

ゲノムインフォマティクス あらかると

佐賀大学農学部
和田康彦

本講義のねらい

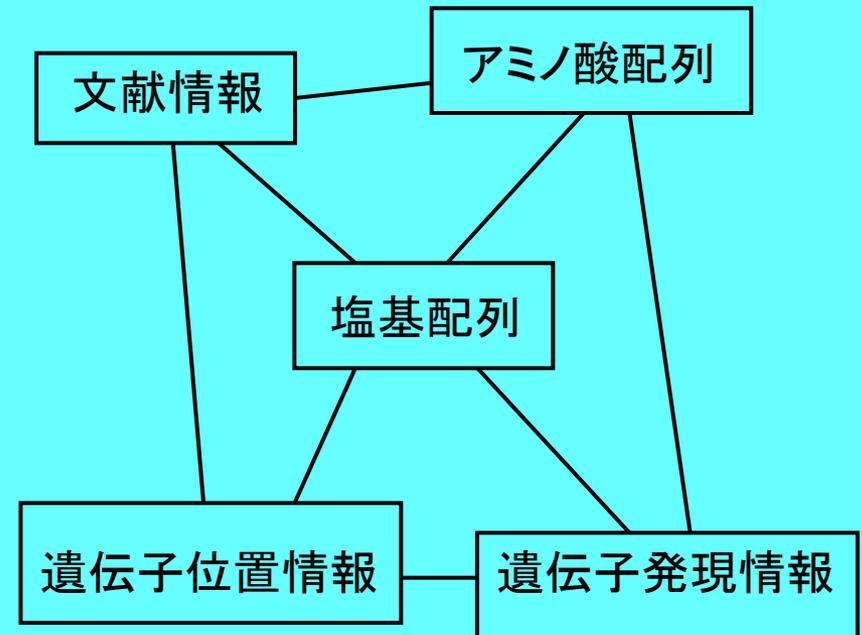
- ゲノムインフォマティクスのおもな手法を理解する
- ゲノム解析の中でのゲノムインフォマティクスの意義を理解する
- ゲノムインフォマティクスの創薬などの応用面での可能性を考察する

ゲノムインフォマティクスとは

- インフォマティクスとは**情報学**のこと
- 技術の進歩によりDNA塩基配列やマッピングなどの**膨大な情報が生産されている**
- それらの情報を**整理し、再編成し、解析して**新たな知識を獲得する必要がある
- 分子生物学の知識に加えて**情報工学、知識工学、統計学**などの技術が必要

ゲノムデータベース

- ゲノム解析によって得られた膨大なデータを集積したデータベース
- DNA塩基配列、アミノ酸配列、染色体上の位置、遺伝子発現などさまざまな情報を格納している
- はじめは個別のデータベースとして開発されたが、次第に統合データベースとして整備
- インターネットを利用した検索が可能なものが多い



塩基配列データベース

- DNAの塩基配列を格納したデータベース
- 当初はフラットファイルの形式で整備された
- **Genbank** (アメリカ)、**EMBL** (EU)、**DDBJ** (日本) の3つのデータベースがあり、**毎日データを交換**している
- 現在では**各研究者がデータを提出**するシステムになっている
- 種単位のシーケンス決定が相次ぎ、塩基配列の登録数は**幾何級数的に増加**している

塩基配列データベースのデータ例 (ブタ 甲状腺ホルモン受容体α mRNA)

タイトル →
アセション番号 →
キーワード →
生物種 →

```
LOCUS       SSAJ5797                1906 bp    RNA        linear    MAM 11-NOV-1999
DEFINITION  Sus scrofa mRNA for thyroid hormone receptor alpha 1 (c-erbA-alpha
            1).
ACCESSION   AJ005797
VERSION     AJ005797.1
KEYWORDS    c-erbA-alpha gene; thyroid hormone receptor alpha 1; transcription
            factor.
SOURCE      pig.
ORGANISM    Sus scrofa
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Cetartiodactyla; Suina; Suidae; Sus.
REFERENCE   1 (bases 1 to 1906)
AUTHORS     White,P.
JOURNAL     Submitted (20-MAY-1998) to the EMBL/GenBank/DBJ databases. White
            P., The Babraham Institute, Department of Neurobiology, Lab 504,
            Babraham, Cambridge, CB2 4AT, UK.
REFERENCE   2
AUTHORS     White,P. and Dauncey,M.J.
TITLE       Differential expression of thyroid hormone receptor isoforms is
            strikingly related to cardiac and skeletal muscle phenotype during
            postnatal development
JOURNAL     J. Mol. Endocrinol. 23, 241-254(1999).
FEATURES    Location/Qualifiers
            source
                1..1906
                /db_xref="taxon:9823"
                /sequenced_mol="cDNA to mRNA"
                /sex="male"
                /organism="Sus scrofa"
                /strain="large white"
                /dev_stage="postnatal"
                /tissue_type="heart"
            5'UTR
                1..136
            CDS
                137..1869
                /codon_start=1
                /db_xref="SWISS-PROT:Q97716"
                /gene="c-erbA-alpha"
                /product="thyroid hormone receptor alpha 1"
                /function="nuclear thyroid hormone receptor"
                /protein_id="CAA06701.1"
                /translation="MEQKPSKYECGSDPEENSARSPDGKRKRKNGQCSLKTMSGYIPS
                YLDKDEQCVYCGDKATGYHYRCITCEGCKGFFRRTIQKNLHPTYSCKYDSCCVYDKITR
                NQCQLCRFKKCIAYGMAMDLVLDSDSKRVAKRKLIEGNRERRRKEEMIRSLQQRPEPTPE
                EWDLIHVATEAHRSTNAQGSHWKQRRKFLPDDIGQSPIYSMPDGDGVLEAFSEFTKII
                TPAITRVYVDFAKKLPMFSELPCEQIILLKGCCMEIMSLRAAVRYDPESDTLTLSGEMA
                YKREQLKNGGLGYVSDAIFELGKSLSAFNLDDTEVALLQAVLLMSTDRSGLLCYDKIEK
                SQEAYLLAFEHYYNHRKHNIPHFWPKLLMKVTDLRMIGACHASRFLHMKYECPELTFPP
                LFLEVFEDQEV"
            3'UTR
                1370..1906
BASE COUNT   424 a    557 c    589 g    336 t
ORIGIN
1  agcgggacac catggstgct gtgcctatgg scctgggtgg cggggggtgg gtggcctgtg
61  cctctggcgg cggggcact gtggcggcgg cctctcttgg cctctctgg cggcctctg
```

文献情報 →

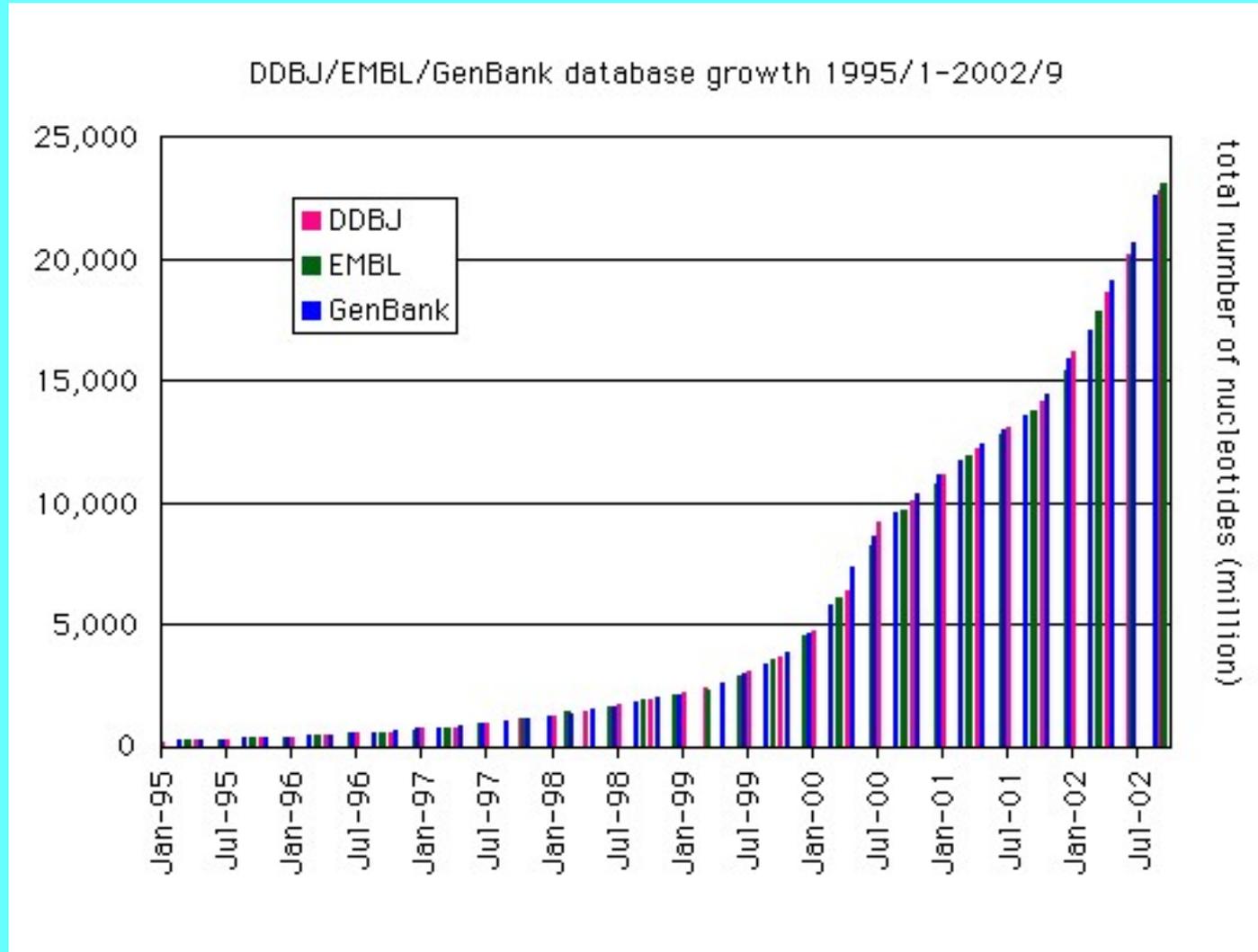
材料 →

アノテーション →

← アミノ酸配列

← 塩基配列

塩基配列の登録数の推移



アミノ酸配列データベース

- 人手でデータを収集、入力しているものが多い
- **SWISS-PROT**, **PIR**, **PRF/SEQDB**などがある
- アミノ酸配列を**ファミリー分類**しているデータベースや、**アミノ酸ドメイン**、**アミノ酸モチーフ**のデータベースもある

アミノ酸ドメインとアミノ酸モチーフ

- タンパク質の**構造や機能**に関して、**特異的なアミノ酸配列**
- ドメインとはアミノ酸の鎖が折りたたまれて**ひとまとまりになっている領域**
- 既に決定された**タンパク質の立体構造**から推定される
- モチーフとは**タンパク質の機能に関連している**と考えられるアミノ酸配列
- In vitroの実験結果などから推定される

遺伝病データベース

- ヒトの遺伝子、遺伝病をカタログ化したOMIMが有名

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search OMIM for [] Go Clear

Limits Preview/Index History Clipboard Details

Display Detailed Save Text Clip Add

***170998** Lit

PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR-ALPHA; PPARA

Alternative titles; symbols

PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR; PPAR

Gene map locus [22q12-q13.1](#)

TEXT

CLONING

Peroxisome proliferators are a diverse group of chemicals which includes hypolipidem...

Gene map

LocusLink

Nomenclature

RefSeq

GenBank

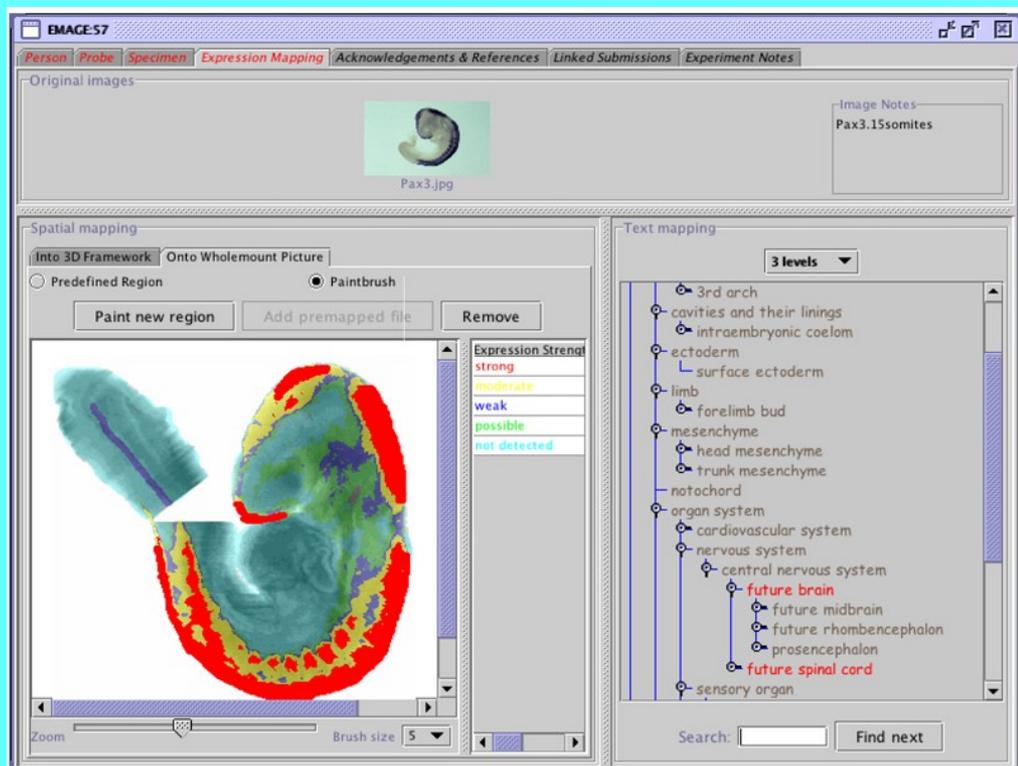
Protein

UniGene

LinkOut

遺伝子発現データベース

- いつ、どの組織でどの遺伝子が発現しているのかをデータベース化したもの



マウスアトラスの表示例

遺伝子 Pax3

マウスゲノムデータベース

- おもにマウスゲノム上の遺伝子の位置を収納したデータベース
- 現在ではマーカー情報や遺伝子カタログ、遺伝子発現情報などを含めた統合データベースとして運用されている
- Jackson Laboratoryが運営している

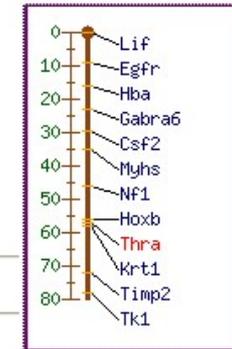
MGD検索例

Type: Gene
Symbol: Thra
Name: thyroid hormone receptor alpha
Chromosome: 11
Cytogenetic Offset: D-E
cM Position: 57.0
MGI Accession ID: MGI:98742

Synonyms: Nr1a1 / T3R[a] / T3Ralpha / TR alpha 1 / c-erbAalpha / Erba / Rvr

Additional Information:

- [Mammalian Homology](#)
- [Marker Mapping Data](#) (26)
- [Phenotypic Alleles](#) (4)
- [RFLP/PCR Polymorphism](#) (8)
- [Gene Expression Data](#) (6 [results](#) in 1 [assay](#))
- [Gene Expression Literature Index Data](#) (6)
- [Molecular Probes and Segments](#) (67)
- [References](#) (90)
- [Mouse Locus Catalog](#)



染色体上の
おおまかな
位置

Gene Classifications: (You can [browse the Gene Ontology \(GO\) Classifications](#))

Category	Classification Term	Evidence	Inferred From	Ref(s)
C	intracellular	IEA		1
C	nucleus	TAS		2
C	nucleus	IEA		2
F	DNA binding	IEA		2
F	ligand-dependent nuclear receptor	IEA		1
F	receptor	IEA		1
F	steroid hormone receptor	IEA		1
F	thyroid hormone receptor	TAS		3

染色体番号 →

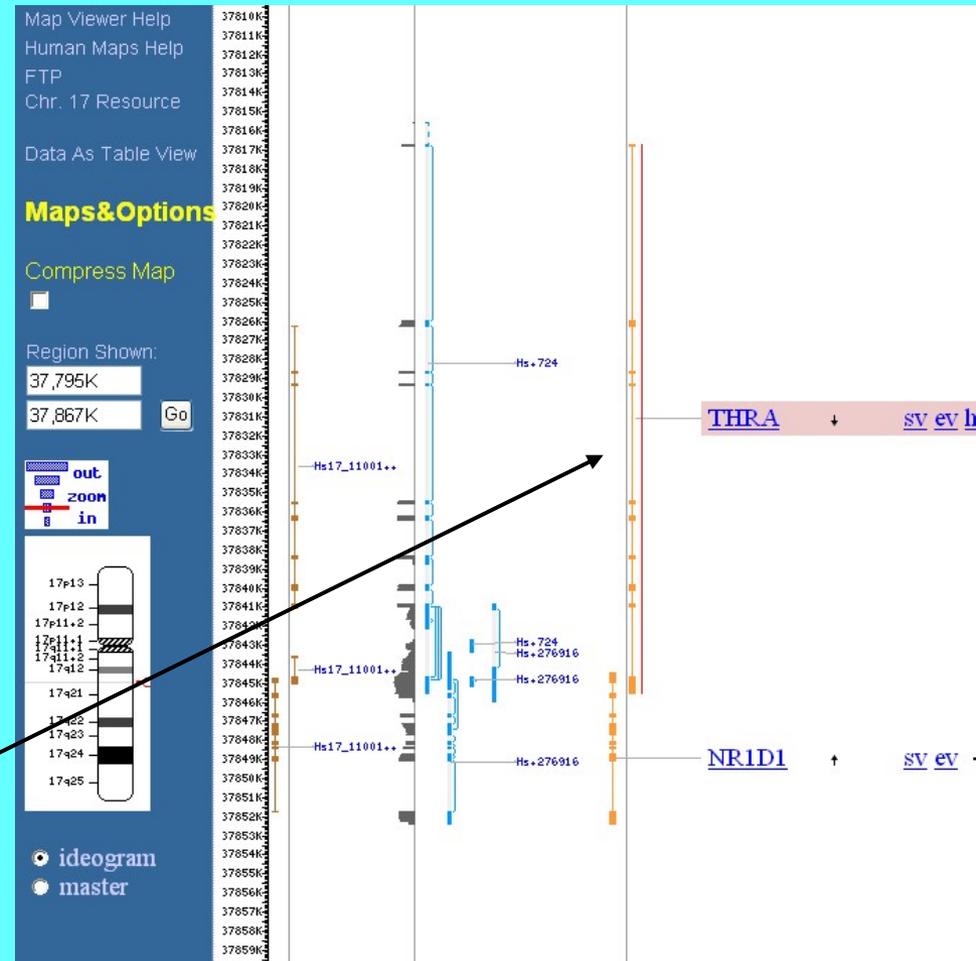
連鎖地図 →

その他の
情報への
リンク →

ヒトの統合データベース

- **LocusLink**と**GDB**が有名
- 遺伝子機能、染色体上での位置表示、塩基配列、アミノ酸配列、OMIM、他の動物種でのホモロジー遺伝子などへのリンクを持っている

THRAの遺伝子領域
30KBにエクソン10個



MapViewerの表示例

アライメント

- 良く似た配列を並置することをアライメントという
- 配列間の関係が良くわかるように並べないといけない
- たくさんの配列を一度に並べることをマルチプルアライメントという
- 配列全体を並べることをグローバルアライメント、配列の一部を並べることをローカルアライメントという

マルチプルアライメントの例

```
NM_000376      MAICIVSPDRPGYQDAALIEAIQDRLSNTLQTYIRCRHPPPGSHLLYAKMIQKLA-----
quail_THRA    QAVLLMSSDRTGLICVDKMEKCOETYLLAFEHYIN--YRKHNIHFHFWPKLLMKVT-----
quail_THRB    QAVLLMSSDRXGLVCYERIEKCOEGFLLAFEHYIN--YRKHNYAHFHWPKLLMKVT-----
NM_000461     QAVLLMSSDRPGLACYERIEKYQDSFLLAFEHYIN--YRKHNYTHFWPKLLMKVT-----
NM_003250     QAVLLMSTDRSGLLCYDKIEKSQEAYLLAFEHYVN--HRKHNIHFHFWPKLLMKEREYQSS
AF109732     QAVILLSSDRPGLTSVERIERCQEEFLLAFEHYIN--YRKHNYAHFHWPKLLMKVT-----
AF146775     QAVLLMSSDRSGLTLVYDKIEKCOETYLLAFEHYIN--HRKHNIHFHFWPKLLMKVT-----
J03933__ra   QAVLLMSSDRPGLACYERIEKYQDSFLLAFEHYIN--YRKHNYTHFWPKLLMKVT-----
M18028__ra   QAVLLMSTDRSGLLCYDKIEKSQEAYLLAFEHYVN--HRKHNIHFHFWPKLLMKVT-----
M35344__Xe   QAVLLMSSDRTGLICTDKIEKCOETYLLAFEHYIN--HRKHNIHFHFWPKLLMKVT-----
M35362__Xe   QAVLLMSSDRPGLSSYERIEKCOEGFLLAFEHYIN--YRKHNIHFHFWPKLLMKVT-----
M65207__ch   QAVLLMSSDRPGLVCYERIEKCOEGFLLAFEHYIN--YRKHNYAHFHWPKLLMKVT-----
U54796__ze   QAVLLMSSDRSGLTCYEKIEKCOEMYLLAFEHYIN--HRKHNIHFHFWPKLLMKVT-----
Y00987__ch   QAVLLMSSDRTGLICVDKIEKCOETYLLAFEHYIN--YRKHNIHFHFWPKLLMKVT-----
AJ002300     QAVLLMSSDRTGLICYEKIEKCOETYLLAFEHYIN--YRKHNIHFHFWPKLLMKVT-----
               * . . * * * * .      . * * .      . . * .      * . . *
```

ホモロジーサーチ

- データベースの中から良く似た配列を探し出すことをホモロジーサーチと呼ぶ
- 厳密だが時間のかかる**FASTA**と、検索時間の早い**blast**などのプログラムがある

```
□ >gi|1695744|dbj|D89547.1 Gallus gallus mRNA for vitellogenin I, complete cds
    Length = 5952
```

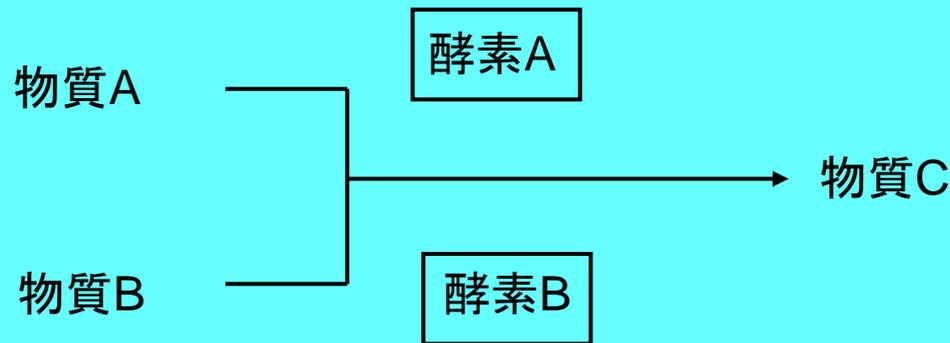
```
Score = 438 bits (221), Expect = e-120
Identities = 256/264 (96%), Gaps = 3/264 (1%)
Strand = Plus / Plus
```

```
Query: 4      ctgaagaatgtgaagtttgatggtgagatccttaaggtt-ctgttgcttcatggatgaga 62
             |||
Sbjct: 5278   ctgaagaatgtgaattttgatggtgagatccttaaggttactgttgcttcatggatgaga 5397

Query: 63      ggaaagac-tgtggagtctgtgaaataatgacagagaaaagcacaatgaactcctaats 121
             |||
Sbjct: 5338   ggaaagacctgtggagtctgtgaaataatgacagagaaaagcacaatgaactcctaats 5397
```

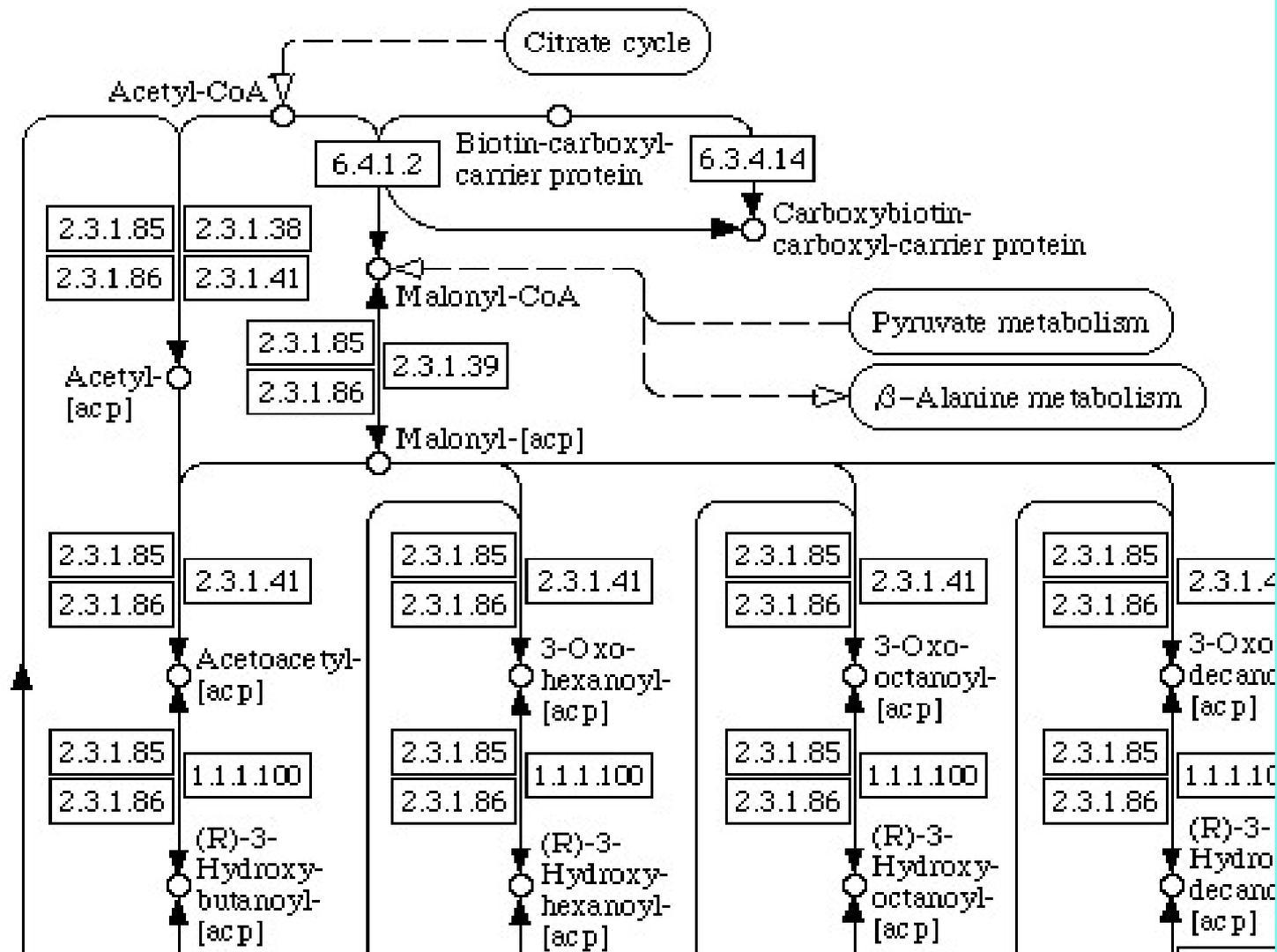
パスウェイ解析

- 生体内の物質の代謝過程を図示したり、検索したりできる
- おもに酵素を中心に記述されている



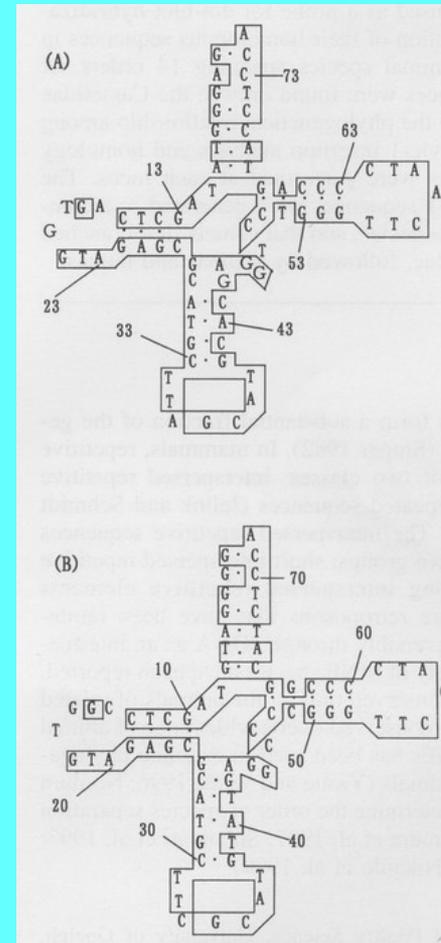
パスウェイの例

(genomenet 脂肪酸の生成)



RNAの2次元構造予測

- 塩基の相補性から1本鎖RNAの形を予測する
- tRNAやSINEのRNAは特徴的なクローバー葉型をしている
- 長大なmRNAの2次元構造を予測するのは困難



マウスのアラニンのtRNA

SINE (Vic-1) の該当領域

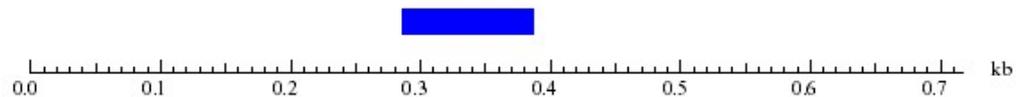
コード領域予測

- 実験的に決定されたDNA塩基配列から**アミノ酸をコードしている部分**をインフォマティクスの技術を用いて予測する
- いろいろな手法が開発されているがmRNAが見つからない場合に**ゲノムDNA塩基配列からの予測精度は80%程度**

コード領域予測の例 (ヒト THRA genescan)

```
>gi|22064780:120206-120924 Homo sapiens chromosome 17 reference genomic contig
TCCCAGTGAGCAGCTGACGTTTATTTTTCCCTCTCCACCATCCTCTGCCTTGTAGATAGATCAAGAAGA
AAGCCCTGAATGGGCAGTCAGAAGAAGTCTTTCATACTCTTAGGGCAGAGATGGACAGGTTCCACC
TCGTGTCTCTCTAGGATTCTACCAGGGAGGCTGGGCTAGGAGAGGGGTCAGAAACAAAAAGTTGAGGGAT
GGGGAAAAGTGTGAGAGCCTCTCTGCTGGGACTCCAAGCCATGCCTGCCTTCCATCCTTTCCCTTCT
CCCAGGGTATATCCCTAGTTACCTGGACAAAAGACGAGCAGTGTGTGTCGTGTGTGGGGACAAGGCAACTGGT
TATCACTACCGCTGTATCACTTGTGAGGGCTGCAAGGTATGGAAGCTACCTCCTGCCCTCCCCTGCCAC
CTGAGCCCCCATATACTCCCTATGTCACCTAAAGCCCGCCTGTTAGGTCATCACTTTTTTTTTTTTTTCT
TTGAGACGGAGTCTCACTCTTGTCCCCCAGGCTGGAGTGCAGTGGCGCGATCATGGCTCACTGCAACCTC
CGCCTTCAGGTTCAAGTGATTCCCCTACCTCGGCCTCCTGAGTAGCTGGGATTACAGGTATGCACCTCCA
CACCCAGCTAGTTTTTTTGTATTTTTAGTAGAGACGGGGTTTCACTATGTGGCCAGGCTGGTCTCGAACT
CCTGACCTCAGGTGACCAC
```

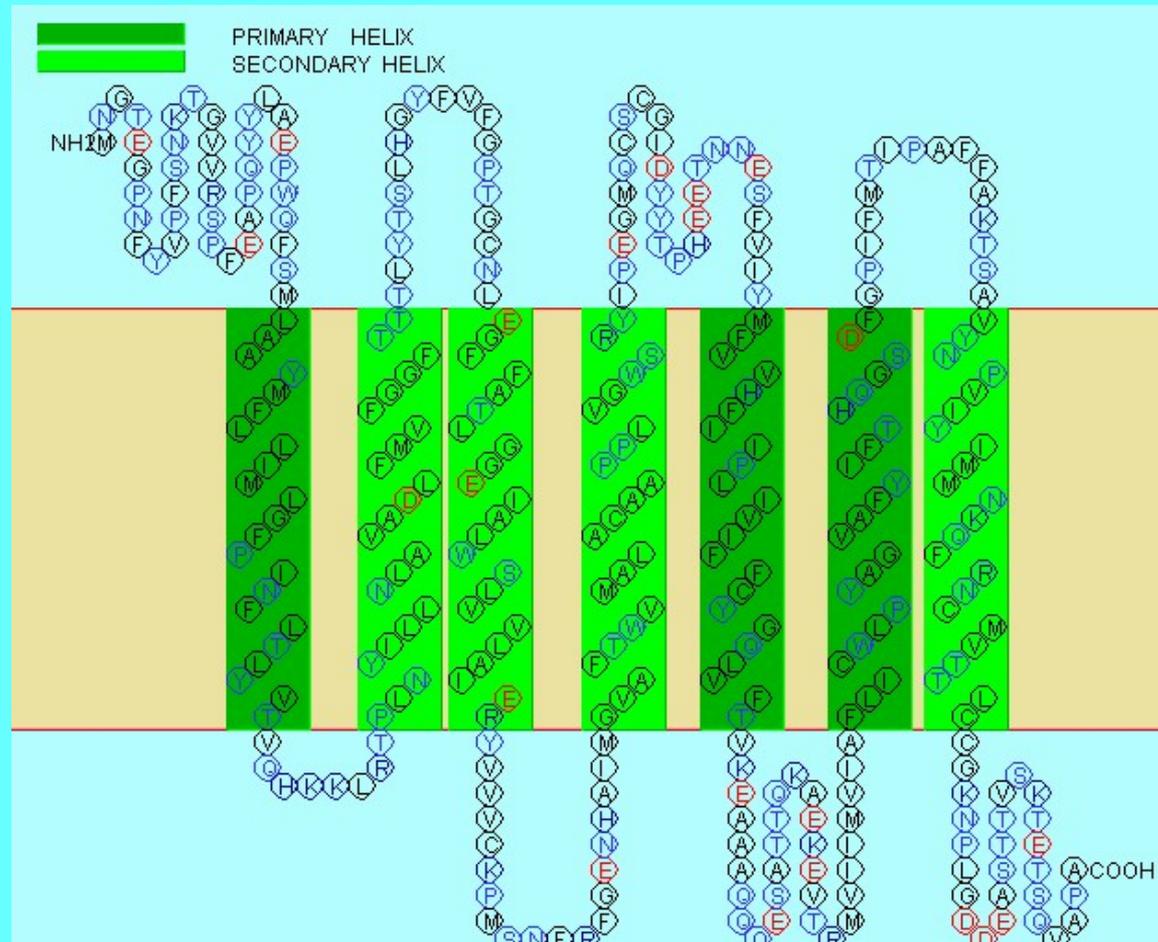
GENSCAN predicted genes in sequence 20:32:08



機能部位予測

- アミノ酸配列などから**遺伝子の機能部位**をある程度予測することができる
- 細胞内局在予測、膜貫通部位予測、修飾部位予測などが実用化されている

膜貫通部位予測の例 (SOSUI ウシ ロドプシン)

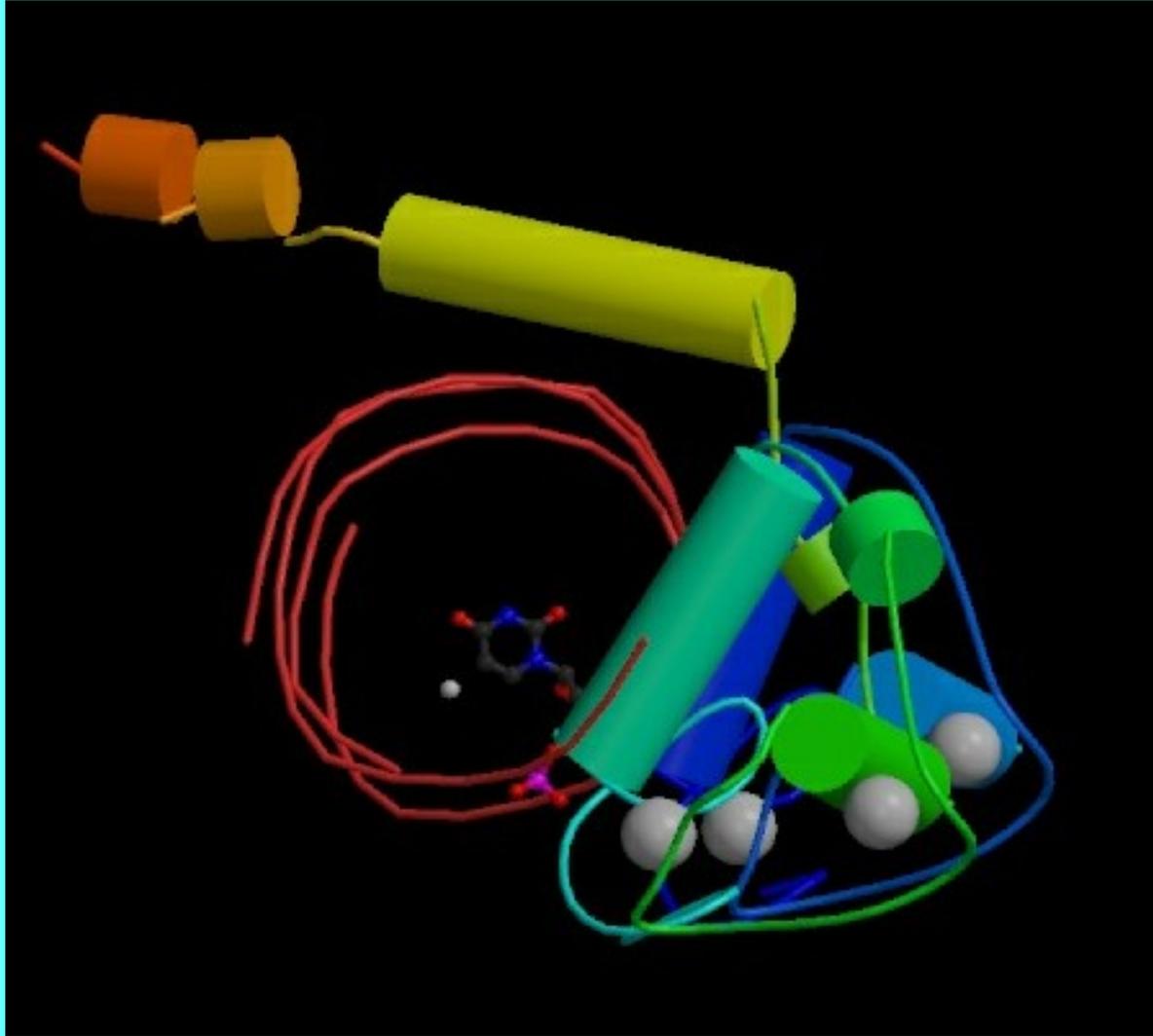


タンパク質の立体構造予測

- タンパク質の立体構造は機能を知る上で非常に重要
- 立体構造を実験的に決定するにはコストがかかる
- 既知のタンパク質のデータをもとに、タンパク質の立体構造を予測することができる
- さまざまな手法が考案されている

タンパク質の立体構造の例

(RXRA+THRA+THRE)



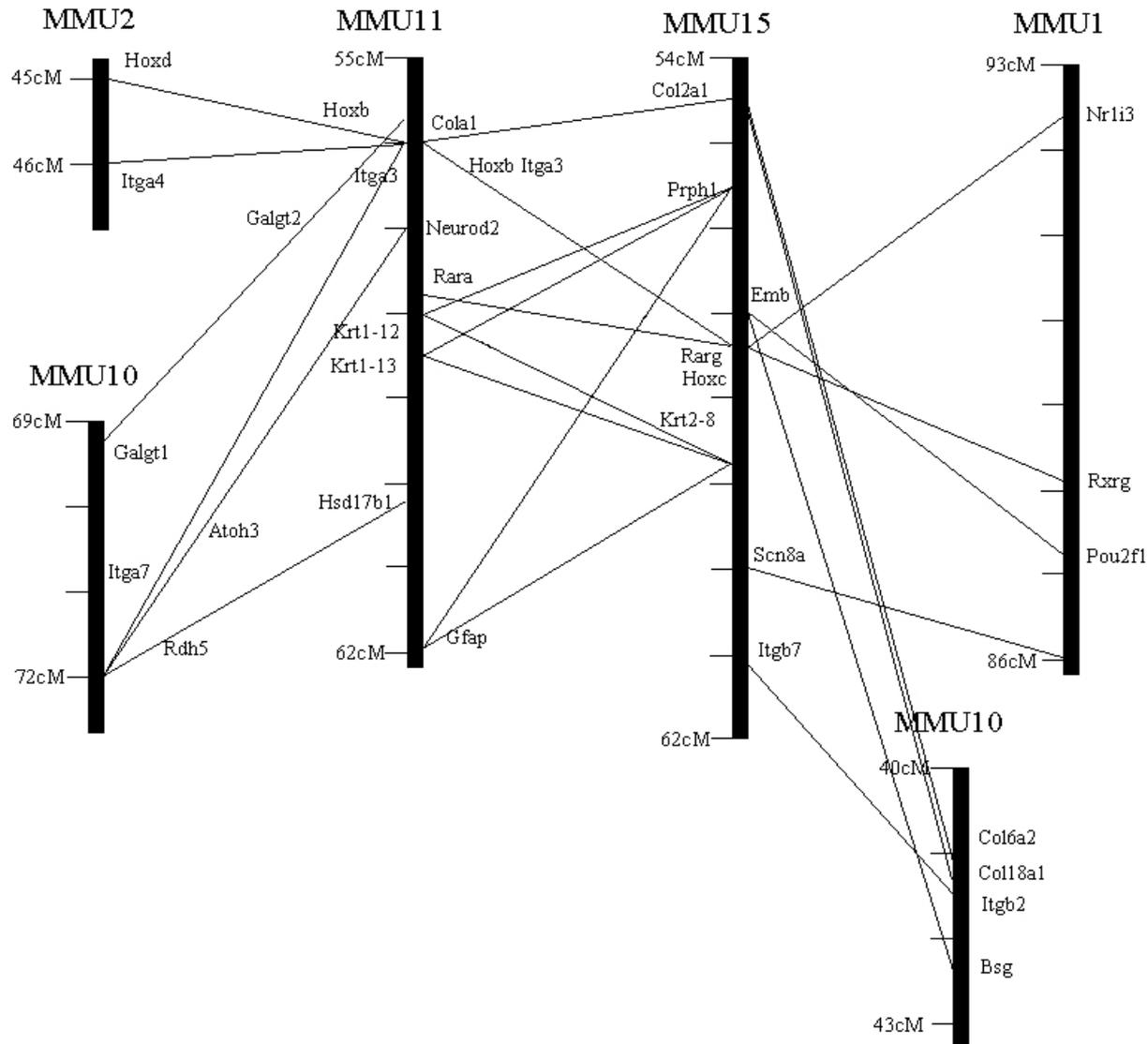
ゲノム上に遺伝子は何個あるか？

- データベースから全長の記録されている mRNA の塩基配列を探す
- **EST** (mRNA の一部) の塩基配列と比較する
- 両者の間にどの程度の重複があるかをもとに遺伝子数を推定する

いろいろな動物の遺伝子数推定値

動物種	ESTデータソース	mRNA	EST	マッチした数	推定遺伝子数
マウス	RIKEN FANTOM-1	3407	19168	1925	33924
ブタ	GB EST database	258	57144	542	27201
ウシ	GB EST database	637	160371	5168	15392
ニワトリ	GB EST database	496	13511	474	17217

マウスゲノム上の重複領域



レポートの課題

- あなたが興味を持ったゲノムインフォマティクスの技術について、800字以内で中学生にもわかるように説明しなさい。