近畿大学・農学部・生命情報学

配列決定と バイオインフォマティクス概論

2009年4月7日（火）

奈良先端大･情報･蛋白質機能予測学講座
川端　猛
takawaba@is.naist.jp

http://isw3.naist.jp/IS/Kawabata-lab/home-ja.html

平成21年度「生命情報学＆生命情報学実習」講義日程

2009.3.17

	講義	生命情報学	演習	生命情報学演習
4/7	川端1	配列決定とバイオインフォマティクス		
4/14	川端2	ペアワイズアライメント	川端	主要WEBデータベースの使用法(BLAST)
4/21	川端3	配列相同性解析	中村	ChemOfficeを用いた計算化学演習
4/28	川端4	マルチプルアライメントとその応用		
5/12	川端5	分子系統学基礎	中村	系統樹作成演習(ClustalX)
5/19	川端6	立体構造データの情報解析	川端	蛋白質立体構造データの可視化(RasMol)
5/26	川端7	>>試験<<		
6/2	金谷1	ポストゲノム解析入門(トランスクリプトーム解析）		
6/9	金谷2	ポストゲノム解析入門(インタラクトーム解析）	金谷１	発現プロフィール解析演習
6/16	金谷3	ポストゲノム解析（統合解析）	金谷２	インタラクトローム・代謝物解析演習
6/23	金谷4	メタボローム解析（その１）		
6/30	金谷5	メタボローム解析（その２）		
7/7	金谷6	メタボローム解析（その３）		
7/14	金谷7	>>試験<<		

全ゲノム配列が決定された生物種

Apr 09,2009

生物種		完了		ドラフト配列	進行中
原核生物	古細菌	56	メタン菌、超好熱菌、高度好塩菌など	9	40
	真正細菌	806	大腸菌、乳酸菌、コレラ菌、結核菌、シアノバクテリアなど	773	741
真核生物	動物	4	ヒト、マウス、ショウジョウバエ、線虫	81	85
	植物	2	シロイヌナズナ、コメ	8	45
	真菌	10	出芽酵母、分裂酵母、カンジダなど	67	41
	原生生物	6	マラリア原虫、赤痢アメーバなど	24	25
合計		884		963	980

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.htmlから転載

全ゲノムが解読された主な生物種(発表年代順)

発表年	生物種	ゲノムサイズ（M(10^6)塩基対)	遺伝子数
1995	マイコプラズマ菌 (*Mycoplasma genitalium*)	0.6	467
	インフルエンザ菌 (*Haemophilus influenzae*)	1.8	1717
1997	出芽酵母 (*Saccharomyces cerevisiae*)	12.1	6140
	大腸菌(*Escherichia coli*)	4.6	4289
1998	線虫(*Caenorhabditis elegans*)	97.0	19099
2002	マウス(*Mus musculus*)	2625.0	25865
2003	ヒト(*Homo sapiens*)	3068.0	26626

・一番小さいマイコプラズマでも0.6x10^6=60万文字の{A,T,G,C}
・フロッピーディスク：1.2M, CD-ROM：600M, DVD：4000Mなので、バクテリアゲノムはフロッピー数枚、ヒトゲノムはDVDに収納可能

今日の講義の内容

分子生物学で扱うデータ(DNA配列、アミノ酸配列)について

（1）そもそもDNAとは？　蛋白質とは？

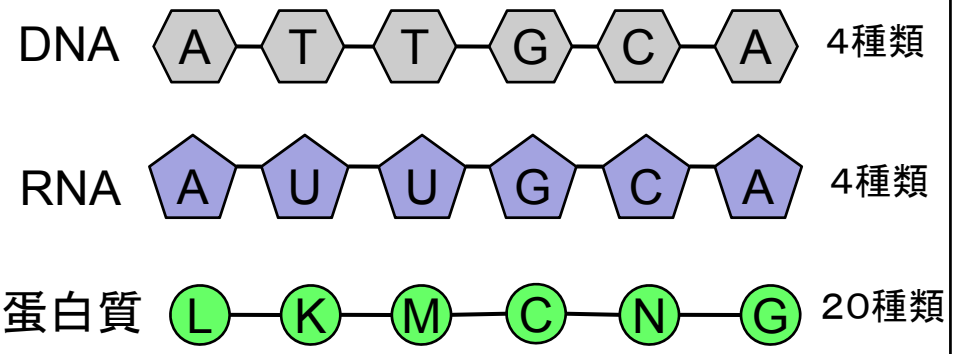
（2）どんなデータベースに、どのように収納されているか？

分子生物学の基礎

DNA→RNA→蛋白質の情報の流れ

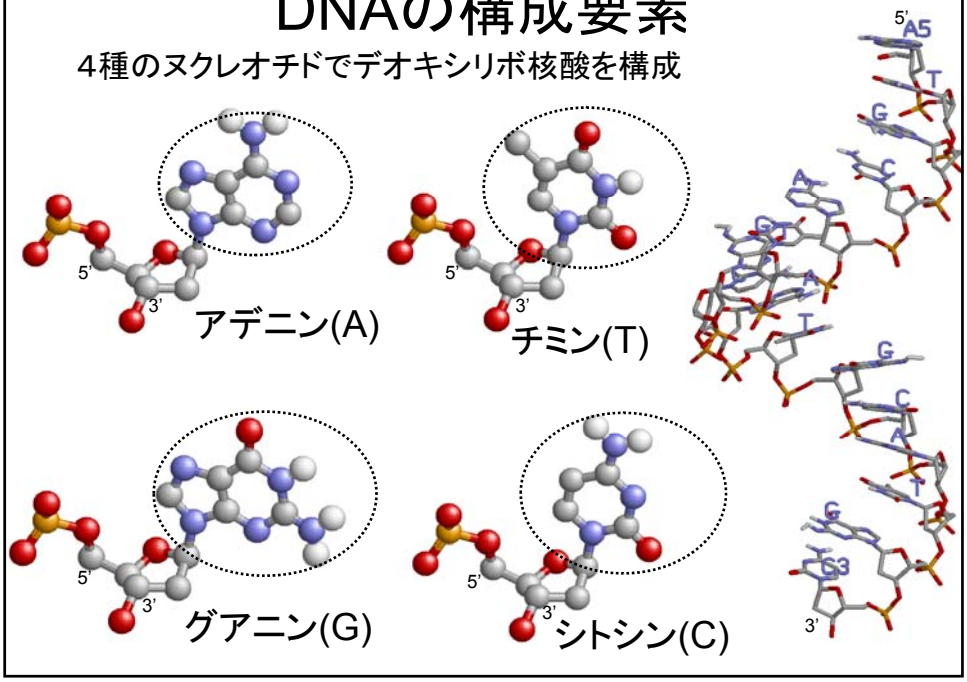
三つの重要な高分子
～ DNA, RNA,蛋白質 ～

これら三つはいずれも重合体(polymer)、つまり
ある単位となる分子（monomer）が一列に並んだ形



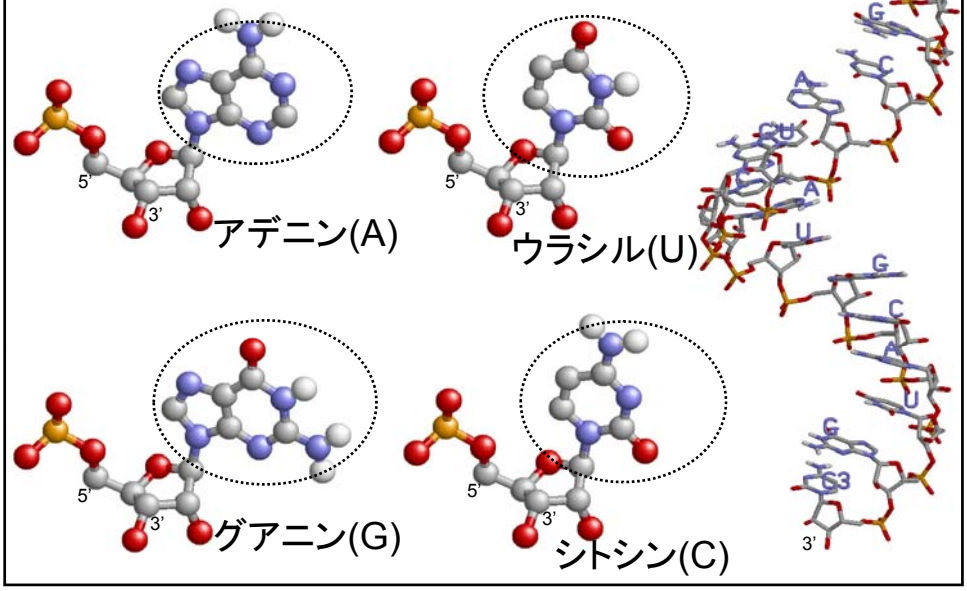
DNAの構成要素

4種のヌクレオチドでデオキシリボ核酸を構成

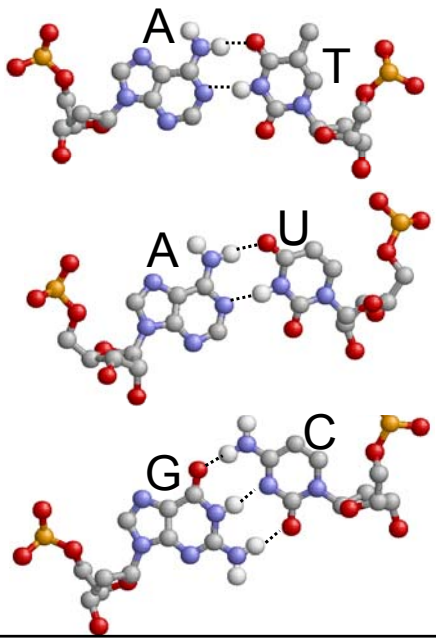


RNAの構成要素

4種のヌクレオチドでリボ核酸を構成



相補的な塩基対構造

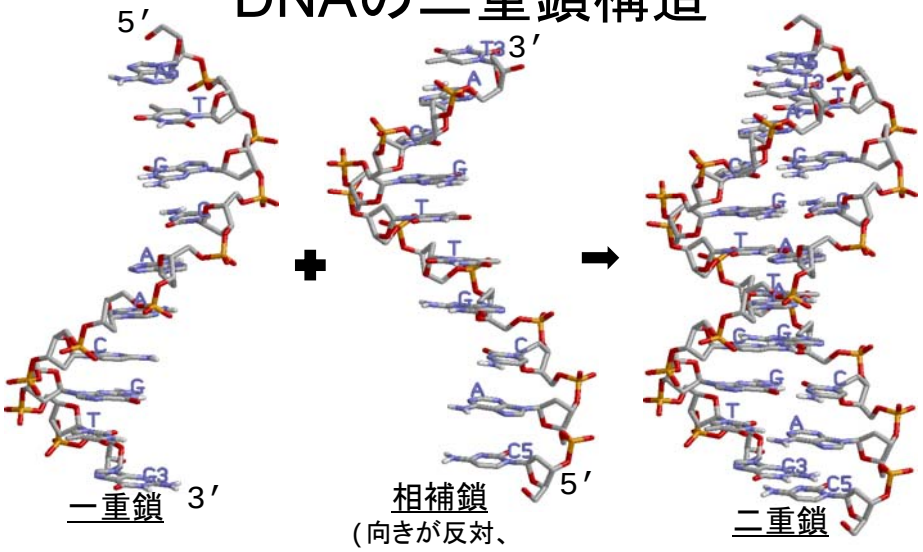


・向かいあう塩基どうしが水素結合を作ることで、相補的な塩基対を作る

・A-T, A-U, G-Cの三種のペアが可能

・DNAどうし、RNAどうしだけでなく、DNAとRNAのペアも可能

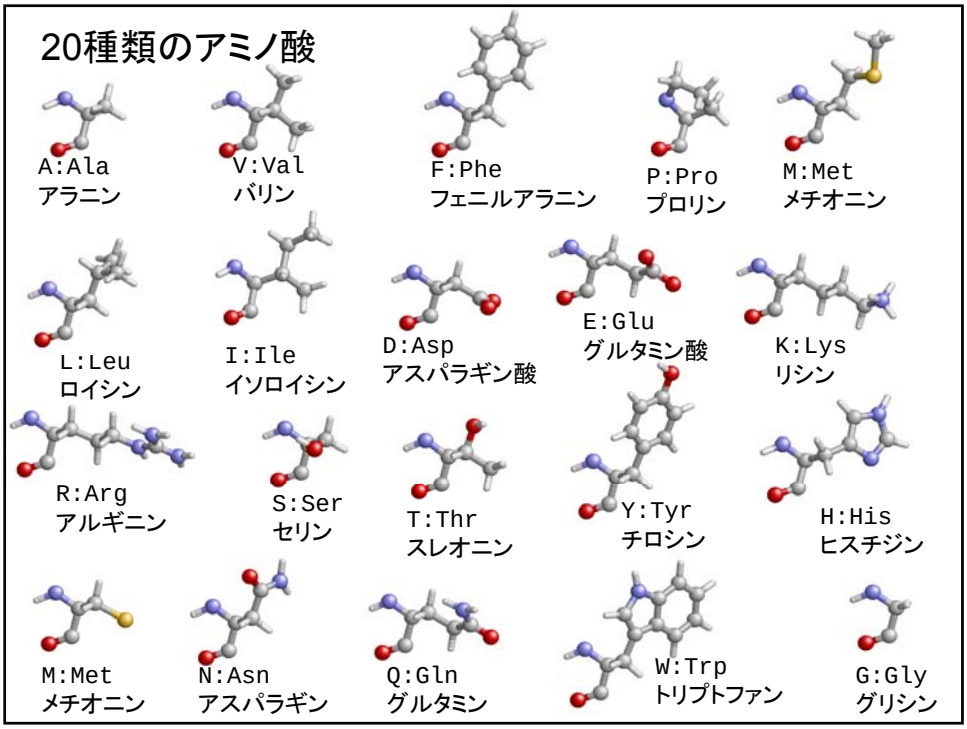
DNAの二重鎖構造



5′ATGCAACGTG3′ ✚ 5′CACGTTGCAT3′ ➡

5′ATGCAACGTG3′
||||||||||
3′TACGTTGCAC5′

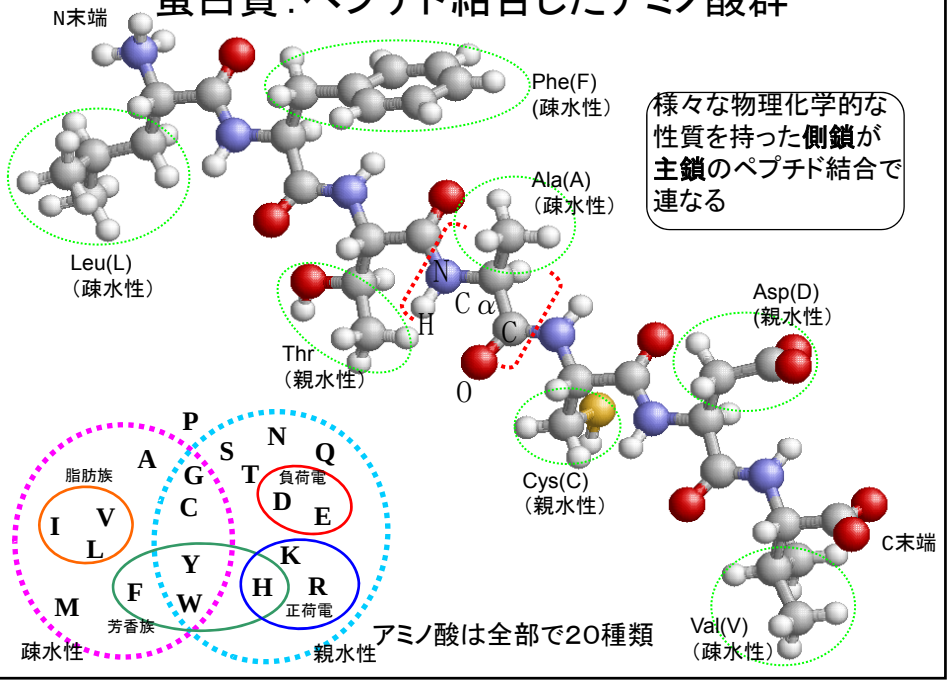
20種類のアミノ酸



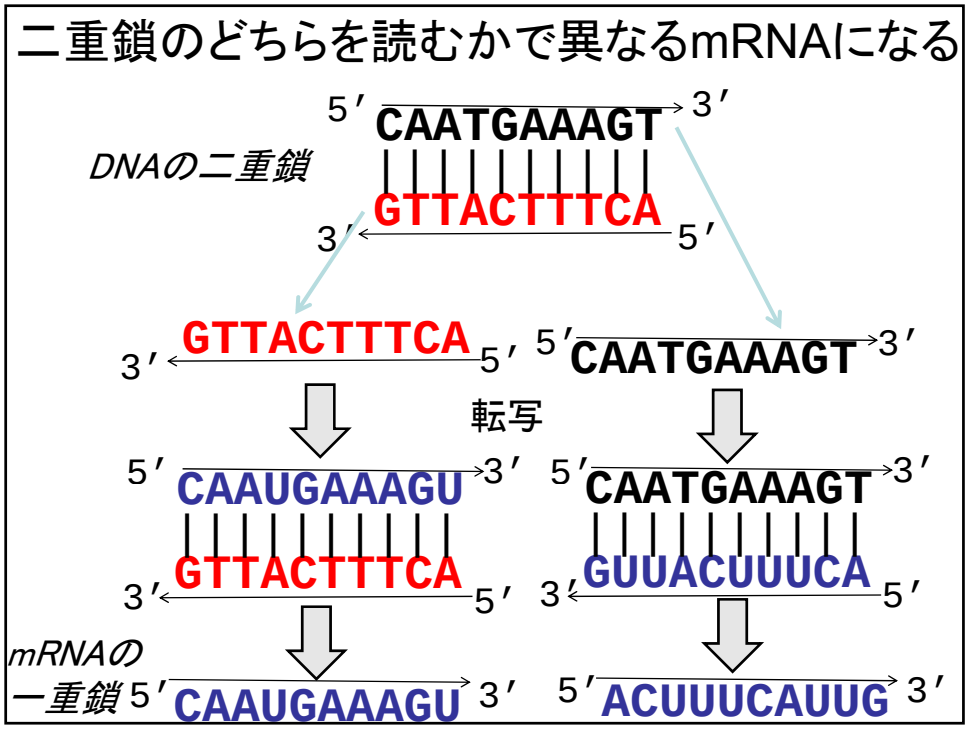
アミノ酸の一文字表記を覚えましょう

- アラニン（A）
- ロイシン（L）
- フェニルアラニン（F）
- トリプトファン（W）
- リジン（K）
- グルタミン（Q）
- グルタミン酸（E）
- アスパラギン（N）
- アスパラギン酸（D）

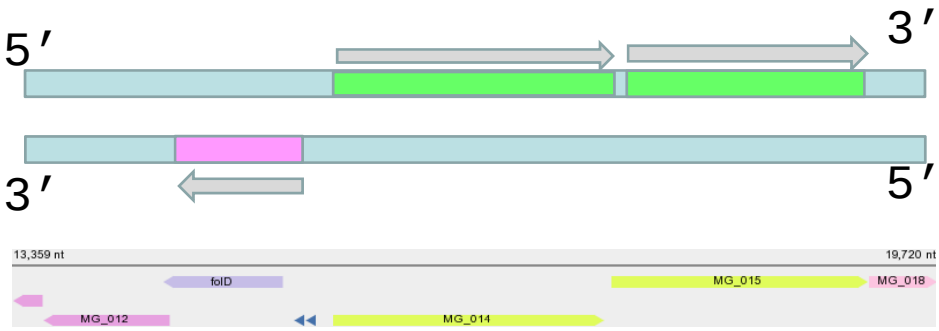
蛋白質：ペプチド結合したアミノ酸群




セントラルドグマ
5′
3′
AAATGAAAATATTAATTAATAAAAAG
DNA
TTTACTTTTATAATTAATTATTTTTC
3′
5′
転写 (T->U)
5′
3′
AAAUGAAAAUAUUAAUUAAUAAAAAG
mRNA
翻訳 (コドン表)
5′
3′
AAAUGAAAAUAUUAAUUAAUAAAAAG
蛋白質
M K I L I N K K


二重鎖のどちらを読むかで異なるmRNAになる
5′ CAATGAAAGT 3′
DNAの二重鎖
GTTACTTTCA
3′ 5′
GTTACTTTCA
3′ 5′
5′ CAATGAAAGT 3′
転写
5′ CAAUGAAAGU 3′
GTTACTTTCA
3′ 5′
5′ CAATGAAAGT 3′
GUUACUUUCA
3′ 5′
mRNAの
一重鎖
5′ CAAUGAAAGU 3′
5′ ACUUUCAUUG 3′

転写されるmRNAの方向は二方向ともある

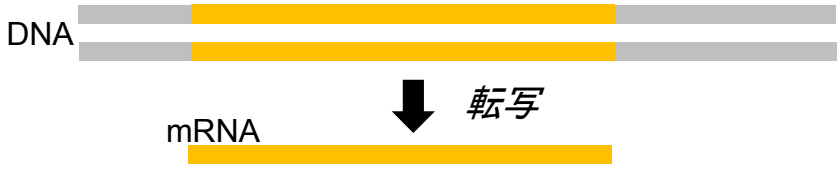


*Mycoplasma genitalium*のゲノムの一部

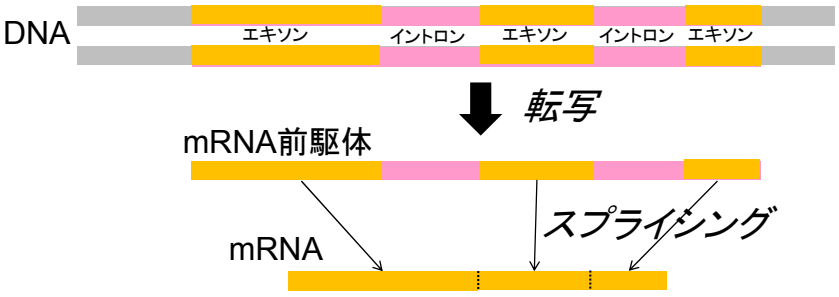
ゲノムデータベースではどちらかの方向を＋、もう一方が－として記載される。

真核生物はエキソン・イントロン構造を持つ

・原核生物(prokaryote)の場合



・真核生物(eukaryote)の場合



遺伝暗号（コドン表）:RNA

UUU	**F:Phe**	UCU	**S:Ser**	UAU	**Y:Tyr**	UGU	**C:Cys**
UUC		UCC		UAC		UGC	
UUA	**L:Leu**	UCA		UAA	終止	UGA	終止
UUG		UCG		UAG		UGG	**W:Trp**
CUU	**L:Leu**	CCU	**P:Pro**	CAU	**H:His**	CGU	**R:Arg**
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	**Q:Gln**	CGA	
CUG		CCG		CAG		CGG	
AUU	**I:Ile**	ACU	**T:Thr**	AAU	**N:Asn**	AGU	**S:Ser**
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	**K:Lys**	AGA	**R:Arg**
AUG	**M:Met**（開始）	ACG		AAG		AGG	
GUU	**V:Val**	GCU	**A:Ala**	GAU	**D:Asp**	GGU	**G:Gly**
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	**E:Glu**	GGA	
GUG		GCG		GAG		GGG	

遺伝暗号（コドン表）:DNA

TTT	**F:Phe**	TCT	**S:Ser**	TAT	**Y:Tyr**	TGT	**C:Cys**
TTC		TCC		TAC		TGC	
TTA	**L:Leu**	TCA		TAA	終止	TGA	終止
TTG		TCG		TAG		TGG	**W:Trp**
CTT	**L:Leu**	CCT	**P:Pro**	CAT	**H:His**	CGT	**R:Arg**
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	**Q:Gln**	CGA	
CTG		CCG		CAG		CGG	
ATT	**I:Ile**	ACT	**T:Thr**	AAT	**N:Asn**	AGT	**S:Ser**
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	**K:Lys**	AGA	**R:Arg**
ATG	**M:Met**（開始）	ACG		AAG		AGG	
GTT	**V:Val**	GCT	**A:Ala**	GAT	**D:Asp**	GGT	**G:Gly**
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	**E:Glu**	GGA	
GTG		GCG		GAG		GGG	

mRNAの翻訳の例

mRNA

AGCAAUGAAAAUAUUAAUUAAUAAAUAACGA

（1）まず開始コドンのAUGを探す

AGCAAUGAAAAUAUUAAUUAAUAAAUAACGA

（2）そのまま3文字ずつスライドしてコドン表に従って翻訳

AGCAAUGAAAAUAUUAAUUAAUAAAUAACGA

M K I

（3）終始コドン（UAA, UAG, UGA）が現れたら終了

AGCAAUGAAAAUAUUAAUUAAUAAAUAACGA

M K I L I N K 終止

翻訳に関係するいくつかの用語

gene（遺伝子）：生物学的情報を含んでいるDNAの部分領域であり、RNAあるいは蛋白質をコードする部分。

CDS：CoDing Sequenceの略。蛋白質をコードしている核酸配列の領域。

ORF：Open Reading Frameの略。開始コドンから始まり、終止コドンで終わる核酸配列の領域

DNA配列からアミノ酸配列を予測できるか？

イントロンのない原核生物の場合

・6通りの読み枠(reading frame)を全て試し、
・開始コドンで始まり終止コドンで終わる領域
(open reading frame)を抽出
・十分長い領域を翻訳されるアミノ酸配列として予測

AGCAAUGAAAAUAUUAAUUAAUAAAUAAC

S N E N I N x x I

A M K I L I N K x

Q x K Y x L I N N

※一つの方向あたり三つの読み枠がある。
相補鎖にも三つあるので、全部で6つの読み枠。

配列決定とバイオインフォマティクス：学籍番号：＿＿＿＿＿＿＿＿＿＿　氏名：＿＿＿＿＿＿＿＿＿＿

問1．以下のDNA配列の3つの読み枠について、それぞれ
対応するアミノ酸を1文字表記で記せ。終止コドンは'x'と書け。

GATGAATGTATTTGCCTGAGTCTTTCTGAAA

GATGAATGTATTTGCCTGAGTCTTTCTGAAA

GATGAATGTATTTGCCTGAGTCTTTCTGAAA

問2．最も長いORFに対応するアミノ酸配列は何か。以下に記せ。

アミノ酸配列：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

生命情報学：2009．4．7

配列決定とバイオインフォマティクス：学籍番号：＿＿＿＿＿＿＿＿＿＿　氏名：＿＿＿＿＿＿＿＿＿＿

問1．以下のDNA配列の3つの読み枠について、それぞれ
対応するアミノ酸を1文字表記で記せ。終止コドンは'x'と書け。

GATGAATGTATTTGCCTGAGTCTTTCTGAAA
D E C I C L S L S E

GATGAATGTATTTGCCTGAGTCTTTCTGAAA
M N V F A x V F L K

GATGAATGTATTTGCCTGAGTCTTTCTGAAA
x M Y L P E S F x

問2．最も長いORFに対応するアミノ酸配列は何か。以下に記せ。

アミノ酸配列：MYLPESF

生命情報学：2009．4．7

より正確に遺伝子を予測するには？

専用の遺伝子予測プログラムの使用が推奨
(GeneHacker, GeneMark, Glimmer)

- 開始コドンの前の配列の特徴
- 遺伝子領域の塩基配列の規則性

真核生物の遺伝子の予測

エキソン部分だけうまく抽出する必要があり、大変難しい

・真核生物用の遺伝子発見プログラムも開発されている
（Genscan, HMMgene, Grail II, GeneParser）

・mRNAのデータ(cDNAやEST)の利用が手堅い

・既知の遺伝子との類似領域の比較(blastxなど)も有効

配列データベースの成り立ち

DNA,RNA,蛋白質の配列を決める実験法

⇒直接配列を計測できるのはDNAだけ

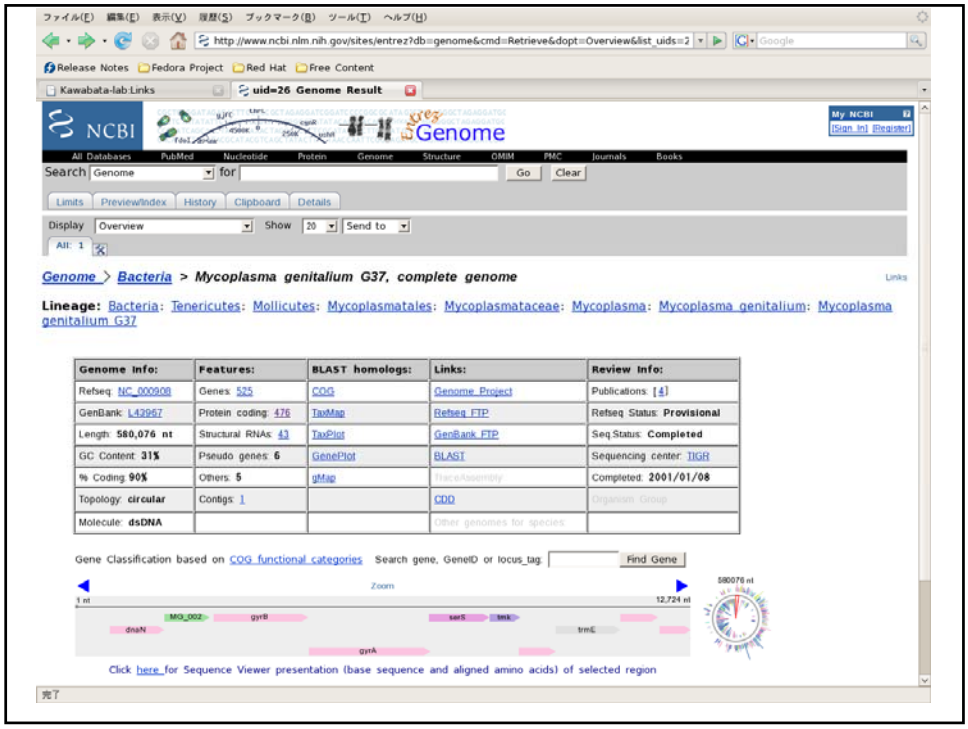
- DNA配列はPCR技術を用いて、注目領域を増幅し、ジデオキシ法を用いたDNAシーケンサを使って決定することができる。
- RNA配列は、RNAをDNAに逆転写し、そのDNA配列を決定することで、得ることができる。
- 蛋白質からそのアミノ酸配列を得るのは難しく、特に全長を得るのは極めて難しい。DNAかRNAの配列を解釈することでアミノ酸配列を得ることが一般的。

国際塩基配列データベース

日本：DDBJ（遺伝研）
米国：GenBank（NCBI）
欧州：EMBL-Bank（EBI）

研究者は決定したDNA配列を三つのデータベースのどれかに登録

どのデータベースに登録しても、データは共有される



GenBankのファイルフォーマット（1）：ヘッダー部

```
LOCUS       NC_000908             580076 bp    DNA     circular BCT 02-FEB-2009
DEFINITION  Mycoplasma genitalium G37, complete genome.
ACCESSION   NC_000908
VERSION     NC_000908.2  GI:108885074
KEYWORDS    .
SOURCE      Mycoplasma genitalium G37
  ORGANISM  Mycoplasma genitalium G37
            Bacteria; Tenericutes; Mollicutes; Mycoplasmataceae; Mycoplasma.
REFERENCE   1  (bases 1 to 580076)
  AUTHORS   Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R.,
            Maruf,M., Hutchison,C.A., Smith,H.O. and Venter,J.C.
  TITLE     Essential genes of a minimal bacterium
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 103 (2), 425-430 (2006)
   PUBMED   16407165
REFERENCE   2  (bases 1 to 580076)
  AUTHORS   Peterson,S.N., Bailey,C.C., Jensen,J.S., Borre,M.B., King,E.S.,
            Bott,K.F. and Hutchison,C.A.III.
  TITLE     Characterization of repetitive DNA in the Mycoplasma genitalium
            genome: possible role in the generation of antigenic variation
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 92 (25), 11829-11833 (1995)
   PUBMED   8524858
REFERENCE   3  (bases 1 to 580076)
  AUTHORS   Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A.,
            Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G.G.,
            Kelley,J.M., Fritchman,J.L., Weidman,J.F., Small,K.V., Sandusky,M.,
            Fuhrmann,J.L., Nguyen,D.T., Utterback,T., Saudek,D.M.,
            Phillips,C.A., Merrick,J.M., Tomb,J., Dougherty,B.A., Bott,K.F.,
```

GenBankのファイルフォーマット(2)：FEATUREの表

```
FEATURES             Location/Qualifiers
     gene            686..1828
                     /gene="dnaN"
                     /locus_tag="MG_001"
                     /db_xref="GeneID:875454"
     CDS             686..1828
                     /gene="dnaN"
                     /locus_tag="MG_001"
                     /EC_number="2.7.7.7"
                     /note="identified by sequence similarity; putative"
                     /codon_start=1
                     /transl_table=4
                     /product="DNA polymerase III, beta subunit"
                     /protein_id="NP_072661.2"
                     /db_xref="GI:108885075"
                     /db_xref="GeneID:875454"
                     /translation="MKILINKSELNKILKKMNNVIISNNKIKPHHSYFLIEAKEKEIN
                     FYANNEYFSVKCNLNKNIDILEQGSLIVKGKIFNDLINGIKEEIITIQEKDQTLLVKT
                     KKTSINLNTINVNEFPRIRFNEKNDLSEFNQFKINYSLLVKGIKKIFHSVSNNREISS
                     KFNGVNFNGSNGKEIFLEASDTYKLSVFEIKQETEPFDFILESNLLSFINSFNPEEDK
                     SIVFYYRKDNKDSFSTEMLISMDNFMISYTSVNEKFPEVNYFFEFEPETKIVVQKNEL
                     KDALQRIQTLAQNERTFLCDMQINSSELKIRAIVNNIGNSLEEISCLKFEGYKLNISF
                     NPSSLLDHIESFESNEINFDFQGNSKYFLITSKSEPELKQILVPSR"
     gene            1828..2760
                     /locus_tag="MG_002"
                     /db_xref="GeneID:875221"
     CDS             1828..2760
```

GenBankのファイルフォーマット(3)：FEATUREの表

```
     gene            complement(12701..13564)
                     /locus_tag="MG_011"
     CDS             complement(12701..13564)
                     /locus_tag="MG_011"
                     /note="identified by sequence similarity; putative"
                     /codon_start=1
                     /product="hypothetical protein"
                     /protein_id="NP_072671.1"
                     /translation="MGKIKLKNRKALVVYDNKDDFEKNQTFALSLIKELQKKKLNAEV
                     LLLENKDINFEAKINEAELILNRSRKVDFLKTNNQINTFLVNPFNVVFIANDKYETYK
                     WLKQNRFLTVNSSLLSKETIKSFPVIVKKRNSHGGKDVHLVNSADEIKHLNIENATEW
                     IVQPFLSIGTVEYRAYILFGKIIKVIKKISNANQFKANFSQGAEVSLFKLKWFTKRKI
                     KKIAKRLREGYYAIDFFLNRYNRVIVNEIEDAAGARALVQLCPDLNITKIIIRTIISK
                     FKKFLKKKLIS"
     gene            complement(15294..15369)
                     /locus_tag="MG_471"
                     /note="MG_t01"
                     /db_xref="GeneID:875702"
     tRNA            complement(15294..15369)
                     /locus_tag="MG_471"
                     /product="tRNA-Ala"
     gene            complement(15375..15451)
                     /locus_tag="MG_472"
                     /old_locus_tag="MGt02"
                     /db_xref="GeneID:875218"
     tRNA            complement(15375..15451)
                     /locus_tag="MG_472"
```

GenBankのファイルフォーマット(4)：塩基配列

```
ORIGIN
        1 taagttatta tttagttaat acttttaaca atattattaa ggtatttaaa aaatactatt
       61 atagtattta acatagttaa atacttcct taatactgtt aaattatatt caatcaatac
      121 atatataata ttattaaaat acttgataag tattatttag atattagaca aatactaatt
      181 ttatattgct ttaatactta ataaatacta cttatgtatt aagtaaatat tactgtaata
      241 ctaataacaa tattattaca atatgctaga ataatattgc tagtatcaat aattactaat
      301 atagtattag gaaaatacca taataatatt tctacataat actaagttaa tactatgtgt
      361 agaataataa ataatcagat taaaaaaatt ttatttatct gaaacatatt taatcaattg
      421 aactgattat tttcagcagt aataattaca tatgtacata gtacatatgt aaaatatcat
      481 taatttctgt tatatataat agtatctatt ttagagagta ttaattatta ctataattaa
      541 gcatttatgc ttaattataa gctttttatg aacaaaatta tagacatttt agttcttata
      601 ataaataata gatattaaag aaaataaaaa aatagaaata aatatcataa cccttgataa
      661 cccagaaatt aatacttaat caaaaatgaa aatattaatt aataaaagtg aattgaataa
      721 aattttgaaa aaaatgaata acgttattat ttccaataac aaaataaaac cacatcattc
      781 atatttttta atagaggcaa aagaaaaaga aataaacttt tatgctaaca atgaatactt
      841 ttctgtcaaa tgtaatttaa ataaaaatat tgatattctt gaacaaggct ccttaattgt
      901 taaaggaaaa atttttaacg atcttattaa tggcataaaa gaagagatta ttactattca
      961 agaaaaagat caaacacttt tggttaaaac aaaaaaaaca agtattaatt taaacacaat
     1021 taatgtgaat gaatttccaa gaataaggtt taatgaaaaa aacgatttaa gtgaatttaa
     1081 tcaattcaaa ataaattatt cacttttagt aaaaggcatt aaaaaaattt ttcactcagt
     1141 ttcaaataat cgtgaaatat cttctaaatt taatggagta aatttcaatg gatccaatgg
     1201 aaaagaaata tttttagaag cttctgacac ttataaacta tctgtttttg agataaagca
     1261 agaaacagaa ccatttgatt tcattttgga gagtaattta cttagtttca ttaattcttt
     1321 taatcctgaa gaagataaat ctattgtttt ttattacaga aaagataata aagatagctt
     1381 tagtacagaa atgttgattt caatggataa ctttatgatt agttacacat cggttaatga
     1441 aaaatttcca gaggtaaact acttttttga atttgaacct gaaactaaaa tagttgttca
:
```

GenBankのファイルフォーマット(5)：塩基配列

```
    :
    12601 aactaagcaa ggatttataa caaaagttat agaaattaaa gctgccgcaa aagactgaaa
    12661 tgatttgttt ttattaaaca actcaaattg atcagcggtt ttaactaatc aacttctttt
    12721 ttaagaattt tttaaattta ctaataattg ttctgataat tattttagtg atatttaaat
    12781 ctggacaaag ctgaactaaa gctctcgcac cagcagcatc ttcaatttca ttaacaataa
    12841 ccctattata tctatttaaa aagaagtcaa tagcataata accttccctt aggcgtttag
    12901 ctatttttctt tatttttctt ttagtaaatc actttaattt aaacaaggaa acttcagcac
    12961 cttgtgaaaa gttagcttta aattgattag cattagaaat ttttttaata actttaatta
    13021 tttttccaaa caaaatataa gcacgatatt caactgtgcc aattgataaa aaaggttgaa
    13081 caattcattc tgttgcattt tcaatgttta aatgtttgat ctcgtcagca ctattaacta
    13141 aatgtacatc ttttccaccg tgtgaattac gtttcttaac gatgacagga aatgatttga
    13201 ttgtttcttt actaagaaga gaagaattga cagttagaaa tctatttgt tttaatcatt
    13261 tatatgtttc gtatttatcg tttgctataa aaacaacatt aaaaggatta actaaaaaag
    13321 tatttatttg attattggtt tttaaaaaat ctacttttct tgaacgattt aaaatcaatt
    13381 cagcttcatt aattttagct tcgaaattaa tgtctttatt ttcaagtaat aagacttcag
    13441 catttagttt tttcttttgt aattccttga ttagacttaa agcaaatgtt tgatttttt
    13501 caaaatcatc cttgttgtca taaacaacta atgcttttct gttttttaat ttaatttttc
    13561 ccattaatct aaattgcttt taaaagctca attgcaagat tagtatttaa atacattgag
    13621 cttcttgtta attgcacatt aggatttact tcacaaaaga tcaatgatct gtcttgatca
    13681 aacaaaaaat caataccgca ataaaaaagt tgcattactt tactaatttt aactgctaaa
    13741 ttttcttgtt ccttattcaa aaaaaagcgt tctgcctttg ccccttttatt gagattagaa
    13801 cgaaaatcac tattattagt tgtatgtaaa gcacctataa ctttattgtt cacaacaata
    :
   579961 atgatcctgc aacattagtt gccattgtag tttttaatac gccgccttta ttatttacaa
   580021 aagaaatgat catatattta aatgattata atatttcttt aatactaaaa aaatac
//
```

核酸配列に付加される主なFEATURE

gene ：遺伝子

CDS ：Coding Sequence

tRNA ：transfer RNA(運搬RNA)

※配列情報以外に付加される情報
のことをアノテーション(annotation)と呼ぶ。

FEATUREの領域の書き方

```
CDS     1828..2760
```

1828～2760番目の塩基配列

```
CDS     complement(1807..2169)
```

1807～2169番目の相補鎖の塩基配列

```
CDS     join(7287..7388,7502..7753)
```

7287～7388番目と7502～7733番目の配列を加えた配列
（複数のエキソンからなる遺伝子の記述に用いる）

```
CDS     complement(join(7287..7388,7502..7753))
```

7287～7388番目と7502～7733番目の配列を加えた配列
の相補鎖の配列。
（複数のエキソンからなる遺伝子が相補鎖にある場合）

問3. *Mycoplasma genitalium*のゲノム配列データ(NC_000908)のデータを見て、以下の問いに答えよ。

(1)DNA配列は全部で_____塩基ペアである。

(2)遺伝子dnaNは、_____番目から_____番目の領域のDNA配列にコードされている。

(3)遺伝子dnaNがコードされているDNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

DNA　：_ _ _ _ _ _ _ _ _

アミノ酸　：_____ _____ _____

(4)遺伝子MG_011がコードされているmRNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

mRNA　：_ _ _ _ _ _ _ _ _

アミノ酸　：_____ _____ _____

問4.今回の講義の中で(重要そうなのに)よくわからなかったことがあれば以下に自由に書いてください。

問3. *Mycoplasma genitalium*のゲノム配列データ(NC_000908)のデータを見て、以下の問いに答えよ。

(1)DNA配列は全部で580076塩基ペアである。

(2)遺伝子dnaNは、_____番目から_____番目の領域のDNA配列にコードされている。

(3)遺伝子dnaNがコードされているDNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

DNA　：_ _ _ _ _ _ _ _ _

アミノ酸　：_____ _____ _____

(4)遺伝子MG_011がコードされているmRNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

mRNA　：_ _ _ _ _ _ _ _ _

アミノ酸　：_____ _____ _____

問4.今回の講義の中で(重要そうなのに)よくわからなかったことがあれば以下に自由に書いてください。

GenBankのファイルフォーマット(2):FEATUREの表

```
FEATURES             Location/Qualifiers
     gene            686..1828
                     /gene="dnaN"
                     /locus_tag="MG_001"
                     /db_xref="GeneID:875454"
     CDS             686..1828
                     /gene="dnaN"
                     /locus_tag="MG_001"
                     /EC_number="2.7.7.7"
                     /note="identified by sequence similarity; putative"
                     /codon_start=1
                     /transl_table=4
                     /product="DNA polymerase III, beta subunit"
                     /protein_id="NP_072661.2"
                     /db_xref="GI:108885075"
                     /db_xref="GeneID:875454"
                     /translation="MKILINKSELNKILKKMNNVIISNNKIKPHHSYFLIEAKEKEIN
                     FYANNEYFSVKCNLNKNIDILEQGSLIVKGKIFNDLINGIKEEIITIQEKDQTLLVKT
                     KKTSINLNTINVNEFPRIRFNEKNDLSEFNQFKINYSLLVKGIKKIFHSVSNNREISS
                     KFNGVNFNGSNGKEIFLEASDTYKLSVFEIKQETEPFDFILESNLLSFINSFNPEEDK
                     SIVFYYRKDNKDSFSTEMLISMDNFMISYTSVNEKFPEVNYFFEFEPETKIVVQKNEL
                     KDALQRIQTLAQNERTFLCDMQINSSELKIRAIVNNIGNSLEEISCLKFEGYKLNISF
                     NPSSLLDHIESFESNEINFDFQGNSKYFLITSKSEPELKQILVPSR"
     gene            1828..2760
                     /locus_tag="MG_002"
                     /db_xref="GeneID:875221"
     CDS             1828..2760
```

問3. *Mycoplasma genitalium*のゲノム配列データ(NC_000908)のデータを見て、以下の問いに答えよ。

(1)DNA配列は全部で____***580076***____塩基ペアである。

(2)遺伝子dnaNは、____***686***____番目から____***1828***____番目の領域のDNA配列にコードされている。

(3)遺伝子dnaNがコードされているDNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

DNA　:_ _ _ _ _ _ _ _ _

アミノ酸　:_____ _____ _____

(4)遺伝子MG_**011**がコードされているmRNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

mRNA　:_ _ _ _ _ _ _ _ _

アミノ酸　:_____ _____ _____

問4.今回の講義の中で(重要そうなのに)よくわからなかったことがあれば以下に自由に書いてください。

GenBankのファイルフォーマット(4):塩基配列

```
ORIGIN
        1 taagttatta tttagttaat acttttaaca atattattaa ggtatttaaa aaatactatt
       61 atagtattta acatagttaa atacttcct taatactgtt aaattatatt caatcaatac
      121 atatataata ttattaaaat acttgataag tattatttag atattagaca aatactaatt
      181 ttatattgct ttaatactta ataaatacta cttatgtatt aagtaaatat tactgtaata
      241 ctaataacaa tattattaca atatgctaga ataatattgc tagtatcaat aattactaat
      301 atagtattag gaaaatacca taataatatt tctacataat actaagttaa tactatgtgt
      361 agaataataa ataatcagat taaaaaaatt ttatttatct gaaacatatt taatcaattg
      421 aactgattat tttcagcagt aataattaca tatgtacata gtacatatgt aaaatatcat
      481 taatttctgt tatatataat agtatctatt ttagagagta ttaattatta ctataattaa
      541 gcatttatgc ttaattataa gctttttatg aacaaaatta tagacatttt agttcttata
      601 ataaataata gatattaaag aaaataaaaa aatagaaata aatatcataa cccttgataa
      661 cccagaaatt aatacttaat caaaaatgaa aatattaatt aataaaagtg aattgaataa
      721 aattttgaaa aaaatgaata acgttattat ttccaataac aaaataaaac cacatcattc
      781 atatttttta atagaggcaa aagaaaaaga aataaacttt tatgctaaca atgaatactt
      841 ttctgtcaaa tgtaatttaa ataaaaatat tgatattctt gaacaaggct ccttaattgt
      901 taaaggaaaa atttttaacg atcttattaa tggcataaaa gaagagatta ttactattca
      961 agaaaaagat caaacacttt tggttaaaac aaaaaaaaca agtattaatt taaacacaat
     1021 taatgtgaat gaatttccaa gaataaggtt taatgaaaaa aacgatttaa gtgaatttaa
     1081 tcaattcaaa ataaattatt cacttttagt aaaaggcatt aaaaaaattt ttcactcagt
     1141 ttcaaataat cgtgaaatat cttctaaatt taatggagta aatttcaatg gatccaatgg
     1201 aaaagaaata tttttagaag cttctgacac ttataaacta tctgtttttg agataaagca
     1261 agaaacagaa ccatttgatt tcattttgga gagtaattta cttagtttca ttaattcttt
     1321 taatcctgaa gaagataaat ctattgtttt ttattacaga aaagataata aagatagctt
     1381 tagtacagaa atgttgattt caatggataa ctttatgatt agttacacat cggttaatga
     1441 aaaatttcca gaggtaaact acttttttga atttgaacct gaaactaaaa tagttgttca
//
```

問3. *Mycoplasma genitalium*のゲノム配列データ(NC_000908)のデータを見て、以下の問いに答えよ。

(1)DNA配列は全部で__580076__塩基ペアである。

(2)遺伝子dnaNは、__686__番目から__1828__番目の領域のDNA配列にコードされている。

(3)遺伝子dnaNがコードされているDNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

DNA :a t g a a a a t a

アミノ酸 : M K I

(4)遺伝子MG_011がコードされているmRNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

mRNA :_ _ _ _ _ _ _ _ _

アミノ酸 :_____ _____ _____

問4.今回の講義の中で(重要そうなのに)よくわからなかったことがあれば以下に自由に書いてください。

GenBankのファイルフォーマット(3):FEATUREの表

```
     gene            complement(12701..13564)
                     /locus_tag="MG_011"
     CDS             complement(12701..13564)
                     /locus_tag="MG_011"
                     /note="identified by sequence similarity; putative"
                     /codon_start=1
                     /product="hypothetical protein"
                     /protein_id="NP_072671.1"
                     /translation="MGKIKLKNRKALVVYDNKDDFEKNQTFALSLIKELQKKKLNAEV
                     LLLENKDINFEAKINEAELILNRSRKVDFLKTNNQINTFLVNPFNVVFIANDKYETYK
                     WLKQNRFLTVNSSLLSKETIKSFPVIVKKRNSHGGKDVHLVNSADEIKHLNIENATEW
                     IVQPFLSIGTVEYRAYILFGKIIKVIKKISNANQFKANFSQGAEVSLFKLKWFTKRKI
                     KKIAKRLREGYYAIDFFLNRYNRVIVNEIEDAAGARALVQLCPDLNITKIIIRTIISK
                     FKKFLKKKLIS"
     gene            complement(15294..15369)
                     /locus_tag="MG_471"
                     /note="MG_t01"
                     /db_xref="GeneID:875702"
     tRNA            complement(15294..15369)
                     /locus_tag="MG_471"
                     /product="tRNA-Ala"
     gene            complement(15375..15451)
                     /locus_tag="MG_472"
                     /old_locus_tag="MGt02"
                     /db_xref="GeneID:875218"
     tRNA            complement(15375..15451)
                     /locus_tag="MG_472"
```

GenBankのファイルフォーマット(5):塩基配列

```
     :
     12601 aactaagcaa ggatttataa caaaagttat agaaattaaa gctgccgcaa aagactgaaa
     12661 tgatttgttt ttattaaaca actcaaattg atcagcggtt ttaactaatc aacttctttt
     12721 ttaagaattt tttaaattta ctaataattg ttctgataat tattttagtg atatttaaat
     12781 ctggacaaag ctgaactaaa gctctcgcac cagcagcatc ttcaatttca ttaacaataa
     12841 ccctattata tctatttaaa aagaagtcaa tagcataata accttccctt aggcgtttag
     12901 ctatttttctt tatttttctt ttagtaaatc actttaattt aaacaaggaa acttcagcac
     12961 cttgtgaaaa gttagcttta aattgattag cattagaaat ttttttaata actttaatta
     13021 tttttccaaa caaaatataa gcacgatatt caactgtgcc aattgataaa aaaggttgaa
     13081 caattcattc tgttgcattt tcaatgttta aatgtttgat ctcgtcagca ctattaacta
     13141 aatgtacatc ttttccaccg tgtgaattac gtttcttaac gatgacagga aatgatttga
     13201 ttgtttcttt actaagaaga gaagaattga cagttagaaa tctattttgt tttaatcatt
     13261 tatatgtttc gtatttatcg tttgctataa aaacaacatt aaaaggatta actaaaaaag
     13321 tatttatttg attattggtt tttaaaaaat ctacttttct tgaacgattt aaaatcaatt
     13381 cagcttcatt aattttagct tcgaaattaa tgtctttatt ttcaagtaat aagacttcag
     13441 catttagttt tttcttttgt aattccttga ttagacttaa agcaaatgtt tgatttttt
     13501 caaaatcatc cttgttgtca taaacaacta atgcttttct gtttttaat ttaattttc
     13561 ccattaatct aaattgcttt taaaagctca attgcaagat tagtatttaa atacattgag
     13621 cttcttgtta attgcacatt aggatttact tcacaaaaga tcaatgatct gtcttgatca
     13681 aacaaaaaat caataccgca ataaaaaagt tgcattactt tactaatttt aactgctaaa
     13741 ttttcttgtt ccttattcaa aaaaaagcgt tctgcctttg ccccttatt gagattagaa
     13801 cgaaaatcac tattattagt tgtatgtaaa gcacctataa ctttattgtt cacaacaata
     :
    579961 atgatcctgc aacattagtt gccattgtag tttttaatac gccgccttta ttatttacaa
    580021 aagaaatgat catatattta aatgattata atatttcttt aatactaaaa aaatac
//
```

問3. *Mycoplasma genitalium*のゲノム配列データ(NC_000908)のデータを見て、以下の問いに答えよ。

(1)DNA配列は全部で__580076__塩基ペアである。

(2)遺伝子dnaNは、__686__番目から__1828__番目の領域のDNA配列にコードされている。

(3)遺伝子dnaNがコードされているDNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。

DNA　:a t g a a a a t a

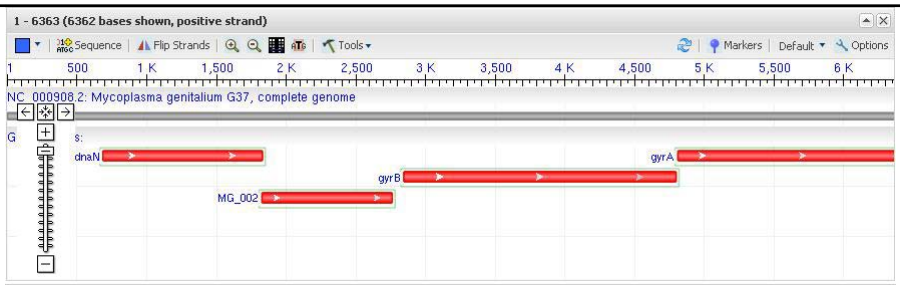
アミノ酸　:__M__　__K__　__Y__

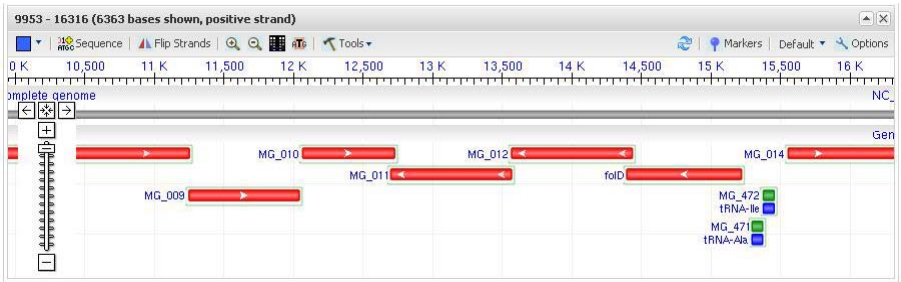
(4)遺伝子MG_011がコードされているmRNA配列の最初の9文字とアミノ酸配列の最初の3文字を書け。t a c c t t t t

mRNA　:a u g g g a a a a

アミノ酸　:__M__　__G__　__K__

問4.今回の講義の中で(重要そうなのに)よくわからなかったことがあれば以下に自由に書いてください。





RefSeq

NCBIで管理されている参照配列データベース

(RefSeqはReference Sequenceの略)

GenBankのデータは、研究者がそれぞれ登録しているので、塩基配列の精度、アノテーションの質・量にばらつきが大きい。また、重複しているデータも多い。

RefSeqでは、いくつかのモデル生物に絞り、それらを専門家が整理して、重複がなく、かつ一定の質のアノテーションになるように、再構成したデータベース

※アクセッション番号の最初の文字が、染色体はNC，mRNAはNM、タンパク質はNPとなっている。

蛋白質の配列のデータベース

nr

・GenBankに登録されている核酸データのFEATUREに記載されているアミノ酸配列を集めたデータベース

・配列数は800万ぐらい（2009年3月）

・蛋白質の名前やその機能については、統一した記載がされていない

・配列の精度が低く、同じような配列が少しずつ異なる長さで登録されていることが多い。

蛋白質の配列のデータベース
Uniprot

・GenBank等のアミノ酸配列データをもとに、専門職員が一つずつチェックし、冗長性を除き、統一した形式で、命名・機能情報を記載したデータベース

・配列数は41万ぐらい（2009年3月）

・以前はSWISS-PROTと呼ばれていたが、他のデータベースを統合してUniprotの名称となった。

・[遺伝子名]_[生物種名]の形式のID.
例)HBA_HUMAN, HBA_MOUSE, TPIS_ECOLI,TPIS_YEAST

Uniprotのデータの例（1）

```
ID   TPIS_CHICK              Reviewed;         248 AA.
AC   P00940;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 2.
DT   20-JAN-2009, entry version 79.
DE   RecName: Full=Triosephosphate isomerase;
DE            Short=TIM;
DE            EC=5.3.1.1;
DE   AltName: Full=Triose-phosphate isomerase;
GN   Name=TPI1;
OS   Gallus gallus (Chicken).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves;
OC   Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus.
OX   NCBI_TaxID=9031;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RX   MEDLINE=85166263; PubMed=3885220;
RA   Straus D., Gilbert W.;
RT   "Chicken triosephosphate isomerase complements an Escherichia coli
RT   deficiency.";
RL   Proc. Natl. Acad. Sci. U.S.A. 82:2014-2018(1985).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=86310829; PubMed=3837846;
RA   Straus D., Gilbert W.;
```

Uniprotのデータの例（2）

```
RN   [4]
RP   X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS), AND SEQUENCE REVISION.
RX   MEDLINE=75175220; PubMed=1134550; DOI=10.1038/255609a0;
RA   Banner D.W., Bloomer A.C., Petsko G.A., Phillips D.C., Pogson C.I.,
RA   Wilson I.A., Corran P.H., Furth A.J., Milman J.D., Offord R.E.,
RA   Priddle J.D., Waley S.G.;
RT   "Structure of chicken muscle triose phosphate isomerase determined
RT   crystallographically at 2.5-A resolution using amino acid sequence
RT   data.";
RL   Nature 255:609-614(1975).
CC   -!- CATALYTIC ACTIVITY: D-glyceraldehyde 3-phosphate = glycerone
CC       phosphate.
CC   -!- PATHWAY: Carbohydrate biosynthesis; gluconeogenesis.
CC   -!- PATHWAY: Carbohydrate degradation; glycolysis; D-glyceraldehyde
3-
CC       phosphate from glycerone phosphate: step 1/1.
CC   -!- SUBUNIT: Homodimer.
CC   -!- SIMILARITY: Belongs to the triosephosphate isomerase family.
DR   EMBL; M11314; AAA49094.1; -; mRNA.
DR   EMBL; M11941; AAA49095.1; -; Genomic_DNA.
DR   PIR; A23448; ISCHT.
DR   RefSeq; NP_990782.1; -.
DR   UniGene; Gga.4148; -.
DR   PDB; 1SPQ; X-ray; 2.16 A; A/B=1-248.
DR   PDB; 1SQ7; X-ray; 2.85 A; A/B=1-248.
DR   PDBsum; 1SPQ; -.
```

Uniprotのデータの例（3）

```
DR   GO; GO:0004807; F:triose-phosphate isomerase activity; IEA:InterPro.
DR   InterPro; IPR013785; Aldolase_TIM..
DR   Gene3D; G3DSA:3.20.20.70; Aldolase_TIM; 1.
DR   PANTHER; PTHR21139; Triophos_ismrse; 1.
DR   Pfam; PF00121; TIM; 1.
DR   ProDom; PD001005; Triophos_ismrse; 1.
DR   TIGRFAMs; TIGR00419; tim; 1.
DR   PROSITE; PS00171; TIM; 1.
PE   1: Evidence at protein level;
KW   3D-structure; Direct protein sequencing; Fatty acid biosynthesis;
KW   Gluconeogenesis; Glycolysis; Isomerase; Lipid synthesis;
FT   INIT_MET      1      1       Removed.
FT   CHAIN         2    248       Triosephosphate isomerase.
FT                               /FTId=PRO_0000090121.
FT   ACT_SITE     95     95       Electrophile.
FT   ACT_SITE    165    165       Proton acceptor.
FT   BINDING      13     13       Substrate.
FT   MUTAGEN      95     95       H->N: Reduces activity 5000-fold.
FT   CONFLICT     17     18       DK -> KR (in Ref. 3; AA sequence).
SQ   SEQUENCE   248 AA;  26620 MW;  AFCC258E574DE982 CRC64;
     MAPRKFFVGG NWKMNGDKKS LGELIHTLNG AKLSADTEVV CGAPSIYLDF ARQKLDAKIG
     VAAQNCYKVP KGAFTGEISP AMIKDIGAAW VILGHSERRH VFGESDELIG QKVAHALAEG
     LGVIACIGEK LDEREAGITE KVVFEQTKAI ADNVKDWSKV VLAYEPVWAI GTGKTATPQQ
     AQEVHEKLRG WLKSHVSDAV AQSTRIIYGG SVTGGNCKEL ASQHDVDGFL VGGASLKPEF
     VDIINAKH
//
```

参考文献

- B.Alberts他著　Essential 細胞生物学　原書第2版　(2005). 南江堂
- T.A.Brown.　ゲノム第3版. (2007). メディカル・サイエンス・インターナショナル.
- 加納　圭　　ヒトゲノムマップ　（2007）. 京都大学出版会