40"
                   /db_xref="taxon:11520"
                   /segment="RNA 6"
    Protein        1..466
                   /product="neuraminidase"
                   /calculated_mol_wt=51311
    Region         103..441
                   /region_name="Sialidase"
                   /note="Sialidases or neuraminidases function to bind and
                   hydrolyze terminal sialic acid residues from various
                   glycoconjugates as well as playing roles in pathogenesis,
                   bacterial nutrition and cellular interactions. They have a
                   six-bladed, beta-propeller fold...; cd00260"
                   /db_xref="CDD:29333"
    Site           order(116,149,276,292,374,409,428)
                   /site_type="other"
                   /note="catalytic site"
                   /db_xref="CDD:29333"
    Site           order(256..258,260..262)
                   /site_type="other"
                   /note="Asp-box motif"
                   /db_xref="CDD:29333"
    CDS            1..466
                   /gene="NA"
                   /locus_tag="FLUBVsRNA-6gp2"
                   /coded_by="NC_002209.1:54..1454"
                   /db_xref="GeneID:956541"
```

B virus glycoprotein from a

All links from this record

- BLink
- Conserved Domains
- BioAssay by target,identical sequence
- Full text in PMC
- Gene
- Identical Proteins
- Nucleotide
- Protein (UniProtKB)
- PubMed (RefSeq)
- PubMed (Weighted)
- Related Structure
- Related Sequences
- Domain Relatives
- Genome
- PubMed
- Taxonomy
- LinkOut

Structure Summary, 3CKZ, 64579 - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

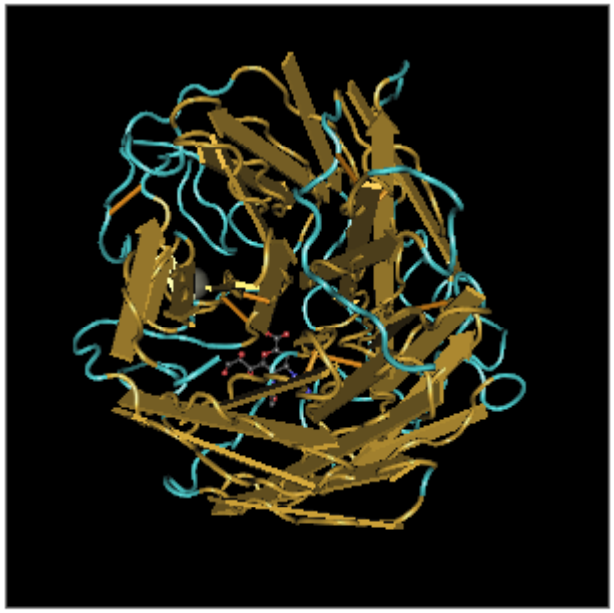
タンパク質立体構造データベース

NCBI Sequence Viewer v... | Structure Summary, ... | Microsoft Application Se...

NCBI

Structure Summary
MMDB

HOME | SEARCH | SITE MAP | Entrez | Structure | Protein | CDD | PubMed | Taxonomy | PubChem | Help | Cn3D



MMDB ID: 64579 **PDB ID:** 3CKZ

Search PDB or MMDB

Reference: Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker PA, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ *Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants* Nature v453, p.1258-1261

The potential impact of pandemic influenza makes effective measures to limit the spread and morbidity of virus infection a public health priority. Antiviral drugs are seen as essential requirements for control of initial influenza outbreaks caused by a new virus, and in pre-pandemic plans there is a heavy reliance on drug stockpiles....

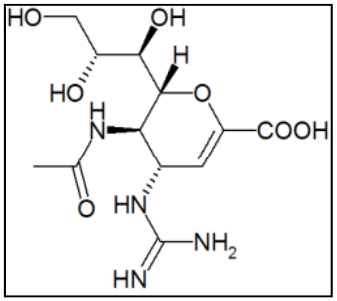
» View full abstract

Description: N1 Neuraminidase H274y + Zanamivir.

Deposition: 2008/3/18

Taxonomy: Influenza A virus

Related Structure: VAST



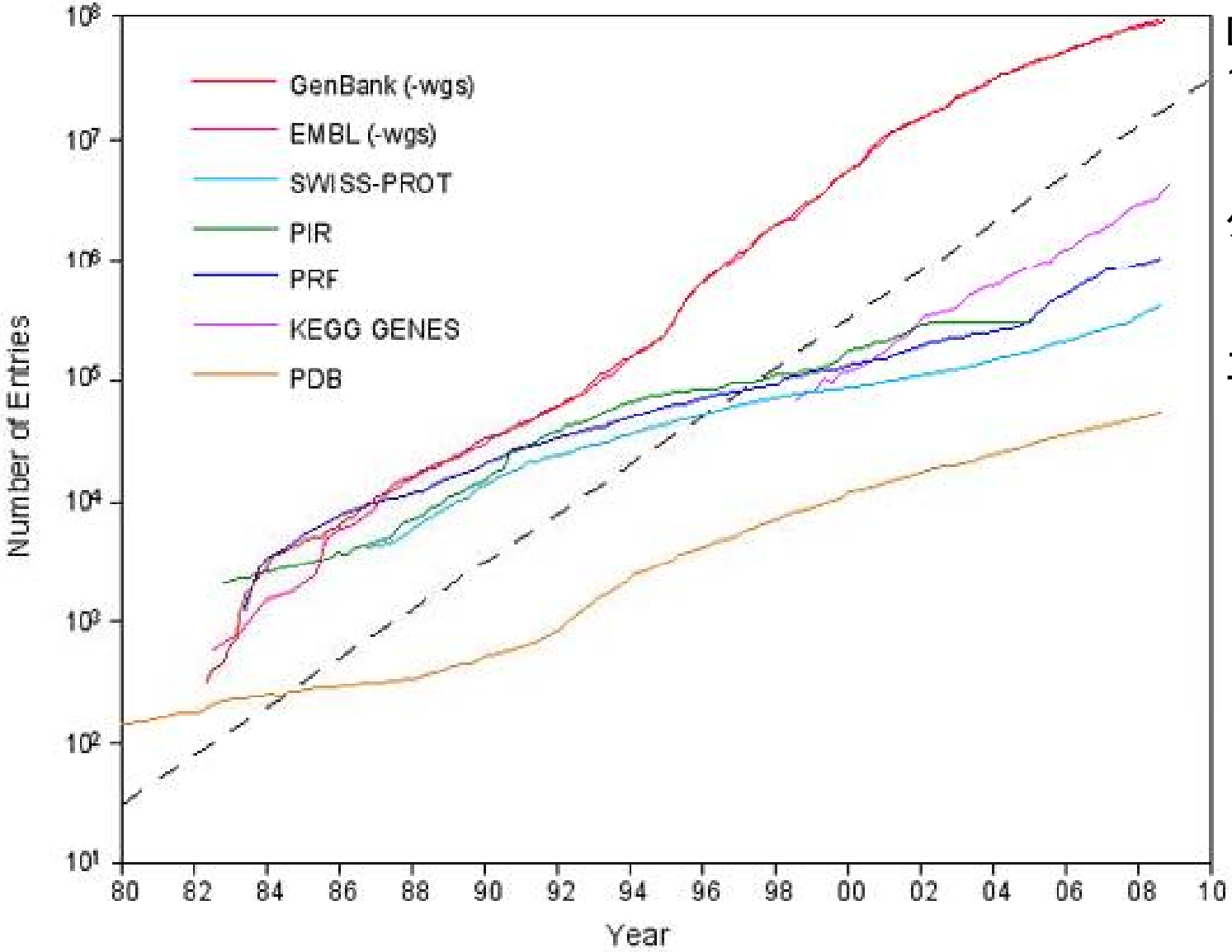
Structure View in Cn3D | Structure View in RasMol

Tasks: Display Drawing: All Atoms

Download Cn3D | View Cn3D Tutorial

Growth of Sequence and 3D Structure Databases

データベース登録数の増加



DNA配列
10年で約100倍

タンパク質配列
10年で約10倍

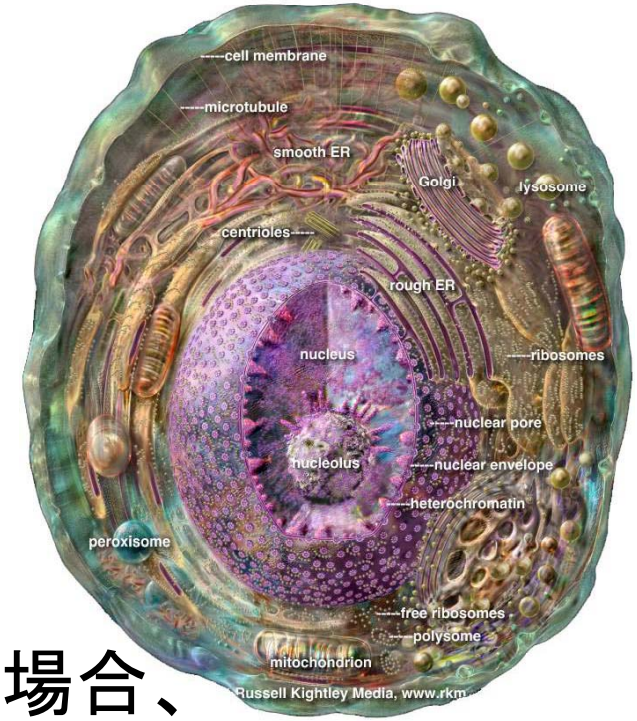
立体構造情報
10年で約10倍

件数ベース
（容量ベース
はより増大）

京都大学 GenomeNet WWW site　(Database Growth

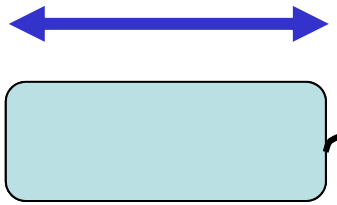
生物：階層的情報システムとしての理解

- 生体 (body)　ヒト個体には60兆個の細胞
- 器官系 (organ system)
- 器官 (organ)
- 組織 (tissue)
- 細胞 (cell)
- 細胞内小器官 (organelle)
- 分子 (molecule)



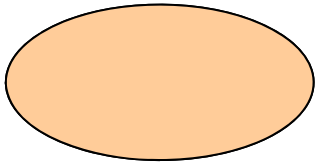
分子論的な理解を前提とする場合、
当面のターゲットは”巨大な”細胞系まで

細胞という“微小”システム



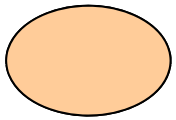
約200μm

ゾウリムシ



ヒトの卵細胞

10～30μm



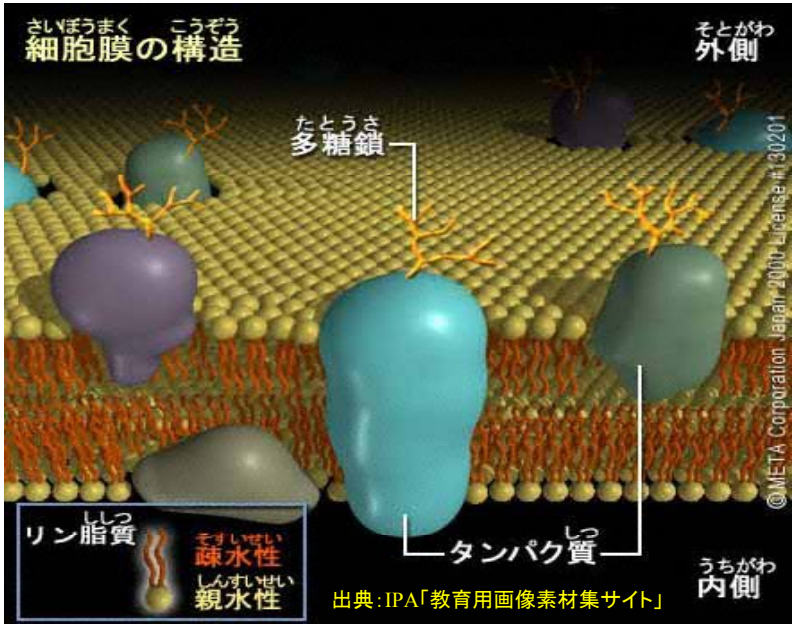
ヒトの体細胞

7～8μm

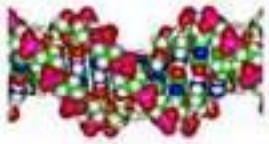
ヒトの赤血球細胞

1～4μm

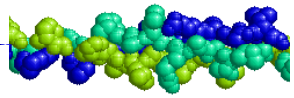
サルモネラ菌、大腸菌



細胞膜：　厚み60～80Å



DNA二重らせん
直径20Å
（ヒト全長約2m）



コラーゲン三本鎖
直径15Å

細胞という“巨大？”システム

スケールを1億倍してみると、

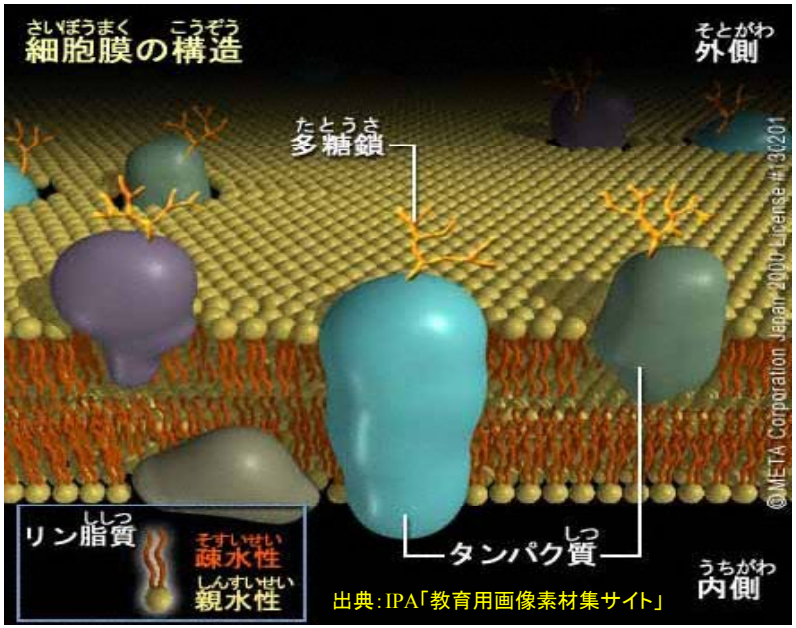
約20km

ゾウリムシ

ヒトの卵細胞

1～3km

ヒトの体細胞



細胞膜：　厚み60～80cm

700～800m

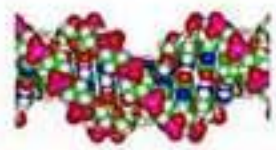
ヒトの赤血球細胞

100～400m

サルモネラ菌、大腸菌

U.S.S. Enterprise

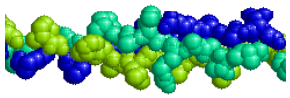
300 – 600 m



DNA二重らせん

直径20cm

（ヒト　全長約20万km ）



コラーゲン三本鎖

直径15cm

細胞システム

・構成部品のスケールから見て“巨大”。
・内部が複雑な部分構造に階層化。

・構成部品の種類が非常に豊富。
　ヒト：２．５万遺伝子、１０万タンパク？

・設計図が少しずつ違う多様な種が存在

・個体ごとにも設計図の差がある。

２１世紀の最先端の研究対象

しかし直接的な、物理化学的
全シミュレーションの道程は遠い

ヒト遺伝子配列の例（ALDH2）

ALDH2は、アセトアルデヒド脱水素酵素。
*2型（変異型）は酵素の効率が悪い。

```
ID   M20455; SV 1; linear; genomic DNA; STD; HUM; 135 BP.
XX
AC   M20455; J03043;
XX
DT   20-FEB-1989 (Rel. 18, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 8)
XX
DE   Human nucleus-encoded mitochondrial aldehyde dehydrogenase 2 (ALDH2) gene,
DE   exon 12.
XX
KW   aldehyde dehydrogenase; aldehyde:NAD+ oxidoreductase;
KW   nuclear-encoded mitochondrial EF-G-like protein.
XX
```

（中略）

```
FT   source          1..135
FT                   /organism="Homo sapiens"
FT                   /map="12q24.2"
```

・・・12番染色体上 24.2の位置にある

（中略）

```
FT   variation       114
FT                   /gene="ALDH2"
FT                   /note="g in ALDH2-1 allele; a in ALDH2-2 allele"
```

（中略）

```
SQ   Sequence 135 BP; 31 A; 25 C; 48 G; 31 T; 0 other;
     caaattacag ggtcaactgc tatgatgtgt ttggagccca gtcacccttt ggtggctaca      60
     agatgtcggg gagtggccgg gagttgggcg agtacgggct gcaggcatac actgaagtga     120
     aaactgtgag tgtgg                                                      135
//
```

g (*1型)
→ a (*2型)

日本人では、*1/*1型の人が56%、*1/*2が38%、*2/*2が4%。
大陸からの渡来系人が *2を持ってきたという説がある。

配列アラインメント

データベースを検索すると、マウス、牛などで、ヒトと良く似た配列が発見できる。
対応する2配列を、並べて比較したものを「アラインメント (alignment)」と呼ぶ。

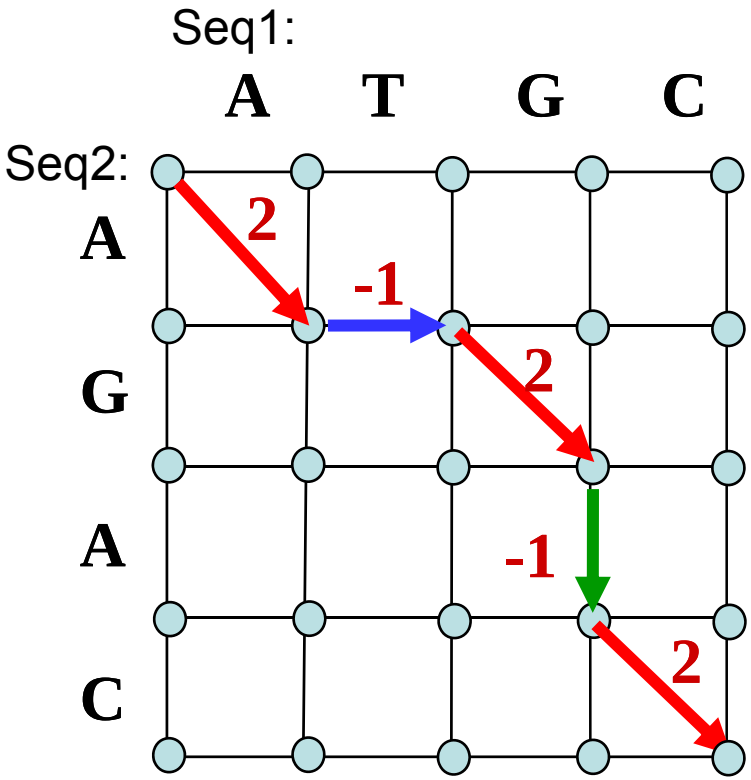
```
>BC116084|BC116084.1 Bos taurus similar to Aldehyde dehydrogenase,
            mitochondrial precursor (ALDH class 2) (ALDH1) (ALDH-E2),
            mRNA (cDNA clone MGC:138997 IMAGE:8094240), complete cds.
          Length = 2068

 Score =  127 bits (64), Expect = 6e-27
 Identities = 100/112 (89%)
 Strand = Plus / Plus


Query: 10    gggtcaactgctatgatgtgtttggagcccagtcacccttttggtggctacaagatgtcgg 69
             |||||||||||||||| |||||||| |||||||| || || ||||||||||||| |||| |
Sbjct: 1499  gggtcaactgctatgacgtgtttggggcccagtcgccgttcggtggctacaagctgtctg 1558


Query: 70    ggagtggccgggagttgggcgagtacgggctgcaggcatacactgaagtgaa 121
              |||||||||||||| |||| || |||||||||||||| ||||| ||||||||
Sbjct: 1559  ggagtggccgggagctggggggaatacgggctgcaggcgtacaccgaagtgaa 1610
```

以下は、第二回への予告編

ウシも*1型??

上記例では、Query（入力）はヒトのALDH2配列。
Sbjct が発見された配列で、この場合はウシ (学名 Bos taurus)の類似配列。

配列アラインメント



ギャップ(gap) を許す

下図のように文字間対応の無い場所のこと

$$2+(-1)+2+(-1)+2 = 4$$

gap

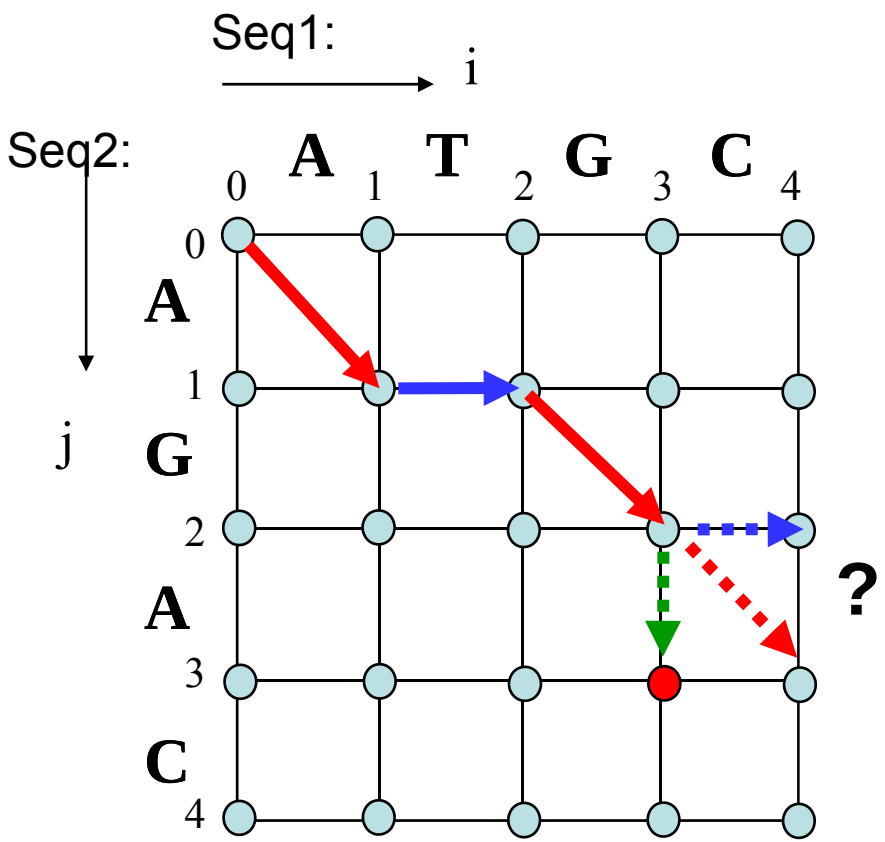
Seq1: A T G — C …

Seq2: A — G A C …

Seq1:	A	T	G	—	C	C
Seq2:	A	—	G	A	C	G
position	*1*	*2*	*3*	*4*	*5*	*6*

1, 3, 5: ***Conserved***
2: ***Deletion (Seq1 → Seq2)***
4: ***Insertion (Seq1 → Seq2)***
6: ***Mutation***

経路探索問題として考える



開始点： 左上(0,0)
終了点： 右下(m,n)

移動可能な方向は3方向のみ。
（後戻りできず、閉路はできない）

deletion（Seq1から見て）

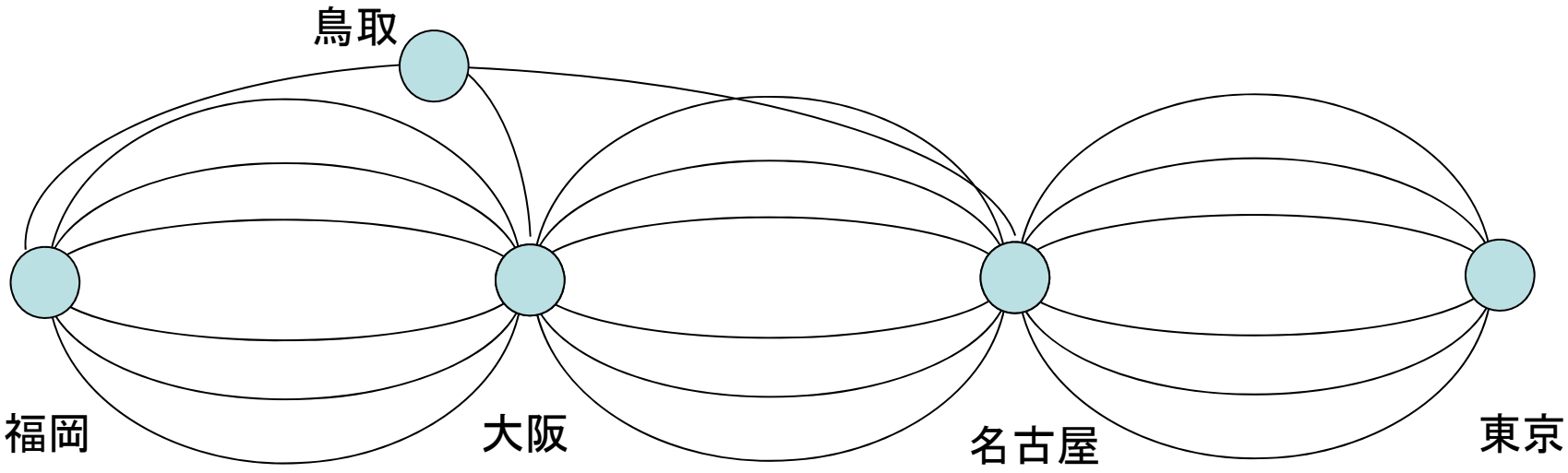
insertion（Seq1から見て）

conserved または
mutation

可能な全経路の中で、
スコアを最大化する経路を探索する。

単純な木探索プログラムを作ると、、、3^N ～ 3^{2N} かかってしまいそうだが・・

動的計画法
(Dynamic Programming: DP)



○ 大きな問題を、いくつかの部分問題に分割して解く。
（部分問題の解の和として、全体問題が解けるという線形性が条件）

○ 部分問題の解をメモリ上に蓄積しておくことにより、同じ計算を
なんども繰り返さなくて良いようにするのが特徴。
（これが速さの源泉）

ローカルアラインメントの計算量は？

Complexity of Local Alignment

(Seq1, Seq2)に対して、その任意の部分配列をとって、
スコア最大となる　アラインメントを探索しなければならない・・・

(もっともNaiveな方法を考えてみると)

Seq1

A T T G C T C G T T G G A ... (length m)

部分生成 ⇒ $O(\mathbf{m}^2)$ **sub-sequences**

Seq2

A A A A C C G T A A... (length n)

部分生成 ⇒ $O(\mathbf{n}^2)$ **sub-sequences**

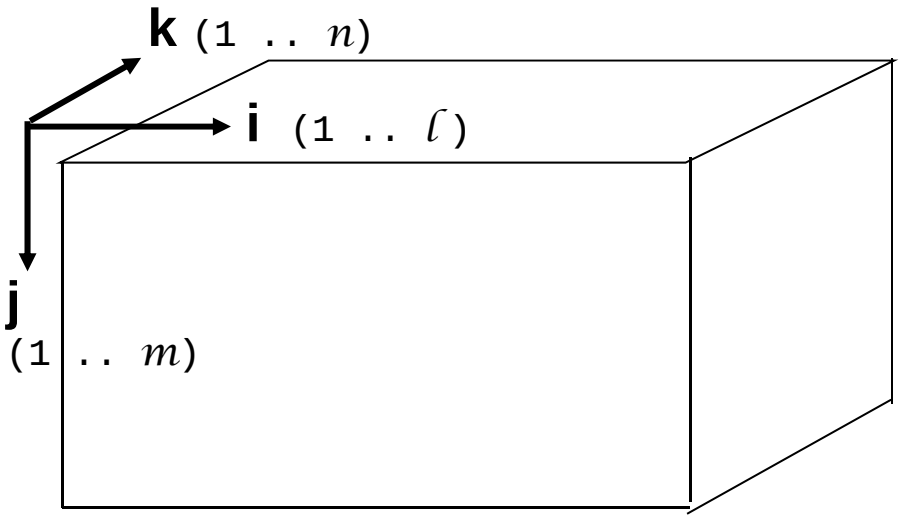
⇒ **Pairwise Alignment by Dynamic Programming**

$O(\mathbf{m\ n})$

⇑ これから最適を選ぶ？

Total: $O(\mathbf{m}^2) \times O(\mathbf{n}^2) \times O(\mathbf{m\ n})$ も必要なのか?!

DPによる多重アライメント



三次元の多重アラインメントの場合

Time complexity:
$O(\ell\ m\ n)$
Space complexity:
$O(\ell\ m\ n)$

$$
M[i, j, k] \leftarrow \max \begin{cases}
M[i-1, j-1, k-1] + \mathrm{Score}(Seq1(i), Seq2(j), Seq3(k)) \\
M[i, j-1, k-1] + \mathrm{Score}(\text{'-'}, Seq2(j), Seq3(k)) \\
M[i-1, j, k-1] + \mathrm{Score}(Seq1(i), \text{'-'}, Seq3(k)) \\
M[i-1, j-1, k] + \mathrm{Score}(Seq1(i), Seq2(j), \text{'-'}) \\
M[i, j, k-1] + \mathrm{Score}(\text{'-'}, \text{'-'}, Seq3(k)) \\
M[i-1, j, k] + \mathrm{Score}(Seq1(i), \text{'-'}, \text{'-'}) \\
M[i, j-1, k] + \mathrm{Score}(\text{'-'}, Seq2(j), \text{'-'})
\end{cases}
$$

注: N次元のDPを直接適用すると、、長さ L の場合は O (L^N) の時間・空間がかかる.